University of Rhode Island

DigitalCommons@URI

Open Access Dissertations

2019

ROBUST ESTIMATION FOR NOISY DATA

Xin Zhou

University of Rhode Island, xin_tz_zhou@hotmail.com

Follow this and additional works at: https://digitalcommons.uri.edu/oa_diss

Recommended Citation

Zhou, Xin, "ROBUST ESTIMATION FOR NOISY DATA" (2019). *Open Access Dissertations*. Paper 1038.
https://digitalcommons.uri.edu/oa_diss/1038

This Dissertation is brought to you for free and open access by DigitalCommons@URI. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons-group@uri.edu.

ROBUST ESTIMATION FOR NOISY DATA

BY

XIN ZHOU

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

ELECTRICAL ENGINEERING

UNIVERSITY OF RHODE ISLAND

2019

DOCTOR OF PHILOSOPHY DISSERTATION

OF

XIN ZHOU

APPROVED:

Dissertation Committee:

Major Professor Steven Kay

Ramdas Kumaresan

Mustafa Kulenovic

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2019

ABSTRACT

When a data set is corrupted by noise, the model for the data generating process is misspecified and can cause parameter estimation problems.

In the case of a Gaussian autoregressive (AR) process corrupted by noise, the data is more accurately modeled as an autoregressive moving average (ARMA) process rather than an AR process. This misspecification leads to bias, and hence, low resolution in AR spectral estimation. However, a new parametric estimator, the realizable information theoretic-estimator (RITE) based on a non-homogeneous Poisson spectral representation, is shown by simulation to be more robust to noise than the asymptotic maximum likelihood estimator (MLE). We therefore conducted an in-depth investigation and analyzed the statistics of RITE and the asymptotic MLE for the misspecified model. For large data records, RITE and the asymptotic MLE are both asymptotically normally distributed. The asymptotic MLE has a slightly lower variance, but RITE exhibits much less bias. Simulation examples of a noise corrupted AR process are provided to support the theoretical properties. This advantage of RITE increases as the SNR decreases.

Another topic of interest is Data fusion for estimation. It is a problem that utilizes information from multiple data sets to estimate an unknown parameter or vector. These data sets are usually from multiple sources, for example, multiple sensors. In this case, this problem is called distributed estimation. It uses data from multiple sensors and a fusion center (FC), or central processor, to achieve a more accurate estimation than using a single sensor observation. In this paper, we propose two estimators for data fusion estimation problem. The Fisher information and the observed Fisher information are used to reduce the negative effects of poor estimations and therefore improve the new estimators' performance in terms of mean square error. At the same time, we found that there is a relationship

between our new estimators and the second order Taylor expansion of l , which is the log likelihood function of data from all sensors. The solution of the maximum of the second order Taylor expansion of l , turns out to be our new estimator that uses the observed Fisher information. Our simulation results showed that the proposed estimators have obvious advantages in both low and intermediate SNR regions, especially when one or many sensors have much lower SNRs than the others.

ACKNOWLEDGMENTS

First of all, I would like to express my deepest gratitude to my advisor Prof. Steven Kay for his guidance and support of my Ph.D. studies. I am lucky and feel honored as his student. I am thankful for the time and patience he spent on me. He is the best Ph.D advisor I could ever imagine. If it is not his invaluable advice, this dissertation would not have been possible. His guidance will benefit me for a life long time in my future career.

Besides my advisor, I would like to thank the rest of my thesis committee members: Prof. Ramdas Kumaresan, Prof. Mustafa Kulenovic, Prof. Marco Alvarez, and Prof. Peter F. Swaszek for their time and efforts in participating in my comprehensive exam and dissertation defense.

I would also like to thank many faculty members, staff, graduate students, undergraduates and friends in the Department of Electrical, Computer and Biomedical Engineering at URI. They offered help and suggestions in several ways.

Last but not the least, I would also like to thank my parents and for their continuous moral and financial support during my study. They have always been very supportive in my life and I am so blessed to have such a wonderful family. Words cannot express how grateful I am for all of the love my family gave and sacrifices they made on my behalf.

PREFACE

This dissertation is constructed in the manuscript format and consists of three manuscripts.

Chapter 1 introduces the RITE spectral estimator proposed by Dr. Steven Kay and also the statistical properties of quasi-MLE and RITE. Both estimators are asymptotically Gaussian distributed but with different means and covariance matrices. Using AR process as simulations examples, this chapter verifies the properties of the two estimators and shows that RITE has advantage over the quasi-MLE when white Gaussian noise is present in data.

Further more, in Chapter 2, the asymptotic statistical properties of the quasi-MLE and RITE are derived. An application to spectral estimation using the AR model is provided in the chapter. Also, for an AR PSD, we prove that the asymptotic Gaussian likelihood function is more sensitive to white noise than the RITE likelihood function. Our experiments show that in comparison to the quasi-MLE, RITE has smaller bias when white noise is present in AR process. This advantage of RITE increases as the signal-to-noise-ratio (SNR) decreases. Besides, it is not limited to white Gaussian noise, but also works for other type of noise like white Laplacian noise and α -Stable modeled impulsive noise.

Chapter 3 proposes two new estimators which use the FIM or the observed FIM in the problem of data fusion for estimation. The former requires the knowledge of the expectation of the log likelihood function, and the latter needs the log likelihood function and data. There is no clear evidence to show which method is better. So in practice, one may choose the method that is easier to implement. The FIM reflects the upper limit of the accuracy of an estimator. By including the FIM or the observed FIM as the weighting factors for each sensor, we can reduce the negative effects of poor estimations and therefore improve the performance of

the final estimator. Two experiments are carried out to test our estimators. Based on our simulation results, the integrated final estimator, which uses the FIM and the one that uses the observed FIM does not have an obvious difference. However, they both reduce the number of outliers and show an improved performance over the averaged estimator in terms of MSE.

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGMENTS	iv
PREFACE	v
TABLE OF CONTENTS	vii
LIST OF FIGURES	x
CHAPTER	
1 A Comparison between Robust Information Theoretic Estimator and Asymptotic Maximum Likelihood Estimator for Misspecified Model	1
Abstract	2
1.1 Introduction	2
1.2 Realizable Information Theoretic Estimator	4
1.3 The Statistical Properties of quasi-MLE and RITE	6
1.3.1 The Statistical Properties of Misspecified MLE	6
1.3.2 The Statistical Properties of RITE	7
1.4 Simulation Examples	8
1.4.1 AR(1) Process Example	8
1.4.2 Noise Sensitivity of Likelihood Function	9
1.5 Discussion and Conclusions	12
List of References	12
2 A Robust Spectral Estimator with Application to a Noise Corrupted Process	18

	Page
Abstract	19
2.1 Introduction	19
2.2 Realizable Information Theoretic Estimator	21
2.3 The Statistical Properties of MLE and RITE	24
2.3.1 The Statistical Properties of Misspecified MLE	24
2.3.2 The Statistical Properties of RITE	26
2.4 Spectral Estimation Application with AR Model	27
2.4.1 White Noise Sensitivity of Likelihood Function	28
2.5 Simulation Examples	31
2.5.1 AR(1) Process Example	32
2.5.2 AR(4) Process Example	34
2.6 Conclusion	37
List of References	38
3 Robust Data Fusion for Estimation	45
Abstract	46
3.1 Introduction	46
3.2 Problem Statement	48
3.2.1 Taylor Expansion of the Log Likelihood Function	49
3.2.2 Definition of Proposed Estimators	51
3.3 Simulation Results	52
3.4 Conclusion and Future Work	58
List of References	58
4 Future Work	60

	Page
APPENDIX	
A Supplementary Material for Chapter 2	61
A.1 Proof of Theorem 1	61
A.2 Proof of Theorem 2	66
A.3	69
A.4 Spectral Estimation for AR Process in Noise Modeled by AR . .	71
A.4.A White Mixture Gaussian Noise	71
A.4.B White Laplacian Noise	71
A.4.C α -Stable Modeled Impulsive Noise	71
A.5 Spectral Estimation for AR process in Noise Modeled by ARMA	73
A.5.A White Gaussian Noise	73
A.5.B α -Stable Modeled Impulsive Noise	73
A.6 Proof of the Statement in Page 23	74
B Supplementary Material for Chapter 3	93
B.1 Analysis of the Estimations at the Third Sensor for Case 1 . . .	93
B.2 Higher Order Taylor Expansion	94
B.3 More Simulation Examples	94
BIBLIOGRAPHY	105

LIST OF FIGURES

Figure		Page
1	AR(1) process, $a[1]=-0.9$, $N=200$	14
2	SNR=40dB, $N=350$	15
3	SNR=35dB, $N=350$	16
4	SNR=15dB, $N=450$	17
5	AR(1) Example: $N \times$ Variance	34
6	AR(1) Example: Mean Square Error, $N=50000$	35
7	AR(1) Example: Variance and Squared Bias, $N=50000$	36
8	100 Overlaid RITE Realizations (WGN)	40
9	Average of RITE Realizations (WGN)	41
10	100 Overlaid Burg Realizations (WGN)	42
11	Average of Burg Realizations (WGN)	43
12	$N=300$, SNR=30dB with WGN	44
13	MSE vs SNR_3 for $N = 100$, $f_0 = 0.1$, $SNR_1 = 1dB$, $SNR_2 = 0dB$. 10,000 simulations for each SNR_3	55
14	MSE vs Δ for $N = 100$, $f_0 = 0.1$, $SNR_1 = \Delta dB$, $SNR_2 = \Delta + 10dB$, $SNR_3 = \Delta + 20dB$. 10,000 simulations for each Δ	56
A1	100 Overlaid RITE Realizations (White Mixture Gaussian Noise)	75
A2	Average of RITE Realizations (White Mixture Gaussian Noise)	76
A3	100 Overlaid Burg Realizations (White Mixture Gaussian Noise)	77
A4	Average of Burg Realizations (White Mixture Gaussian Noise)	78
A5	100 Overlaid RITE Realizations (White Laplacian Noise)	79
A6	Average of RITE Realizations (White Laplacian Noise)	80

Figure		Page
A7	100 Overlaid Burg Realizations (White Laplacian Noise)	81
A8	Average of Burg Realizations (White Laplacian Noise)	82
A9	100 Overlaid RITE Realizations (Impulsive Noise)	83
A10	Average of RITE Realizations (Impulsive Noise)	84
A11	100 Overlaid Burg Realizations (Impulsive Noise)	85
A12	Average of Burg Realizations (Impulsive Noise)	86
A13	100 Overlaid LSMYWE Realizations (WGN)	87
A14	Average of LSMYWE Realizations (WGN)	88
A15	100 Overlaid LSMYWE Realizations (Impulsive Noise)	89
A16	Average of LSMYWE Realizations (Impulsive Noise)	90
A17	100 Overlaid RITE (LSMYWE Initial) Realizations (Impulsive Noise)	91
A18	Average of RITE (LSMYWE Initial) Realizations (Impulsive Noise)	92
B1	The number of good estimations. 10^5 realizations for $N = 100$, $f_0 = 0.1$, $SNR_1 = 1dB$, $SNR_2 = 0dB$, $SNR_3 = -9dB$	96
B2	The number of of outliers. 10^5 realizations for $N = 100$, $f_0 = 0.1$, $SNR_1 = 1dB$, $SNR_2 = 0dB$, $SNR_3 = -9dB$	97
B3	Taylor expansion up to 3rd order	98
B4	Taylor expansion up to 4th order	99
B5	Taylor expansion up to 4th order at the region near $f=0.1$. . .	100
B6	The MSE vs different value of f_0 for $N = 50$, $\phi_1 = 0.314$ (5000 realizations).	101
B7	$N = 50$, $\phi_1 = 0.314$, $f_1 = 0.05$ (5000 realizations).	102
B8	$N = 50$, $\phi_1 = 0.314$, $f_1 = 0.05$ (5000 realizations).	103
B9	$N = 50$, $\phi_1 = 0.314$, $f_1 = 0.05$ (5000 realizations).	104

CHAPTER 1

A Comparison between Robust Information Theoretic Estimator and Asymptotic Maximum Likelihood Estimator for Misspecified Model

by

Xin Zhou and Steven Kay

Dept. of Electrical, Computer and Biomedical Engineering

University of Rhode Island, Kingston, RI, USA

published in International Society for Optics and Photonics Conference, 2018.

Abstract

A robust information-theoretic estimator (RITE) is based on a non-homogeneous Poisson spectral representation. When an autoregressive (AR) Gaussian wide sense stationary (WSS) process is corrupted by noise, RITE is analyzed and shown by simulation to be more robust to noise than the asymptotic maximum likelihood estimator (MLE). The statistics of RITE and asymptotic MLE are analyzed for the misspecified model. For large data records, RITE and MLE are asymptotically normally distributed. MLE has lower variance, but RITE exhibits much less bias. Simulation examples of a noise corrupted AR process are provided to support the theoretical properties and show the advantage of RITE for low signal-to-noise ratios (SNR).

1.1 Introduction

The modeling approach to spectral estimation produces less biased and lower variance spectral estimators if the model chosen is an accurate representation of the power spectral density (PSD). The most popular modeling approach is the AR spectral estimator since it can be found by solving a set of linear equations, the Yule-Walker equations[1]. On the contrary, autoregressive moving average (ARMA) or moving average (MA) estimation requires one to solve a set of nonlinear equations. When the AR modeling assumption is correct, spectral estimators such as the covariance method, Burg method, and recursive MLE method are approximately maximum likelihood estimators (MLE)[2] and attain the Cramer-Rao Lower Bound (CRLB) [3] [4]. However, modeling errors are always present to some extent. This problem is said to be a misspecification, which is difficult to avoid. When the observations are corrupted by noise, the various AR estimators mentioned above are no longer MLEs but quasi-MLEs [5], which produce smoothed AR spectral estimates and are unable to resolve the peaks in the power spectral

density (PSD). Numerous studies indicate that the resolution of estimated AR spectra decreases as the SNR decreases [6] [7] [8]. The sensitivity to noise results in a severe bias, therefore, limiting the utility of AR spectrum estimation. Several suboptimum approaches have been proposed for the misspecified model. One method is to recognize the true model as an ARMA process and use the modified Yule-Walker equations [9]. A second solution is to increase the order of AR model to reduce the bias due to the misspecification. A third approach is to compensate the AR parameters or reflection coefficients for the biasing effect of noise [10]. One more option is filtering of the data with a Wiener filter to enhance the signal. However, none of these existing approaches have met with great success. A realizable information theoretic estimator (RITE) [11] is proposed and is shown by theory and simulation to be more robust to noise than the asymptotic MLE.

Different from the general spectrum representation, which is a sum of sinusoids of fixed frequencies with random amplitudes and random phases, RITE is proposed by modeling the frequencies as random events distributed according to a nonhomogeneous Poisson process. Once this representation is adopted, we can derive its likelihood function. Since the random frequency events are not observable, we propose a realizable approximation to the likelihood function. Both RITE and MLE are obtained as the maximum of likelihood functions. They are special cases of M-estimators, which are introduced and whose asymptotic properties are analyzed by Huber [12]. The properties of a correctly specified model have been well studied. A misspecified model has also been investigated but to a lesser extent [5] [13]. In our study, the detailed statistical theorems for quasi-MLE and RITE are based on the theory of M-estimators. More detailed information about the M-estimator can be found in "The calculus of m-estimators" [14].

The paper is organized as follows. Section 1.2 gives a brief introduction to

RITE. In Section 1.3, the theoretical properties of quasi-MLE and RITE are given. In Section 1.4, simulation examples to verify the theory in Section 1.3 and a simple explanation of the robustness of RITE are provided. Section 1.5 summarizes our results and discusses future work.

1.2 Realizable Information Theoretic Estimator

The background for this section can be found in references [11] [15]. A real discrete-time wide sense stationary (WSS) random process can be written in the spectral form using random frequencies as

$$x[n] = \frac{1}{\sqrt{\lambda_0/2}} \sum_{k=1}^{N_p} A_k \cos(2\pi F_k n + \Phi_k) \quad -\infty < n < \infty$$

The representation can be viewed as a marked Poisson process. In this study, instead of using uniformly spaced frequencies with a fixed number, we take N_p as a nonhomogeneous Poisson random process in frequency with intensity $\lambda(f) = \lambda_0 p(f)$. F_k is the k^{th} point event in the frequency interval $0 \leq f \leq 0.5$ with "marks" (A_k, Φ_k) . A_i 's are independent and identically distributed (IID) positive amplitude random variables. Φ_i 's are uniformly distributed on $[0, 2\pi)$ phase random variables. The amplitude, phase, and frequency random variables are independent of each other.

We normalize the intensity by $\lambda(f) = \lambda_0 p(f)$, where $\int_0^{1/2} p(f) df = 1$. With this normalization, $p(f)$ can be interpreted as a probability density function (PDF). The PSD of $x[n]$ can be shown to be $P(f) = \frac{E(A^2)}{2} p(|f|)$ for $-0.5 \leq f \leq 0.5$ [11]. Here, we are only interested in the case that the total power is 1, i.e. $E(A^2) = 1$. It follows that $\lambda(f) = 2\lambda_0 P(f)$, where $\int_{-\frac{1}{2}}^{\frac{1}{2}} P(f) df = 1$. It can be shown that the part of the log-likelihood that depends on $\lambda(f)$ is [16]

$$l = - \int_0^{\frac{1}{2}} \lambda(f) df + \int_0^{\frac{1}{2}} \ln \lambda(f) N(df)$$

where $N(df)$ is the random variable indicating the number of frequency events

in the interval $(f, f + df)$. Since the frequency events are in general not observable but only $x[n]$ is observed, we proceed by replacing $N(df)$ with a realizable approximation:

$$E(N(df)) = \lambda(f)df = 2\lambda_0 P(f)df \approx 2\lambda_0 \bar{I}(f)df$$

where $\bar{I}(f)$ is the normalized periodogram, which is given by

$$\bar{I}(f) = \frac{I(f)}{\int_{-\frac{1}{2}}^{\frac{1}{2}} I(f)df}$$

and $I(f)$ is the unnormalized periodogram

$$I(f) = \frac{1}{N} \left| \sum_{n=0}^{N-1} x[n] \exp(-j2\pi n) \right|^2$$

In accordance with $\int_{-\frac{1}{2}}^{\frac{1}{2}} P(f)df = 1$, we normalize the periodogram to ensure $\int_{-\frac{1}{2}}^{\frac{1}{2}} \bar{I}(f)df = 1$. We now have

$$\begin{aligned} l' &\approx - \int_0^{\frac{1}{2}} \lambda(f)df + \int_0^{\frac{1}{2}} \ln(\lambda(f))2\lambda_0 \bar{I}(f)df \\ &= -\lambda_0 + 2\lambda_0 \int_0^{\frac{1}{2}} \ln(2\lambda_0 P(f))\bar{I}(f)df \\ &= -\lambda_0 + 2\lambda_0 \ln(2\lambda_0) \int_0^{\frac{1}{2}} \bar{I}(f)df + 2\lambda_0 \int_0^{\frac{1}{2}} \bar{I}(f) \ln P(f)df \\ &= -\lambda_0 + \lambda_0 \ln(2\lambda_0) + 2\lambda_0 \int_0^{\frac{1}{2}} \bar{I}(f) \ln P(f)df \end{aligned}$$

Ignoring the terms that do not depend on the PSD and the scaling parameter λ_0 , we have the realizable likelihood function

$$l_R = \int_{-\frac{1}{2}}^{\frac{1}{2}} \bar{I}(f) \ln P(f)df$$

The function $\int_{-\frac{1}{2}}^{\frac{1}{2}} \bar{I}(f) \ln P(f)df$ achieves its maximum when $P(f)$ is identical to $\bar{I}(f)$. When we assume that the PSD depends on a set of parameters, the estimation of those parameters is chosen to maximize l_R . Note that since

$\bar{I}(f) = \frac{I(f)}{\int_{-\frac{1}{2}}^{\frac{1}{2}} I(f) df} = \frac{I(f)}{\frac{1}{N} \sum_{n=0}^{N-1} x^2[n]}$, the maximization result does not depend on the normalization term $\frac{1}{\frac{1}{N} \sum_{n=0}^{N-1} x^2[n]}$. We finally have the RITE likelihood function as

$$l_R = \int_{-\frac{1}{2}}^{\frac{1}{2}} I(f) \ln P(f) df$$

1.3 The Statistical Properties of quasi-MLE and RITE

Let the real signal $s[n]$ be a wide sense stationary (WSS) Gaussian random process whose power equals 1. Let N be the data record length. The observed data set $\{x[0], x[1], \dots, x[N-1]\}$ is generated from the noise corrupted signal, with PSD function $Q(f; \boldsymbol{\theta}^*)$. We propose $P(f; \boldsymbol{\theta})$ to be the PSD model of the signal, where $\boldsymbol{\theta}$ is a $p \times 1$ vector parameter. Assume $P(f; \boldsymbol{\theta})$ is suitably smooth on $\boldsymbol{\theta}$. In accordance with the fact that signal power equals 1, we constrain $\int_{-\frac{1}{2}}^{\frac{1}{2}} P(f; \boldsymbol{\theta}) df = 1$, or equivalently, the autocorrelation satisfies $r[0] = 1$.

1.3.1 The Statistical Properties of Misspecified MLE

For large data records, the asymptotic Gaussian log likelihood function is [17]

$$l_M = -\frac{N}{2} \ln 2\pi - \frac{N}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\ln P(f; \boldsymbol{\theta}) + \frac{I(f)}{P(f; \boldsymbol{\theta})} \right) df \quad (1)$$

If $E_{\boldsymbol{\theta}^*}(l_M)$ exists, where $E_{\boldsymbol{\theta}^*}$ represents the expected value with respect to the true model, we define $\boldsymbol{\theta}_0$ to be the one that maximizes $E_{\boldsymbol{\theta}^*}(l_M)$.

Theorem 1: The estimator $\hat{\boldsymbol{\theta}}$ that maximizes (4) is asymptotically normally distributed with mean $\boldsymbol{\theta}_0$ and covariance matrix $\mathbf{A}^{-1}(\boldsymbol{\theta}_0)\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}^{-T}(\boldsymbol{\theta}_0)$, i.e.

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}(\boldsymbol{\theta}_0)^{-T})$$

where

$$[\mathbf{A}(\boldsymbol{\theta}_0)]_{ul} = \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{\partial^2 \ln P(f; \boldsymbol{\theta})}{\partial \theta_u \partial \theta_l} \left(1 - \frac{Q(f; \boldsymbol{\theta}^*)}{P(f; \boldsymbol{\theta})} \right) + \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_l} \frac{Q(f; \boldsymbol{\theta}^*)}{P(f; \boldsymbol{\theta})} \right) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} df$$

$$[\mathbf{B}(\boldsymbol{\theta}_0)]_{ul} = \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_l} \frac{Q^2(f; \boldsymbol{\theta}^*)}{P^2(f; \boldsymbol{\theta})} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} df$$

Corollary 1 In the case of a scalar parameter, for large data records, the quasi-MLE $\hat{\theta}$ is asymptotically normally distributed with mean θ_0 and variance σ^2 , or

$$\sqrt{N}(\hat{\theta} - \theta_0) \stackrel{a}{\sim} \mathcal{N}(0, \sigma^2)$$

where

$$\sigma^2 = \frac{2 \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{Q(f; \boldsymbol{\theta}^*)}{P(f; \boldsymbol{\theta})} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta} \right)^2 \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} df}{\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{\partial^2 \ln P(f; \boldsymbol{\theta})}{\partial \theta^2} (1 - \frac{Q(f; \boldsymbol{\theta}^*)}{P(f; \boldsymbol{\theta})}) + \left(\frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta} \right)^2 \frac{Q(f; \boldsymbol{\theta}^*)}{P(f; \boldsymbol{\theta})} \right) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} df \right)^2}$$

1.3.2 The Statistical Properties of RITE

The RITE log likelihood function is

$$l_R = \int_{-\frac{1}{2}}^{\frac{1}{2}} I(f) \ln P(f; \boldsymbol{\theta}) df \quad (2)$$

Assume $E_{\boldsymbol{\theta}^*}(l_R)$ exists, here we define $\boldsymbol{\theta}_0$ to be the one that maximizes $E_{\boldsymbol{\theta}^*}(l_R)$.

Theorem 2: The estimator $\hat{\boldsymbol{\theta}}$ that maximizes (3) is asymptotically normally distributed with mean $\boldsymbol{\theta}_0$ and covariance matrix $\mathbf{A}^{-1}(\boldsymbol{\theta}_0)\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}^{-T}(\boldsymbol{\theta}_0)$, i.e.

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}(\boldsymbol{\theta}_0)^{-T})$$

where

$$[\mathbf{A}(\boldsymbol{\theta}_0)]_{ul} = \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\partial^2 \ln P(f; \boldsymbol{\theta})}{\partial \theta_u \partial \theta_l} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} Q(f; \boldsymbol{\theta}^*) df$$

$$[\mathbf{B}(\boldsymbol{\theta}_0)]_{ul} = \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_l} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} Q^2(f; \boldsymbol{\theta}^*) df$$

Corollary 2 In the case of a scalar parameter, RITE $\hat{\theta}$ is asymptotically normally distributed with mean θ_0 and variance σ^2 , or

$$\sqrt{N}(\hat{\theta} - \theta_0) \stackrel{a}{\sim} \mathcal{N}(0, \sigma^2)$$

where

$$\sigma^2 = \frac{2 \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(Q(f; \theta^*) \frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 \Big|_{\theta=\theta_0} df}{\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} Q(f; \theta^*) \frac{\partial^2 \ln P(f; \theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} df \right)^2}$$

1.4 Simulation Examples

Assume an AR Gaussian process $s[n]$

$$s[n] = - \sum_{k=1}^p a[k] s[n-k] + u[n]$$

where $u[n]$ has variance σ_u^2 , p is the order of the AR process, and $a[k]$ is the k^{th} autoregressive coefficient. The PSD of $s[n]$ is:

$$P(f) = \frac{\sigma_u^2}{|1 + a[1] \exp(-j2\pi f) + \dots + a[p] \exp(-j2\pi fp)|^2}$$

As we stated in Section 1.3, the PSD is restricted to yield a power of 1, which makes σ_u^2 a function of the AR coefficients. If white noise $w[n]$ is present in addition to the signal $s[n]$, then the data is $x[n] = s[n] + w[n]$. The true PSD $Q(f)$ of the observed data becomes

$$Q(f) = P(f) + \sigma_w^2$$

where σ_w^2 is the variance of the observation noise. In our simulations, we assume the AR model order is known. Thus $a[1], a[2], \dots, a[p]$ are the p parameters to be estimated. No analytical expression is available for the RITE estimator.

1.4.1 AR(1) Process Example

The PSD of an AR(1) process with restriction $r[0] = 1$ is

$$P(f) = \frac{1 - a^2[1]}{1 + 2a[1] \cos(2\pi f) + a^2[1]}$$

We generate a Gaussian AR(1) process with $a[1] = -0.9$. We use a grid search on k_1 to find RITE since the reflection coefficient k_1 is limited in $(-1, 1)$ to guarantee

a stable AR process. To be fair, quasi-MLE is also calculated using a grid search. When the observed data is embedded in noise, quasi-MLE and RITE converge to means other than the true value. Thus to compare the performance of the two estimators, we need to evaluate the mean square error (MSE), which equals the variance plus the squared bias. The theoretical variance is computed by using Corollary 1 and Corollary 2, where $\theta^* = -0.9$. By definition, θ_0 is the value that maximizes the expected value of the likelihood function. As illustrated in Figure 6, the MSE of RITE grows slower than the one of quasi-MLE and this advantage increases as the SNR decreases. As shown in Figure 7, although the quasi-MLE has a smaller variance, the bias weakens its performance for low SNR range. Therefore, when the squared bias exceeds the variance, the RITE exhibits more noise robustness than the quasi-MLE.

1.4.2 Noise Sensitivity of Likelihood Function

An important problem in AR spectral estimation is its sensitivity to observation noise. The effect of noise flattens the estimated PSD and reduces the resolution, resulting from the severe bias of the misspecified MLE. However, the RITE is shown to have less bias than the MLE when observation noise is present, and this results in improved resolution. The derivative of expected likelihood function with respect to noise power is an evaluation for noise sensitivity. It can be proved that

$$\left| \frac{\partial E_{\theta^*}(l_M)}{\partial \sigma_w^2} \right| > \left| \frac{\partial E_{\theta^*}(l_R)}{\partial \sigma_w^2} \right|$$

Consider a narrow-band process, for which the k_i 's may be close to 1. The closer they are to 1, the larger is $\left| \frac{\partial E_{\theta^*}(l_R)}{\partial \sigma_w^2} \right|$ and $\left| \frac{\partial E_{\theta^*}(l_M)}{\partial \sigma_w^2} \right|$, and the more effect the noise has. The MLE likelihood function is more severely affected by noise when the AR random process is narrow-band. As an example, the simulation example we

employ here, which is an AR(4) process, for

$$a[1] = -2.7428, a[2] = 3.7906, a[3] = -2.6454, a[4] = 0.93$$

or equivalently,

$$k_1 = -0.71, k_2 = 0.98, k_3 = -0.70, k_4 = 0.93$$

we have that

$$\left| \frac{\partial E_{\theta^*}(l_R)}{\partial \sigma_w^2} \right|_{\theta=\theta^*} = 6.6$$

$$\left| \frac{\partial E_{\theta^*}(l_M)}{\partial \sigma_w^2} \right|_{\theta=\theta^*} = 2.2 \times 10^4$$

This is a sensitivity difference of several orders of magnitude.

Burg and RITE AR Spectral Estimator

For a higher order AR process, we need to find the global maximum of l_R , but an efficient algorithm is still to be found. Instead of performing a grid search which involves a high computation cost, we use the Matlab optimization function *fmincon* with a proper initial point generated by Burg method and constraint $-1 < k_i < 1$ to hopefully find the global maximum solution as the RITE estimator. Once $\hat{k}_1, \hat{k}_2, \dots, \hat{k}_p$ are available, the AR parameters are estimated as follows:

$$\hat{a}[i] = \hat{a}_p[i] \quad \text{for } i = 1, \dots, p$$

For $k = 1$

$$\hat{a}_1[1] = \hat{k}_1$$

For $k = 2, 3, \dots, p$

$$\hat{a}_k[i] = \begin{cases} \hat{a}_{k-1}[i] + \hat{k}_k \hat{a}_{k-1}[k-i] & \text{for } i = 1, 2, \dots, k-1 \\ \hat{k}_k & \text{for } i = k \end{cases}$$

$$\hat{\sigma}_u^2 = \prod_{i=1}^p (1 - \hat{k}_i^2)$$

We denote this approach as the modified Levinson algorithm since we constrain $r[0] = 1$. Once the $\hat{a}[i]$'s and $\hat{\sigma}_u^2$ are obtained, we substitute the estimated AR parameters into the theoretical PSD to obtain the spectral estimate.

The Burg method is approximately an MLE and has been shown to have good resolution for a narrow-band PSD when the data is not noise corrupted [1]. Similar to our approach of computing RITE, the Burg method first estimates reflection coefficients, and then calculates the AR parameters by the Levinson algorithm. Here to compare the Burg method and RITE, instead of using the Levinson algorithm, we use the modified Levinson algorithm described above, to obtain AR parameters and σ_u^2 .

Results are presented in Figures 2, 3, 4 for different SNRs and data record length N . The studies illustrated in the figures demonstrate that the Burg method has less variance than RITE. However, when noise is present, there is a large bias for Burg method. For SNR= 40dB, the resolution of the Burg method is degraded due to noise. When SNR decreases to 35dB, Burg method is unable to resolve the peaks at all. As SNR reduces further, Burg method generate even more flattened PSDs. On the other hand, when SNR reduces from 40dB to 35dB, RITE still provides good resolution although the bias is increased somewhat. Even if the SNR is reduced to 15dB, RITE still produces estimates with good resolution. The overlaid plots show that RITE has more variance than Burg. However, the average plots verify that RITE has less bias and higher resolution, demonstrating that RITE is indeed more robust to noise as compared to the Burg method.

1.5 Discussion and Conclusions

We have introduced RITE as a new method for spectral estimation. The statistical properties of RITE and the quasi-MLE are stated when the data is misspecified. In particular, the misspecification example of additive noise is studied in detail. Theorems have been verified via simulation for a Gaussian AR(1) process. Simulation examples of higher order AR process employed in this study demonstrate that RITE is indeed more robust than the quasi-MLE when noise is present, due to its smaller bias. In this study, for higher order AR examples, we had to rely on an iterative search algorithm to find the global maximum of the RITE likelihood function. A potential problem of using *fmincon* to find RITE is that, it may produce some outlier, which is actually a local maximum not the true RITE. These studies have established a solid foundation to further our goal of searching for the global maximum which is the true RITE, and which may have an even better performance. Therefore, an efficient method of computing RITE will be explored in future work.

List of References

- [1] S. M. Kay, *Modern spectral estimation*. Pearson Education India, 1988.
- [2] R. A. Fisher, “Theory of statistical estimation,” in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 22, no. 5. Cambridge University Press, 1925, pp. 700–725.
- [3] H. Cramér, “Mathematical methods of statistics,” vol. 9, 1999.
- [4] S. Kay and J. Makhoul, “On the statistics of the estimated reflection coefficients of an autoregressive process,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 31, no. 6, pp. 1447–1455, 1983.
- [5] H. White, “Maximum likelihood estimation of misspecified models,” *Econometrica: Journal of the Econometric Society*, pp. 1–25, 1982.
- [6] R. T. Lacoss, “Data adaptive spectral analysis methods,” *Geophysics*, vol. 36, no. 4, pp. 661–675, 1971.

- [7] W. Chen and G. Stegen, “Experiments with maximum entropy power spectra of sinusoids,” *Journal of Geophysical Research*, vol. 79, no. 20, pp. 3019–3022, 1974.
- [8] L. Marple, “Resolution of conventional fourier, autoregressive, and special arma methods of spectrum analysis,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP’77.*, vol. 2. IEEE, 1977, pp. 74–77.
- [9] W. Gersch, “Estimation of the autoregressive parameters of a mixed autoregressive moving-average time series,” *IEEE Transactions on Automatic Control*, vol. 15, no. 5, pp. 583–588, 1970.
- [10] S. Kay, “Noise compensation for autoregressive spectral estimates,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 3, pp. 292–303, 1980.
- [11] S. Kay, “Poisson maximum likelihood spectral inference,” 2017, unpublished <http://www.ele.uri.edu/faculty/kay/New%20web/Books.htm>.
- [12] P. J. Huber *et al.*, “Robust estimation of a location parameter,” *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.
- [13] S. Fortunati, F. Gini, M. S. Greco, and C. D. Richmond, “Performance bounds for parameter estimation under misspecified models: Fundamental findings and applications,” *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 142–157, 2017.
- [14] L. A. Stefanski and D. D. Boos, “The calculus of m-estimation,” *The American Statistician*, vol. 56, no. 1, pp. 29–38, 2002.
- [15] J. F. C. Kingman, *Poisson processes*. Wiley Online Library, 1993.
- [16] D. L. Snyder and M. I. Miller, *Random point processes in time and space*. Springer Science & Business Media, 2012.
- [17] S. M. Kay, “Fundamentals of statistical signal processing, volume i: Estimation theory (v. 1),” *PTR Prentice-Hall, Englewood Cliffs*, 1993.

(a) Mean Square Error of $\hat{a}[1]$

(b) Variance and Squared Bias of $\hat{a}[1]$

Figure 1: AR(1) process, $a[1]=-0.9$, $N=200$

(a) 100 Overlaid Realizations.

(b) Average of Realizations.

(c) 100 Overlaid Realizations.

(d) Average of Realizations.

Figure 2: SNR=40dB, N=350

(a) 100 Overlaid Realizations.

(b) Average of Realizations.

(c) 100 Overlaid Realizations.

(d) Average of Realizations.

Figure 3: SNR=35dB, N=350

(a) 100 Overlaid Realizations.

(b) Average of Realizations.

(c) 100 Overlaid Realizations.

(d) Average of Realizations.

Figure 4: SNR=15dB, N=450

CHAPTER 2

A Robust Spectral Estimator with Application to a Noise Corrupted Process

by

Xin Zhou and Steven Kay

Dept. of Electrical, Computer and Biomedical Engineering

University of Rhode Island, Kingston, RI, USA

published in IEEE Transactions on Signal Processing, 2019.

Abstract

When a data set is corrupted by noise, the model for the data generating process is misspecified and can cause parameter estimation problems. As an example, in the case of a Gaussian autoregressive (AR) process corrupted by noise, the data is more accurately modeled as an autoregressive moving average (ARMA) process rather than an AR process. This misspecification leads to bias, and hence, low resolution in AR spectral estimation. However, a new parametric spectral estimator, the realizable information theoretic estimator (RITE) based on a nonhomogeneous Poisson spectral representation, is shown by simulation to be more robust to white noise than the asymptotic maximum likelihood estimator (MLE). We therefore conducted an in-depth investigation and analyzed the statistics of RITE and the asymptotic MLE for the misspecified model. For large data records, RITE and the asymptotic MLE are both asymptotically normally distributed. The asymptotic MLE has a slightly lower variance, but RITE exhibits much less bias. Simulation examples of a white noise corrupted AR process are provided to support the theoretical properties. This advantage of RITE increases as the signal-to-noise-ratio (SNR) decreases.

2.1 Introduction

The spectral representation for a wide sense stationary (WSS) random process relies on the time representation which is a sum of sinusoids with fixed frequencies, random phases and random amplitudes [1]. It forms the basis for spectral estimation. Another less well known representation models the frequencies as random point events distributed according to a nonhomogeneous Poisson process. The likelihood function can be derived for this spectral representation. Since the frequency events are usually not observable, some modifications are applied to the likelihood function. The estimator that maximizes the approximated likelihood function is

called the realizable information theoretic estimator (RITE) [2]. It can be used in model-based spectral estimation.

The autoregressive (AR) model is widely used in spectral estimation. The maximum likelihood estimator (MLE) is usually employed for a good estimate of the AR parameters. If we assume a real Gaussian random process, the autocorrelation method, which requires solving the Yule-Walker equations with a suitable autocorrelation function (ACF), can be found efficiently and is equivalent to the approximate MLE [3]. Many other methods for AR parameters estimation that produce the same numerical values for large data records, like the Burg method and the covariance method, are also approximate MLEs [4]. Hence those methods share the desirable properties of the MLE that for large data records they are consistent, asymptotically Gaussian, unbiased and attain the Cramer-Rao lower bound (CRLB) [5] [6]. However, when the observations are corrupted by additive noise, the various methods for AR parameters estimation mentioned above produce severe biases. This sensitivity to the noise addition results in a smoothed AR spectral estimate. Numerous studies indicate that the resolution of estimated AR spectra decreases as the signal-to-noise-ratio (SNR) decreases [7] [8] [9]. This is because the additive noise changes the true model to an autoregressive moving average (ARMA) where AR and moving average (MA) parameters are linked, instead of an AR. Hence the methods above are no longer the true MLEs. To get better resolution, one option is to use an ARMA model estimated by the least squares modified Yule-Walker equations (LSMYWE) [10], but the model order of the MA part depends on the noise type (see supplementary material A.5), and therefore limits its utility in practice.

The MLE for a misspecified model is called a quasi-MLE [11]. A misspecified model has been investigated but to a lesser extent in [11] [12]. Hence more

analysis on misspecification is necessary. Compared with the asymptotic Gaussian MLE, RITE shows a robustness to white noise in PSD classification problems [2]. Therefore it would be of interest to investigate how RITE performs in AR spectral estimation.

In this paper, the asymptotic statistical properties of the quasi-MLE and RITE are derived and verified by simulation examples. Both estimators are asymptotically Gaussian distributed but with different means and covariance matrices. An application to spectral estimation using the AR model is provided in the paper. For an AR PSD, we prove that the asymptotic Gaussian likelihood function is more sensitive to white noise than the RITE likelihood function. Our experiments show that in comparison to the quasi-MLE, RITE has smaller bias when white noise is present in AR process.

The paper is organized as follows. Section 2.2 gives a brief introduction to RITE. In Section 2.3, the theoretical properties of MLE and RITE are given. Section 2.4 reviews the AR model and gives a simple explanation of the white noise robustness of RITE. In Section 2.5, spectral estimation application using AR model are provided to verify the theory in Section 2.3 and to show the robustness of RITE for white noise corrupted data. Section 2.6 summarizes our results and discusses future work.

2.2 Realizable Information Theoretic Estimator

The background for this section can be found in [2] and [13]. A real discrete-time WSS random process can be represented in the spectral form as a sum of sinusoids with random frequencies, amplitudes and phases:

$$x[n] = \frac{1}{\sqrt{\lambda_0/2}} \sum_{k=1}^M A_k \cos(2\pi F_k n + \Phi_k) \quad -\infty < n < \infty$$

A similar representation that uses two independent Poisson point processes can be found in [14]. The representation herein can be viewed as a marked Poisson

process. If the number of events M is fixed, then the model reduces to that in [15]. In this study we take M as the number of events of a nonhomogeneous Poisson random process in frequency with intensity $\lambda(f)$ on the interval $[0, 0.5]$. F_k is the k^{th} point event on the frequency interval $0 \leq f \leq 0.5$ with "marks" (A_k, Φ_k) . $A_1, A_2, \dots, A_M$ are independent and identically distributed (IID) positive amplitude random variables. $\Phi_1, \Phi_2, \dots, \Phi_M$ are phase random variables uniformly IID on $[0, 2\pi)$. The amplitude, phase, and frequency random variables are independent of each other.

We normalize the intensity by $\lambda(f) = \lambda_0 p(f)$. With this normalization, the integral of $p(f)$ over $[0, 0.5]$ is equal to 1. This property allows $p(f)$ to be interpreted as a probability density function (PDF) on $0 \leq f \leq 0.5$. The power spectral density (PSD) of $x[n]$ can be shown to be $P(f) = \frac{E(A^2)}{2} p(|f|)$ on $-0.5 \leq f \leq 0.5$ [2]. Here, we are only interested in the case that the total power is 1, i.e. $E(A^2) = 1$ or equivalently $\int_{-\frac{1}{2}}^{\frac{1}{2}} P(f) df = 1$. From the above relations and conditions, we can write the intensity function in terms of the PSD function as $\lambda(f) = 2\lambda_0 P(f)$ on $0 \leq f \leq 0.5$, and use $P(-f) = P(f)$.

It can be shown that the part of the log-likelihood that depends on $\lambda(f)$ is

$$l = - \int_0^{\frac{1}{2}} \lambda(f) df + \int_0^{\frac{1}{2}} \ln \lambda(f) N(df)$$

with $N(df)$ is the random variable indicating the number of frequency events on the interval $[f, f + df)$. Since we cannot observe the frequency events but only $x[n]$ in general, we proceed by replacing $N(df)$ with its approximate mean:

$$E(N(df)) = \lambda(f) df = 2\lambda_0 P(f) df \approx 2\lambda_0 \bar{I}(f) df$$

where $\bar{I}(f)$ is the normalized periodogram, which is given by

$$\bar{I}(f) = \frac{I(f)}{\int_{-\frac{1}{2}}^{\frac{1}{2}} I(f) df}$$

and $I(f)$ is the unnormalized periodogram

$$I(f) = \frac{1}{N} \left| \sum_{n=0}^{N-1} x[n] \exp(-j2\pi n f) \right|^2$$

In accordance with $\int_{-\frac{1}{2}}^{\frac{1}{2}} P(f) df = 1$, the periodogram is normalized to ensure $\int_{-\frac{1}{2}}^{\frac{1}{2}} \bar{I}(f) df = 1$. We now have the approximated likelihood function:

$$\begin{aligned} l' &\approx - \int_0^{\frac{1}{2}} \lambda(f) df + \int_0^{\frac{1}{2}} \ln(\lambda(f)) 2\lambda_0 \bar{I}(f) df \\ &= -\lambda_0 + 2\lambda_0 \int_0^{\frac{1}{2}} \ln(2\lambda_0 P(f)) \bar{I}(f) df \\ &= -\lambda_0 + 2\lambda_0 \ln(2\lambda_0) \int_0^{\frac{1}{2}} \bar{I}(f) df + 2\lambda_0 \int_0^{\frac{1}{2}} \bar{I}(f) \ln P(f) df \\ &= -\lambda_0 + \lambda_0 \ln(2\lambda_0) + 2\lambda_0 \int_0^{\frac{1}{2}} \bar{I}(f) \ln P(f) df \end{aligned}$$

Ignoring the terms that do not depend on the PSD and the scaling λ_0 , we have the realizable likelihood function

$$l_R = \int_{-\frac{1}{2}}^{\frac{1}{2}} \bar{I}(f) \ln P(f) df$$

The function $\int_{-\frac{1}{2}}^{\frac{1}{2}} \bar{I}(f) \ln P(f) df$ achieves its maximum when $P(f)$ is identical to $\bar{I}(f)$ (*This is proved in Appendix A, but not included in the original paper*). If we assume that the PSD depends on a set of parameters, then the estimation of those parameters is chosen to maximize l_R . Note that $\int_{-\frac{1}{2}}^{\frac{1}{2}} I(f) df = \frac{1}{N} \sum_{n=0}^{N-1} x^2[n]$, it follows that

$$\begin{aligned} l_R &= \frac{1}{\int_{-\frac{1}{2}}^{\frac{1}{2}} I(f) df} \int_{-\frac{1}{2}}^{\frac{1}{2}} I(f) \ln P(f) df \\ &= \frac{1}{\frac{1}{N} \sum_{n=0}^{N-1} x^2[n]} \int_{-\frac{1}{2}}^{\frac{1}{2}} I(f) \ln P(f) df \end{aligned}$$

Since the maximization result does not depend on the normalization term

$\frac{1}{\frac{1}{N} \sum_{n=0}^{N-1} x^2[n]}$, we finally have the RITE likelihood function as

$$l_R = \int_{-\frac{1}{2}}^{\frac{1}{2}} I(f) \ln P(f) df \quad (3)$$

2.3 The Statistical Properties of MLE and RITE

The MLE and RITE are both obtained by maximizing their likelihood functions. They are both special cases of M-estimators. Huber introduced M-estimators and analyzed their asymptotic properties [16]. The derivation of the statistical properties for MLE and RITE are based on the theory of M-estimators. More detailed information about the M-estimator can be found in [17].

Let the signal $s[n]$ be a wide sense stationary (WSS) Gaussian random process whose power equals 1. Let $\{x[0], x[1], \dots, x[N-1]\}$ be an observed data set generated from the noise corrupted signal, with PSD function $Q(f; \boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ is the true value of a $q \times 1$ vector parameter. We propose $P(f; \boldsymbol{\theta})$ to be the PSD model of the signal, where $\boldsymbol{\theta}$ is a $p \times 1$ vector parameter. Assume $P(f; \boldsymbol{\theta})$ is suitably smooth on $\boldsymbol{\theta}$, i.e., $P(f; \boldsymbol{\theta})$ has continuous derivatives with respect to $\boldsymbol{\theta}$ up to some desired order. In accordance with the fact that signal power equals 1, we constrain $\int_{-\frac{1}{2}}^{\frac{1}{2}} P(f; \boldsymbol{\theta}) df = 1$, or equivalently, the autocorrelation satisfies $r[0] = 1$.

2.3.1 The Statistical Properties of Misspecified MLE

For large data records, the asymptotic Gaussian log likelihood function is [18]

$$l_M = -\frac{N}{2} \ln 2\pi - \frac{N}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\ln P(f; \boldsymbol{\theta}) + \frac{I(f)}{P(f; \boldsymbol{\theta})} \right) df \quad (4)$$

If $E_{\boldsymbol{\theta}^*}(l_M)$ exists, where $E_{\boldsymbol{\theta}^*}$ represents the expected value with respect to the true model, then we define $\boldsymbol{\theta}_0$ to be the one that maximizes $E_{\boldsymbol{\theta}^*}(l_M)$. The following theorem and corollaries are valid under the assumption that $\{\boldsymbol{c} \cdot \frac{\partial l_M}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\}$ satisfies the Lyapunov condition for any $p \times 1$ vector $\boldsymbol{c}$ at any frequency. The theorem below applies more generally to a vector parameter for misspecified problems. The corollaries are simplified for a scalar parameter and the correct model.

Theorem 1. *The estimator $\hat{\boldsymbol{\theta}}$ that maximizes (4) is asymptotically normally distributed with mean $\boldsymbol{\theta}_0$ and covariance matrix $\boldsymbol{A}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{B}(\boldsymbol{\theta}_0)\boldsymbol{A}^{-T}(\boldsymbol{\theta}_0)$, i.e.,*

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{A}(\boldsymbol{\theta}_0)^{-1} \mathbf{B}(\boldsymbol{\theta}_0) \mathbf{A}(\boldsymbol{\theta}_0)^{-T})$$

where

$$\begin{aligned} [\mathbf{A}(\boldsymbol{\theta}_0)]_{ul} &= \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{\partial^2 \ln P(f; \boldsymbol{\theta})}{\partial \theta_u \partial \theta_l} \left(1 - \frac{Q(f; \boldsymbol{\theta}^*)}{P(f; \boldsymbol{\theta})} \right) \right. \\ &\quad \left. + \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_l} \frac{Q(f; \boldsymbol{\theta}^*)}{P(f; \boldsymbol{\theta})} \right) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} df \\ [\mathbf{B}(\boldsymbol{\theta}_0)]_{ul} &= \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_l} \frac{Q^2(f; \boldsymbol{\theta}^*)}{P^2(f; \boldsymbol{\theta})} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} df \end{aligned}$$

$[\cdot]_{ul}$ denotes the elements at row u column l . The proof is given in supplementary material A.1.

Corollary 1.1. *If the proposed model is the correct one, then $Q(f; \boldsymbol{\theta}^*) = P(f; \boldsymbol{\theta}_0)$.*

It follows that

$$\mathbf{A}(\boldsymbol{\theta}^*) = \mathbf{B}(\boldsymbol{\theta}^*)$$

and

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \stackrel{a}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{A}(\boldsymbol{\theta}^*)^{-1})$$

where

$$[\mathbf{A}(\boldsymbol{\theta}^*)]_{ul} = \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_l} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} df$$

This result agrees with the asymptotic CRLB [5] [19] and implies that for large data records the MLE estimator is the one that has the minimum variance among all estimators.

Corollary 1.2. *In the case of a scalar parameter, the quasi-MLE $\hat{\theta}$ is asymptotically normally distributed with mean θ_0 and variance σ^2 , i.e.,*

$$\sqrt{N}(\hat{\theta} - \theta_0) \stackrel{a}{\sim} \mathcal{N}(0, \sigma^2)$$

where

$$\sigma^2 = \frac{2 \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{Q(f; \theta^*)}{P(f; \theta)} \frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 \big|_{\theta=\theta_0} df}{\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{\partial^2 \ln P(f; \theta)}{\partial \theta^2} \left(1 - \frac{Q(f; \theta^*)}{P(f; \theta)}\right) + \left(\frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 \frac{Q(f; \theta^*)}{P(f; \theta)} \right) \big|_{\theta=\theta_0} df \right)^2}$$

Corollary 1.3. *In the case of a scalar parameter, if the proposed model is correct, so that $Q(f; \theta^*) = P(f; \theta_0)$, then the MLE $\hat{\theta}$ is asymptotically normally distributed with mean θ^* and variance σ^2 , i.e.,*

$$\sqrt{N}(\hat{\theta} - \theta^*) \stackrel{a}{\sim} \mathcal{N}(0, \sigma^2)$$

where

$$\sigma^2 = \frac{2}{\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 \big|_{\theta=\theta^*} df}$$

2.3.2 The Statistical Properties of RITE

The RITE likelihood function l_R is given in (3). Assume $E_{\theta^*}(l_R)$ exists, here we define θ_0 to be the one that maximizes $E_{\theta^*}(l_R)$. The following theorems are valid under the assumption that $\{\mathbf{c} \cdot \frac{\partial l_R}{\partial \theta} \big|_{\theta=\theta_0}\}$ satisfies the Lyapunov condition for any $p \times 1$ vector $\mathbf{c}$ at any frequency.

Theorem 2. *The estimator $\hat{\theta}$ that maximizes (3) is asymptotically normally distributed with mean θ_0 and covariance matrix $\mathbf{A}^{-1}(\theta_0)\mathbf{B}(\theta_0)\mathbf{A}^{-T}(\theta_0)$, i.e.,*

$$\sqrt{N}(\hat{\theta} - \theta_0) \stackrel{a}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{A}(\theta_0)^{-1}\mathbf{B}(\theta_0)\mathbf{A}(\theta_0)^{-T})$$

where

$$[\mathbf{A}(\theta_0)]_{ul} = \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\partial^2 \ln P(f; \theta)}{\partial \theta_u \partial \theta_l} \big|_{\theta=\theta_0} Q(f; \theta^*) df$$

$$[\mathbf{B}(\theta_0)]_{ul} = \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\partial \ln P(f; \theta)}{\partial \theta_u} \frac{\partial \ln P(f; \theta)}{\partial \theta_l} \big|_{\theta=\theta_0} Q^2(f; \theta^*) df$$

The proof is given in supplementary material A.2. When model is correct, unlike the MLE case, the equality $\mathbf{A}(\boldsymbol{\theta}_0) = \mathbf{B}(\boldsymbol{\theta}_0)$ does not hold. Therefore, the expressions for a correct model cannot be simplified.

Corollary 2.1. *In the case of a scalar parameter, RITE $\hat{\theta}$ is asymptotically normally distributed with mean θ_0 and variance σ^2 , i.e.,*

$$\sqrt{N}(\hat{\theta} - \theta_0) \stackrel{a}{\sim} \mathcal{N}(0, \sigma^2)$$

where

$$\sigma^2 = \frac{2 \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(Q(f; \theta^*) \frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 |_{\theta=\theta_0} df}{\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} Q(f; \theta^*) \frac{\partial^2 \ln P(f; \theta)}{\partial \theta^2} |_{\theta=\theta_0} df \right)^2}$$

2.4 Spectral Estimation Application with AR Model

We assume an AR Gaussian process $s[n]$

$$s[n] = - \sum_{k=1}^p a[k] s[n-k] + u[n]$$

where $u[n]$ is the driving noise of the model with variance σ_u^2 , p is the order of the AR process, and $a[k]$ is the k^{th} AR coefficient. The AR PSD $P(f)$ is [6]:

$$\frac{\sigma_u^2}{|1 + a[1] \exp(-j2\pi f) + \dots + a[p] \exp(-j2\pi fp)|^2} \quad (5)$$

As we stated in Section 2.3, the PSD is restricted to yield a power of 1, so σ_u^2 is not actually a parameter, but a function of $a[1], a[2], \dots, a[p]$.

Alternatively, an AR process can be expressed by $r[0]$, which in our case equals 1, and the reflection coefficients $k_1, k_2, \dots, k_p$, which are restricted in $(-1, 1)$ to guarantee a stable process. The Levinson algorithm transfers the reflection coefficients to the AR parameters [6]. It recursively computes the parameter sets $\{a_1[1], \rho_1\}, \{a_2[1], a_2[2], \rho_2\}, \dots, \{a_p[1], a_p[2], \dots, a_p[p], \rho_p\}$. The reflection coefficients are given by $k_i = a_i[i]$. In the final set, $a_p[i]$'s are the AR parameters $a[i]$'s,

and ρ_p is σ_u^2 . The algorithm is initialized by:

$$a_1[1] = -\frac{r[1]}{r[0]}$$

$$\rho_1 = (1 - a_1^2[1])r[0]$$

The recursion for $j = 2, 3, \dots, p$ is

$$a_j[i] = \begin{cases} a_{j-1}[i] + a_j[j]a_{j-1}[j-i] & \text{for } i = 1, 2, \dots, j-1 \\ -\frac{r[j] + \sum_{l=1}^{j-1} a_{j-1}[l]r[j-l]}{\rho_{j-1}} & \text{for } i = j \end{cases}$$

$$\rho_j = (1 - a_j^2[j])\rho_{j-1}$$

In our case, since $r[0] = 1$, the initial set should be:

$$a_1[1] = -r[1]$$

$$\rho_1 = 1 - a_1^2[1]$$

We denote the Levinson recursion with the above initial set as the modified Levinson algorithm. Note that the general Levinson algorithm has $\sigma_u^2 = r[0] \prod_{i=1}^p (1 - k_i^2)$ while here it is $\sigma_u^2 = \prod_{i=1}^p (1 - k_i^2)$.

2.4.1 White Noise Sensitivity of Likelihood Function

Many existing AR spectral estimators (Burg method, covariance method, etc.) are approximate MLEs. They are unbiased and have minimum variances *if there is no modeling error*. An important problem is their sensitivity to observation noise. The effect of noise flattens the estimated PSD and reduces the resolution. This is due to the severe bias of the misspecified, i.e., quasi-MLE. However, RITE is shown to have less bias than the quasi-MLE when white noise is present, and this results in improved resolution.

If white noise $w[n]$ is present in addition to the signal $s[n]$, then the data is $x[n] = s[n] + w[n]$. We assume that $w[n]$ is independent of $s[n]$. The true PSD $Q(f)$ of the observed data becomes

$$Q(f) = P(f) + \sigma_w^2$$

where σ_w^2 is the variance of the observation noise. To analyze how robust the estimator is, we could try analyzing how the white noise affects the likelihood function. By taking the expected value of the likelihood function of RITE, we get from (3)

$$\begin{aligned} E_{\theta^*}(l_R) &= \int_{-\frac{1}{2}}^{\frac{1}{2}} (P(f; \theta^*) + \sigma_w^2) \ln P(f; \theta) df \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} P(f; \theta^*) \ln P(f; \theta) df + \int_{-\frac{1}{2}}^{\frac{1}{2}} \sigma_w^2 \ln \frac{\sigma_u^2}{|A(f)|^2} df \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} P(f; \theta^*) \ln P(f; \theta) df + \int_{-\frac{1}{2}}^{\frac{1}{2}} \sigma_w^2 \ln \sigma_u^2 df - \int_{-\frac{1}{2}}^{\frac{1}{2}} \sigma_w^2 \ln |A(f)|^2 df \end{aligned}$$

where $\theta = [a[1], a[2], \dots, a[p]]^T$ and $A(f)$ is the Fourier transform of $[1, a[1], \dots, a[p]]$. Since a stable AR process has all its poles inside the unit circle, $A(f)$ is minimum phase [6] and

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln |A(f)|^2 df = 0$$

which leads to

$$\begin{aligned} E_{\theta^*}(l_R) &= \int_{-\frac{1}{2}}^{\frac{1}{2}} P(f; \theta^*) \ln P(f; \theta) df + \int_{-\frac{1}{2}}^{\frac{1}{2}} \sigma_w^2 \ln \sigma_u^2 df \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} P(f; \theta^*) \ln P(f; \theta) df + \sigma_w^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} \ln \left(\prod_{i=1}^p (1 - k_i^2) \right) df \end{aligned}$$

Here k_i is the true reflection coefficient. Taking the derivative with respect to σ_w^2 , we get

$$\frac{\partial E_{\theta^*}(l_R)}{\partial \sigma_w^2} = \ln \left(\prod_{i=1}^p (1 - k_i^2) \right)$$

If we do the same operations to l_M , we have from (4)

$$\begin{aligned}
E_{\theta^*}(l_M) &= - \int_{-\frac{1}{2}}^{\frac{1}{2}} \ln P(f; \theta) + \frac{P(f; \theta^*) + \sigma_w^2}{P(f; \theta)} df \\
&= - \int_{-\frac{1}{2}}^{\frac{1}{2}} \ln P(f; \theta) + \frac{P(f; \theta^*)}{P(f; \theta)} df - \sigma_w^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{P(f; \theta)} df \\
&= - \int_{-\frac{1}{2}}^{\frac{1}{2}} \ln P(f; \theta) + \frac{P(f; \theta^*)}{P(f; \theta)} df - \frac{\sigma_w^2}{\sigma_u^2} \int_{-\frac{1}{2}}^{\frac{1}{2}} |A(f)|^2 df
\end{aligned}$$

Hence,

$$\frac{\partial E_{\theta^*}(l_M)}{\partial \sigma_w^2} = -\frac{1}{\sigma_u^2} \int_{-\frac{1}{2}}^{\frac{1}{2}} |A(f)|^2 df$$

By Parseval's theorem, $\int_{-\frac{1}{2}}^{\frac{1}{2}} |A(f)|^2 df = 1 + \sum_{i=1}^p a^2[i]$. Thus,

$$\frac{\partial E_{\theta^*}(l_M)}{\partial \sigma_w^2} = -\frac{1}{\sigma_u^2} \left(1 + \sum_{i=1}^p a^2[i] \right) < -\frac{1}{\sigma_u^2} = -\prod_{i=1}^p \frac{1}{1 - k_i^2}$$

Since $\prod_{i=1}^p \frac{1}{1 - k_i^2} > 1$, we have

$$\begin{aligned}
\left| \frac{\partial E_{\theta^*}(l_M)}{\partial \sigma_w^2} \right| &> \prod_{i=1}^p \frac{1}{1 - k_i^2} \\
&> \ln \left(\prod_{i=1}^p \frac{1}{1 - k_i^2} \right) = \left| \ln \prod_{i=1}^p (1 - k_i^2) \right| = \left| \frac{\partial E_{\theta^*}(l_R)}{\partial \sigma_w^2} \right|
\end{aligned}$$

Consider a narrow-band process, for which the k_i 's may be close to 1. The closer they are to 1, the larger are $\left| \frac{\partial E_{\theta^*}(l_R)}{\partial \sigma_w^2} \right|$ and $\left| \frac{\partial E_{\theta^*}(l_M)}{\partial \sigma_w^2} \right|$, and the more effect the noise has. The MLE likelihood function is more severely affected by noise when the AR random process is narrow-band. As an example, take the AR(4) process that we employed in the next section, for

$$[k_1, k_2, k_3, k_4] = [-0.71, 0.98, -0.70, 0.93]$$

we have that

$$\begin{aligned}
\left| \frac{\partial E_{\theta^*}(l_R)}{\partial \sigma_w^2} \right|_{\theta=\theta^*} &= 6.6 \\
\left| \frac{\partial E_{\theta^*}(l_M)}{\partial \sigma_w^2} \right|_{\theta=\theta^*} &= 2.2 \times 10^4
\end{aligned}$$

This is a sensitivity difference of several orders of magnitude.

2.5 Simulation Examples

We consider spectral estimation of an AR process in noise to test our estimator. White Gaussian noise (WGN), white mixture Gaussian noise and white Laplacian noise give similar results for both RITE and asymptotic MLE. In the case of IID impulsive noise modeled by α -Stable noise, RITE does not perform as well as in white noise case, but still outperforms the asymptotic MLE. In this section therefore, we present only the results for WGN. The other simulations are included in supplemental material A.4. In the simulation, we assume the AR model order is known. Thus $a[1], a[2], \dots, a[p]$ are the p parameters to be estimated. Alternatively, we can estimate the reflection coefficients $k_1, k_2, \dots, k_p$.

No analytical solution is available for RITE with AR model. To find the global maximum of l_R , one option is a grid search. Since the reflection coefficients are guaranteed to give a valid AR process, we do the search over the reflection coefficients. The estimation procedure is:

- a) Create a p -dimensional grid with each dimension in the range $(-1, 1)$.
- b) Assign a value from the gridded domain to the reflection coefficients.
- c) Transform the reflection coefficients to the AR parameters by using the modified Levinson algorithm.
- d) Plug the AR parameters into (5) to get the PSD.
- e) Plug the PSD into (3) and get the value of l_R .
- f) Repeat b) to e) over the valid grid and find the one that maximizes l_R .
- g) The final estimated PSD is obtained by repeating c) to e) with the solution of the reflection coefficients found in f).

If the grid is fine enough, then the solution of a grid search should be very close to the global maximum. This method is recommended when p is small. However, as p increases, a grid search suffers the curse of dimensionality. Hence we

recommend an alternative option, using the Matlab optimization toolbox function *fmincon*, which is a gradient-based method which finds the local minimum of an objective function with constraints. In our procedure, we

- a) Create a function that transfers the reflection coefficients to the objective function $-l_R$.
- b) Use a standard AR spectral estimator (Burg method, covariance method, etc.), to get an estimate of the reflection coefficients and assign it as the initial value.
- c) Given the above function, the constraints $-1 < k_i < 1$, and the proper initial value, *fmincon* outputs the solution of a *local minimum* near the initial value.
- d) Transfer the solution of the reflection coefficients found in c) to the estimated PSD (like the step g in grid search procedure).

For a higher order AR process, this method is more efficient, but it only gives the local minimum. Hopefully, with a proper initial value, this local solution will also yield the true global solution.

Next we use a grid search for an AR(1) example and *fmincon* for an AR(4) example.

2.5.1 AR(1) Process Example

The PSD of an AR(1) process is

$$P(f) = \frac{\sigma_u^2}{1 + 2a[1] \cos(2\pi f) + a^2[1]}$$

Note that our restriction $r[0] = 1$ leads to $\sigma_u^2 = 1 - a^2[1]$. Hence

$$P(f) = \frac{1 - a^2[1]}{1 + 2a[1] \cos(2\pi f) + a^2[1]}$$

We generated a Gaussian AR(1) process with $a[1] = -0.9$. If the observed process is not corrupted by noise, then the MLE (stands for the asymptotic MLE in this section) and RITE are unbiased estimators but with different variances. In this case, MLE is the optimal estimator since it has a smaller variance and attains the

CRLB. It is proved that the RITE variance is larger than the MLE variance as shown in supplementary material A.3. By Corollary 1.3 and Corollary 2.1, the MLE has variance:

$$\frac{1}{\frac{N}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 |_{\theta=\theta_0} df}$$

and the RITE variance is:

$$\frac{\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(P(f; \theta^*) \frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 |_{\theta=\theta_0} df}{\frac{N}{2} \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} P(f; \theta^*) \frac{\partial^2 \ln P(f; \theta)}{\partial \theta^2} |_{\theta=\theta_0} df \right)^2}$$

where

$$\theta = a[1]$$

$$\theta_0 = \theta^* = -0.9$$

$$\frac{\partial \ln P(f; \theta)}{\partial \theta} = -\frac{2\theta}{1 - \theta^2} - \frac{2\theta + 2 \cos(2\pi f)}{1 + \theta^2 + 2\theta \cos(2\pi f)} \quad (6)$$

$$\begin{aligned} \frac{\partial^2 \ln P(f; \theta)}{\partial \theta^2} = & -\frac{1}{(1 - \theta)^2} - \frac{1}{(1 + \theta)^2} \\ & - \frac{2}{1 + \theta^2 + 2\theta \cos(2\pi f)} + \frac{4 \cos(2\pi f) + 4\theta}{(1 + \theta^2 + 2\theta \cos(2\pi f))^2} \end{aligned} \quad (7)$$

To be fair, RITE and MLE are both calculated using a fine grid search. The theoretical $N \times \text{variance}$ vs N values are plotted in solid lines as Fig. 5. Simulated results are shown as circles.

When the observed data is embedded in WGN, MLE and RITE converge to means other than the true value. Thus to compare the performance of the two estimators, we need to evaluate the mean square error (MSE), which equals the variance plus the squared bias. The theoretical variance is computed by using Corollary 1.2 and Corollary 2.1, where $\theta^* = -0.9$. By definition, θ_0 is the value that maximizes the expected value of the likelihood function. $\frac{\partial \ln P(f; \theta)}{\partial \theta}$, $\frac{\partial^2 \ln P(f; \theta)}{\partial \theta^2}$ are listed in (6), (7), and $Q(f; \theta^*) = P(f; \theta^*) + \sigma_w^2$. As illustrated in Fig. 6, the MSE of RITE grows slower than the one of the quasi-MLE and this advantage

increases as the SNR decreases. As shown in Fig. 7, although the quasi-MLE has a smaller variance, the bias weakens its performance for low SNR range. Therefore, when the squared bias exceeds the variance, RITE exhibits more noise robustness than the quasi-MLE. This example verifies the Theorems 1 and 2, at least for an AR(1) process in WGN.

Figure 5: AR(1) Example: $N \times \text{Variance}$

2.5.2 AR(4) Process Example Burg and RITE AR Spectral Estimator

The Burg method is approximately an MLE and therefore is asymptotically unbiased with variance that attains the CRLB. It has been shown to have good resolution for a narrow-band PSD when the data is not noise corrupted [6]. Similar to our approach of computing RITE, the Burg method first estimates reflection coefficients, and then calculates the AR parameters by the Levinson algorithm. For a fair comparison between Burg method and RITE, instead of using the Levinson algorithm, we use the modified Levinson algorithm, described in Section 2.4, to

Figure 6: AR(1) Example: Mean Square Error, N=50000

obtain the Burg estimation. For RITE estimation, instead of performing a grid search, which requires a high computational cost, we use the Matlab *fmincon* function with the reflection coefficients estimated by the Burg method as our initial point to hopefully find the global maximum solution.

Simulations

The AR(4) parameters are set to be

$$[a[1], a[2], a[3], a[4]] = [-2.7428, 3.7906, -2.6454, 0.93]$$

or equivalently,

$$[k_1, k_2, k_3, k_4] = [-0.71, 0.98, -0.70, 0.93]$$

The data length N is 350. Results are presented in Figs. 8 to 11. The studies illustrated in the figures demonstrate that the Burg method has less variance than

Figure 7: AR(1) Example: Variance and Squared Bias, N=50000

RITE, but when WGN is present, there is a large bias. For SNR= 40dB, the resolution of the Burg method (Figs. 10(a), 11(a)) is degraded due to noise, while RITE (Figs. 8 (a), 9(a)) has very good resolution. When SNR= 35, 15dB, the Burg method (Figs. 10(b), 10(c), 11(b), 11(c)) is unable to resolve the two peaks. As shown in the Figs. 8 (b) and 9(b), RITE is not clearly affected as SNR decreases to 35dB. Even if the SNR is reduced to 15dB (Figs. 8 (c), 9(c)), RITE still produces some estimates with good resolution. The overlaid plots show that RITE has more variance than Burg. However, the average plots verify that RITE has less bias and higher resolution, demonstrating that RITE is indeed more robust to WGN as compared to the Burg method. However, a potential problem of using *fmincon* to find RITE is that, the iterative optimization may only produce a local maximum and not the true RITE. Actually, most of the poor estimates for RITE (flattened

PSDs) are due to local minima, as evidenced by results of a fine grid search which yield larger values of the likelihood. As an example, there is a single outlier, which is only a local maximum, as shown in Fig. 12(a) when $SNR = 30dB$ and $N = 300$. This outlier $\mathbf{k}_o = [-0.713, 0.908, -0.221, 0.186]$ produces $l_R = 0.78$, while another possible solution, found by *fmincon* with $[-0.7, 0.7, -0.7, 0.7]$ as the initial value, $\mathbf{k}_g = [-0.724, 0.968, -0.689, 0.698]$ has $l_R = 0.82$. Therefore the outlier is not the true RITE. The estimation results generated by $\mathbf{k}_o$ and $\mathbf{k}_g$ are shown in Fig. 12(b).

2.6 Conclusion

We have introduced RITE as a new method for PSD estimation. RITE and the quasi-MLE are compared analytically when the data is misspecified. In particular, the misspecification example of additive noise is studied in detail. Theoretical results have been verified via simulation for a Gaussian AR(1) process. Examples employed in this study demonstrate that RITE is indeed more robust than the quasi-MLE when WGN is present, resulting in a smaller bias. This improvement has been demonstrated for AR spectral estimation when observation noise is present. In this study, higher order AR examples had to rely on an iterative search algorithm to find the global maximum. It is not clear if this was attained. These studies have established a solid foundation to further our goal of searching for the global maximum which is the true RITE, and which may have an even better performance. Therefore, an efficient method of computing RITE will be explored in future works. It should be emphasized that RITE is a general approach to model-based spectral estimation in the presence of model inaccuracies. Its robustness properties need to be explored for other scenarios in which data models are inaccurate, which is the "rule rather than the exception".

List of References

- [1] P. J. Brockwell and R. A. Davis, *Time series: theory and methods*. Springer Science & Business Media, 2013.
- [2] S. Kay, “Poisson maximum likelihood spectral inference,” 2017, unpublished <http://www.ele.uri.edu/faculty/kay/New%20web/Books.htm>.
- [3] G. M. Jenkins and D. G. Watts, “Spectral analysis,” 1968.
- [4] S. Kay and J. Makhoul, “On the statistics of the estimated reflection coefficients of an autoregressive process,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 31, no. 6, pp. 1447–1455, 1983.
- [5] H. Cramér, “Mathematical methods of statistics,” vol. 9, 1999.
- [6] S. M. Kay, *Modern spectral estimation*. Pearson Education India, 1988.
- [7] R. T. Lacoss, “Data adaptive spectral analysis methods,” *Geophysics*, vol. 36, no. 4, pp. 661–675, 1971.
- [8] W. Chen and G. Stegen, “Experiments with maximum entropy power spectra of sinusoids,” *Journal of Geophysical Research*, vol. 79, no. 20, pp. 3019–3022, 1974.
- [9] L. Marple, “Resolution of conventional fourier, autoregressive, and special arma methods of spectrum analysis,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP’77.*, vol. 2. IEEE, 1977, pp. 74–77.
- [10] J. Cadzow, “High performance spectral estimation—a new arma method,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 5, pp. 524–529, 1980.
- [11] H. White, “Maximum likelihood estimation of misspecified models,” *Econometrica: Journal of the Econometric Society*, pp. 1–25, 1982.
- [12] S. Fortunati, F. Gini, M. S. Greco, and C. D. Richmond, “Performance bounds for parameter estimation under misspecified models: Fundamental findings and applications,” *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 142–157, 2017.
- [13] J. F. C. Kingman, *Poisson processes*. Wiley Online Library, 1993.
- [14] M. Grigoriu, “A spectral representation based model for monte carlo simulation,” *Probabilistic Engineering Mechanics*, vol. 15, no. 4, pp. 365–370, 2000.

- [15] S. Kay, "Representation and generation of non-gaussian wide-sense stationary random processes with arbitrary psds and a class of pdfs," *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3448–3458, 2010.
- [16] P. J. Huber *et al.*, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.
- [17] L. A. Stefanski and D. D. Boos, "The calculus of m-estimation," *The American Statistician*, vol. 56, no. 1, pp. 29–38, 2002.
- [18] S. M. Kay, "Fundamentals of statistical signal processing, volume i: Estimation theory (v. 1)," *PTR Prentice-Hall, Englewood Cliffs*, 1993.
- [19] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters," in *Breakthroughs in statistics*. Springer, 1992, pp. 235–247.

Figure 8: 100 Overlaid RITE Realizations (WGN)

(a) SNR=40dB

(b) SNR=35dB

(c) SNR=15dB

Figure 9: Average of RITE Realizations (WGN)

(a) SNR=40dB

(b) SNR=35dB

(c) SNR=15dB

Figure 10: 100 Overlaid Burg Realizations (WGN)

(a) SNR=40dB

(b) SNR=35dB

(c) SNR=35dB

Figure 11: Average of Burg Realizations (WGN)

(a) 100 Overlaid Realizations.

(b) Outlier and Better Solution

Figure 12: $N=300$, $SNR=30dB$ with WGN

CHAPTER 3

Robust Data Fusion for Estimation

by

Xin Zhou and Steven Kay

Dept. of Electrical, Computer and Biomedical Engineering

University of Rhode Island, Kingston, RI, USA

submitted to IEEE Signal Processing Letters.

Abstract

In this paper, we propose two estimators for data fusion estimation problem. The Fisher information and the observed Fisher information are used to reduce the negative effects of poor estimations and therefore improve the new estimators' performance in terms of mean square error. At the same time, we found that there is a relationship between our new estimators and the second order Taylor expansion of l , which is the log likelihood function of data from all sensors. The solution of the maximum of the second order Taylor expansion of l , turns out to be our new estimator that uses the observed Fisher information. Our simulation results showed that the proposed estimators have obvious advantages in both low and intermediate SNR regions, especially when one or many sensors have much lower SNRs than the others. This paper presents only the scalar case, but it should be easily extended to the multivariate case.

3.1 Introduction

Data fusion for estimation is a problem that utilizes information from multiple data sets to estimate an unknown parameter or vector. These data sets are usually from multiple sources, for example, multiple sensors. In this case, this problem is called distributed estimation. It uses data from multiple sensors and a fusion center (FC), or central processor, to achieve a more accurate estimation than using a single sensor observation [1]. It has been actively researched for decades. In late 1970's, researchers started with optimal fusion to reconstruct the global estimate [2]. Around the 2000, wireless ad hoc sensor networks information processing became a very active area [3] [4] [5]. The communication cost is expensive for these networks, so distributed estimation is preferred to require only local network information and minimized communication. Recently more interests are focused on distributed estimation algorithms that handle process noise [6] [7] [8]. Most of

those studies assumed that the local estimates are unbiased. However, this is not always the case in the real world. For example, when the noise or inference is located closely to one sensor but far away from the other sensors, that sensor may generate outliers due to low signal-to-noise ratio (SNR), which causes the bias in estimation. Currently, there aren't many investigation with the assumption that the local estimation is not ideal.

This paper will use multi-sensor estimation as an example to illustrate our method. However, our method is not limited to the distributed estimation problem. It can also be used for data from the same source at different time intervals. Also, here we focus on the algorithm at the central processor, which merges all the local estimations and the robustness against noise. An obvious way to integrate data is simply averaging the estimations from all sensors. This method doesn't require any additional information and is easy to implement. However, it's not a robust estimator. When one or some of the sensors have low SNRs, it affects not only the local estimation but also the final estimation at the central processor. A robust estimator should be able to reduce the effects of poor estimations. In the real world, poor SNR conditions are very likely to occur and may be caused by multiple reasons. For example, some sensors, compared to others, may have larger noise interference, or be farther from to the target. The SNR information is ignored in this estimator, but it is important and can be used to calculate the Fisher information (FIM). If the local estimator is an optimal one, then the FIM can be used as a measure of the estimation accuracy. By using this measure as the weighting factor for all the local results, less accurate estimates will have smaller contributions to the global estimation. Researchers have published new methods using FIM for specific applications or problems [9] [10]. Here we introduce a new one, which works for more general problems.

We propose two new estimators which use the FIM or the observed FIM. The former requires the knowledge of the expectation of the log likelihood function, and the latter needs the log likelihood function and data. There is no clear evidence to show which method is better. So in practice, one may choose the method that is easier to implement. The FIM reflects the upper limit of the accuracy of an estimator [11]. By including the FIM or the observed FIM as the weighting factors for each sensor, we can reduce the negative effects of poor estimations and therefore improve the performance of the final estimator.

Two experiments are carried out to test our estimators. Note that, below a certain SNR threshold, the mean square error (MSE) doesn't follow the Cramer Rao lower bound (CRLB) any more, and outliers show up with some probability [12]. Those outliers are the main factor that causes the rapid increase in MSE. Based on our simulation results, the integrated final estimator, which uses the FIM and the one that uses the observed FIM does not have an obvious difference. However, they both reduce the number of outliers and show an improved performance over the averaged estimator in terms of MSE.

The paper structure is as follows: Section 2 gives a detailed description of the problem, our method and reveals the relationship between the overall log likelihood function and our new estimators. Section 3 is the simulation results that shows our estimator is better than the averaged one. Section 4 is conclusion and future work.

3.2 Problem Statement

Consider M sensors are linked with a fusion center. Each sensor observes a real-valued vector $\mathbf{x}_i[n]$, $n = 0, 2, \dots, N-1$, that consists of a deterministic signal $\mathbf{s}[n; \theta]$ and white Gaussian noise $\mathbf{w}_i[n] \sim \mathcal{N}(0, \sigma_i^2)$. Here we assume that σ_i^2 's are all known. θ is the unknown parameter to be estimated, $\mathbf{w}_i$'s are independent

across the sensors.

Maximum likelihood estimation (MLE) is widely used in practice, since it is asymptotically optimal for large data record [11]. If we put aside the communication cost and consider only the estimation performance, then for large data record, the overall MLE that involves the data from all the sensors reaches the CRLB and gives the efficient estimation. Denote the overall log likelihood as l , the log likelihood at the i^{th} sensor as l_i . We have $l = \sum_{i=1}^M l_i$ and

$$l_i = -\frac{N}{2} \ln 2\pi - \frac{N}{2} \ln \sigma_i^2 - \frac{1}{2\sigma_i^2} \sum_{n=0}^{N-1} (x_i[n] - s[n; \theta])^2$$

3.2.1 Taylor Expansion of the Log Likelihood Function

By doing Taylor expansion at point $\theta = \hat{\theta}_i$, which is the MLE at the i^{th} sensor, l_i can be written as

$$\begin{aligned} l_i \approx \ln p(\mathbf{x}_i; \hat{\theta}_i) &+ \frac{1}{2!} \left. \frac{\partial^2 \ln p(\mathbf{x}_i; \theta)}{\partial \theta^2} \right|_{\theta=\hat{\theta}_i} (\theta - \hat{\theta}_i)^2 \\ &+ \frac{1}{3!} \left. \frac{\partial^3 \ln p(\mathbf{x}_i; \theta)}{\partial \theta^3} \right|_{\theta=\hat{\theta}_i} (\theta - \hat{\theta}_i)^3 + \dots \end{aligned}$$

Note that $l = \sum_{i=1}^M l_i$, so the overall log likelihood function can be expressed by

$$\begin{aligned} l \approx \sum_{i=1}^M \ln p(\mathbf{x}_i; \hat{\theta}_i) &+ \frac{1}{2!} \sum_{i=1}^M \left. \frac{\partial^2 \ln p(\mathbf{x}_i; \theta)}{\partial \theta^2} \right|_{\theta=\hat{\theta}_i} (\theta - \hat{\theta}_i)^2 \\ &+ \frac{1}{3!} \sum_{i=1}^M \left. \frac{\partial^3 \ln p(\mathbf{x}_i; \theta)}{\partial \theta^3} \right|_{\theta=\hat{\theta}_i} (\theta - \hat{\theta}_i)^3 + \dots \end{aligned}$$

Since the first term $\sum_{i=1}^M \ln p(\mathbf{x}_i; \hat{\theta}_i)$ has only data, finding the maximum of l is equivalent to finding the maximum of

$$\begin{aligned} l' = \frac{1}{2!} \sum_{i=1}^M \left. \frac{\partial^2 \ln p(\mathbf{x}_i; \theta)}{\partial \theta^2} \right|_{\theta=\hat{\theta}_i} &(\theta - \hat{\theta}_i)^2 \\ &+ \frac{1}{3!} \sum_{i=1}^M \left. \frac{\partial^3 \ln p(\mathbf{x}_i; \theta)}{\partial \theta^3} \right|_{\theta=\hat{\theta}_i} (\theta - \hat{\theta}_i)^3 + \dots \end{aligned}$$

Second Order Taylor Expansion

With the second order Taylor expansion approximated log likelihood function, the "approximated" MLE is the one that maximizes

$$\begin{aligned}
l' &= \frac{1}{2!} \sum_{i=1}^M \frac{\partial^2 \ln p(\mathbf{x}_i; \theta)}{\partial \theta^2} \Big|_{\theta=\hat{\theta}_i} (\theta - \hat{\theta}_i)^2 \\
&= \frac{1}{2} \sum_{i=1}^M \frac{\partial^2 \ln p(\mathbf{x}_i; \theta)}{\partial \theta^2} \Big|_{\theta=\hat{\theta}_i} \theta^2 - \sum_{i=1}^M \frac{\partial^2 \ln p(\mathbf{x}_i; \theta)}{\partial \theta^2} \Big|_{\theta=\hat{\theta}_i} \hat{\theta}_i \theta \\
&\quad + \frac{1}{2} \sum_{i=1}^M \frac{\partial^2 \ln p(\mathbf{x}_i; \theta)}{\partial \theta^2} \Big|_{\theta=\hat{\theta}_i} (\hat{\theta}_i)^2
\end{aligned}$$

Hence the solution of the maximum of the approximated log likelihood function is

$$\frac{\sum_{i=1}^M \frac{\partial^2 \ln p(\mathbf{x}_i; \theta)}{\partial \theta^2} \Big|_{\theta=\hat{\theta}_i} \hat{\theta}_i}{\sum_{i=1}^M \frac{\partial^2 \ln p(\mathbf{x}_i; \theta)}{\partial \theta^2} \Big|_{\theta=\hat{\theta}_i}} \quad (8)$$

In this case, $\frac{\partial^2 \ln p(\mathbf{x}_i; \theta)}{\partial \theta^2} \Big|_{\theta=\hat{\theta}_i}$ and $\hat{\theta}_i$ are enough to reconstruct an approximated log likelihood function at the central processor. The approximation may sacrifice some performance, but the communication cost can be greatly reduced.

Higher Order Taylor Expansion

By using third or higher order expansion to represent the log likelihood function, we may approximate l better in a region of $\hat{\theta}_i$'s. Unlike the second order expansion, higher order expansion does not guarantee a convex function in θ and the maximum may be difficult to find. When this happens, higher order expansion will show some wrong estimations. Hence, in practice, the second order expansion should be more useful than higher order expansion. More detailed discussion can be found in Appendix B.

3.2.2 Definition of Proposed Estimators

Without considering the communication cost, for large data record, the optimal estimator is the maximum likelihood estimator, which is defined as

$$\begin{aligned}\hat{\theta}_M &= \max_{\theta} \sum_{i=1}^M -\frac{N}{2} \ln 2\pi - N \ln \sigma_i - \frac{1}{2\sigma_i} \sum_{n=0}^{N-1} (x_i[n] - s[n; \theta])^2 \\ &= \max_{\theta} \sum_{i=1}^M -\frac{1}{2\sigma_i^2} \sum_{n=0}^{N-1} (x_i[n] - s[n; \theta])^2\end{aligned}$$

M in $\hat{\theta}_M$ stands for MLE. This estimator requires to transmit M functions of θ to the central processor. To reduce the cost, one obvious way is to send the local estimates to the central processor and then average them. By doing this, it requires to pass the estimation results only. We denote this estimator as

$$\hat{\theta}_A = 1/M \sum_{i=1}^M \hat{\theta}_i$$

A in $\hat{\theta}_A$ represents averaging since that what this estimator does is averaging the local estimations. Note that $\hat{\theta}_A$ is actually a special case of (8). It requires $\frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial \theta^2}$ to be a constant in θ and $\mathbf{x}$. This estimator is easy to process. However, when one or many sensors generate poor estimations, the performance of the final estimator $\hat{\theta}_A$ will be degraded. To improve the performance, we introduce the new estimator (8), which is the maximum of overall log likelihood approximated by second order Taylor expansion at point $\hat{\theta}_i$'s:

$$\hat{\theta}_O = \frac{\sum_{i=1}^M \left. \frac{\partial^2 \ln p(\mathbf{x}_i; \theta)}{\partial \theta^2} \right|_{\theta=\hat{\theta}_i} \hat{\theta}_i}{\sum_{i=1}^M \left. \frac{\partial^2 \ln p(\mathbf{x}_i; \theta)}{\partial \theta^2} \right|_{\theta=\hat{\theta}_i}} = \frac{\sum_{i=1}^M \beta_i \hat{\theta}_i}{\sum_{i=1}^M \beta_i} \quad (9)$$

with

$$\beta_i = - \left. \frac{\partial^2 \ln p(\mathbf{x}_i; \theta)}{\partial \theta^2} \right|_{\theta=\hat{\theta}_i}$$

Let the O in $\hat{\theta}_O$ represents the observed Fisher information. Note that $\mathcal{I}(\mathbf{x}; \theta) = -\frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial \theta^2}$ is called the observed FIM. As we can see from the previous section,

our estimator $\hat{\theta}_O$ uses $\hat{\theta}_i$ and the second derivative to reconstruct the approximated overall log likelihood function at the central processor. It greatly reduces the communication cost since we need to send β_i 's and $\hat{\theta}_i$'s only.

If we have the FIM $I(\theta) = -E[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}]$ available, then we don't need to calculate β_i for each sensor, since it depends on data $\mathbf{x}_i$. This case, the weighting factor β_i is replaced by α_i , which is the Fisher information at $\hat{\theta}_i$. We define this estimator to be $\hat{\theta}_F$:

$$\hat{\theta}_F = \frac{\sum_{i=1}^M \alpha_i \cdot \hat{\theta}_i}{\sum_{i=1}^M \alpha_i}$$

with

$$\alpha_i = \frac{1}{I^{-1}(\hat{\theta}_i)} = I(\hat{\theta}_i)$$

F in $\hat{\theta}_F$ is short for Fisher information. The Cramer-Rao Lower Bound shows that the minimum variance of all $\hat{\theta}$ is $I^{-1}(\theta)$, which indicates the accuracy of the optimal estimator [11]. The smaller the variance is, the more reliable the estimator is. Therefore, it is reasonable to use $I(\hat{\theta}_i)$ as the weighting factor in front of $\hat{\theta}_i$. It has not been proven which method is better, the FIM or the observed FIM. The observed FIM is used in many studies [13] [14]. It may be more practical since the expectation of $\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}$ is not needed.

3.3 Simulation Results

In this section, we use the sinusoid frequency estimation problem as an example. Assume we have data $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ from 3 sensors. At each sensor

$$x_i[n] = s[n; f_0] + w_i[n]$$

$$s[n; f_0] = \sin(2\pi f_0 n + \phi)$$

where $w_i[n]$ is white Gaussian noise (WGN) with variance σ_i^2 and independent from sensor to sensor. Here ϕ is known. f_0 is the unknown parameter, which is assumed not to be near 0 or 0.5. To calculate the FIM and the observed FIM, we

need to know the log likelihood function and its second derivative with respect to f_0 :

$$\ln p(\mathbf{x}_i; f_0) = -\frac{N}{2} \ln 2\pi - \frac{N}{2} \ln \sigma_i^2 - \frac{1}{2\sigma_i^2} \sum_{n=0}^{N-1} (x_i[n] - s[n; f_0])^2 \quad (10)$$

$$\begin{aligned} \frac{\partial^2 \ln p(\mathbf{x}_i; f_0)}{\partial f_0^2} &= \frac{1}{\sigma_i^2} \sum_{n=0}^{N-1} \left((x_i[n] - s[n; f_0]) \frac{\partial^2 s[n; f_0]}{\partial f_0^2} - \left(\frac{\partial s[n; f_0]}{\partial f_0} \right)^2 \right) \\ &= -\frac{1}{\sigma_i^2} \sum_{n=0}^{N-1} 4\pi^2 n^2 \cdot (\cos(4\pi f_0 n + 2\phi) + x_i[n] \sin(2\pi f_0 n + \phi)) \end{aligned}$$

By using the approximation

$$\sum_{n=0}^{N-1} n^2 \cos(4\pi f_0 n + 2\phi) \approx 0$$

we have the FIM to be

$$\begin{aligned} I_i(f_0) &= -E \left[\frac{\partial^2 \ln p(\mathbf{x}_i; f_0)}{\partial f_0^2} \right] \\ &\approx \frac{4\pi^2}{\sigma_i^2} \sum_{n=0}^{N-1} n^2 \sin^2(2\pi f_0 n + \phi) \\ &= \frac{4\pi^2}{\sigma_i^2} \sum_{n=0}^{N-1} n^2 \left(\frac{1}{2} - \frac{1}{2} \cos(4\pi f_0 n + 2\phi) \right) \\ &\approx \frac{2\pi^2}{\sigma_i^2} \sum_{n=0}^{N-1} n^2 \\ &= \frac{\pi^2 N(N-1)(2N-1)}{3\sigma_i^2} \end{aligned}$$

which varies linearly with respect to the SNR, or $\frac{1}{\sigma_i^2}$. Note that, since the data from different sensors are independent, the overall FIM is

$$I(f_0) = \sum_{i=1}^3 I_i(f_0) = \sum_{i=1}^3 \frac{1}{\sigma_i^2} \sum_{n=0}^{N-1} \frac{\pi^2 N(N-1)(2N-1)}{3}$$

Together with the two new estimators, we have 4 estimators to compare:

$$\begin{aligned}\hat{f}_A &= 1/3 \sum_{i=1}^3 \hat{f}_0^{(i)} \\ \hat{f}_F &= \frac{\sum_{i=1}^3 \hat{f}_0^{(i)} \cdot \alpha_i}{\sum_{i=1}^3 \alpha_i} \\ \hat{f}_O &= \frac{\sum_{i=1}^3 \hat{f}_0^{(i)} \cdot \beta_i}{\sum_{i=1}^3 \beta_i} \\ \hat{f}_M &= \max_{f_0} \sum_{i=1}^3 \ln p(\mathbf{x}_i; f_0)\end{aligned}$$

Here $\hat{f}_0^{(i)}$ is the MLE of the i^{th} sensor, which maximizes (10). α_i and β_i are the weighting factors of the $\hat{f}_0^{(i)}$:

$$\begin{aligned}\alpha_i &= \frac{1}{I_i^{-1}(f_0)} = I_i(f_0) \\ \beta_i &= -\left. \frac{\partial^2 \ln p(\mathbf{x}_i; f_0)}{\partial f_0^2} \right|_{f_0=\hat{f}_0^{(i)}} \\ &\approx -\frac{4\pi^2}{\sigma_i^2} \sum_{n=0}^{N-1} n^2 x_i[n] \sin(2\pi \hat{f}_0^{(i)} + \phi)\end{aligned}$$

Since $\frac{\pi^2 N(N-1)(2N-1)}{3}$ is a constant from sensor to sensor, we simplify α_i to be $\frac{1}{\sigma_i^2}$. We test the three estimators in two cases. Case 1, one of the sensors has extremely low SNR compared to the other two. In the real world, this case represents that one sensor (or some of the sensors) is (are) heavily corrupted by noise while the rest of the sensors are working properly. Case 2, the SNRs of all the sensors decrease at the same rate. In practice, this SNR reduction over all sensors may be due to the weakening of the signal or the departing of the target. The true values of parameters we use in this section are: $f_0 = 0.1, \phi = 0.1\pi$.

Case 1

Let $SNR_i = \frac{1}{2\sigma_i^2}$ denotes the SNR at the i^{th} sensor. We fix the SNR of the first two sensors to be $1dB$ and $0dB$. Let SNR_3 varies from $-10dB$ to $-1dB$.

(a) MSE vs SNR_3

(b) Zoom in on (a) at $SNR_3 = -2, -1, 0, 1, 2$

Figure 13: MSE vs SNR_3 for $N = 100$, $f_0 = 0.1$, $SNR_1 = 1dB$, $SNR_2 = 0dB$. 10,000 simulations for each SNR_3 .

(a) MSE vs Δ

(b) Zoom in on (a) at $\Delta = -5, -4, -3$

Figure 14: MSE vs Δ for $N = 100$, $f_0 = 0.1$, $SNR_1 = \Delta dB$, $SNR_2 = \Delta + 10dB$, $SNR_3 = \Delta + 20dB$. 10,000 simulations for each Δ .

For each SNR_3 , we do 10,000 simulations. Fig. 13 shows how the MSE(dB) changes when SNR_3 changes. $\hat{f}_M$ is universally optimal in the simulation. $\hat{f}_A$ has larger MSE compared with the other two estimators over all ranges, not only below the threshold $SNR_3 = -2dB$, but also above it. The MSEs of $\hat{f}_O$ and $\hat{f}_F$ are smaller than that of $\hat{f}_A$. $\hat{f}_F$ has a slight advantage over $\hat{f}_O$ when SNR_3 is below $-9dB$. However, in $-9dB < SNR_3 < 2dB$, $\hat{f}_O$ shows a smaller MSE than $\hat{f}_F$. However, we cannot conclude which one is better from one experiment, but they both outperform $\hat{f}_A$ regardless of the value of SNR_3 . The improvement of the performance in terms of MSE results in the reduction of outliers for $\hat{f}_O$ and $\hat{f}_F$. Note that in Fig. 13(b), the gap between $\hat{f}_A$ and other estimators is narrower at $SNR_3 = 0, 1$. This is because the SNR_i 's are close to each other, which means the FIM or the observed FIM are close to $1/3$, the weighting factor of $\hat{f}_A$. For more detailed analysis of the estimations at the third sensor and the MSE curve, please refer to the supplementary material.

Case 2

We fix the initial SNR at the three sensors at $\{0, 10, 20\}$ dB. For each new experiment, we subtract $1dB$ from all sensors and each experiment contains 10,000 simulations. Fig. 14 plots the MSE vs the value of the subtracted SNR, Δ . $\hat{f}_M$ is optimal and attains the CRLB in all the experiments at high enough SNR. $\hat{f}_F$ and $\hat{f}_O$ outperform $\hat{f}_A$ on all trials in terms of MSE. For example, $\hat{f}_A$ has about $30dB$ disadvantage over the other two estimators for $\Delta = -12$. This gap decreases as Δ increases in Fig. 14(b) since the number of outliers is greatly reduced when SNR is above a certain threshold. It is not obvious if $\hat{f}_O$ or $\hat{f}_F$ is better, but $\hat{f}_O$ extends the threshold to be $-4dB$ other than $-3dB$. However, they both work better than $\hat{f}_A$.

3.4 Conclusion and Future Work

We have proposed two estimators $\hat{\theta}_O$ and $\hat{\theta}_F$ for the distributed estimation problem. $\hat{\theta}_O$ turns out to be the one that maximizes the second order Taylor expansion of the overall log likelihood function. In Section 3, we compare our new estimators with the estimator $\hat{\theta}_A$ in terms of MSE. Based on the simulation results, our estimators have obvious advantages over $\hat{\theta}_A$, especially when the SNR at one or some sensors are much lower than the others. This improvement is due to the use of the FIM or the observed FIM. Since the weighting factors α_i 's or β_i 's for low SNRs are smaller than for high SNRs, the effect of poor estimations are reduced in $\hat{\theta}_O$ and $\hat{\theta}_F$. Therefore we have better performances in terms of MSE. We defined and simulated the scalar case only. It can be potentially extended to the vector case. In this paper, we assume the data come from independent sensors. This assumption can be extended to, for example, independent data sets that come from the same sensor but at different time intervals. Also, here σ_i^2 's are assumed to be known. In the future, we may further study the case that σ_i^2 's are unknown.

List of References

- [1] Y. Liu, C. Li, W. K. Tang, and Z. Zhang, "Distributed estimation over complex networks," *Information Sciences*, vol. 197, pp. 91–104, 2012.
- [2] C.-Y. Chong, "Hierarchical estimation," in *Proc. MIT/ONR Workshop on C3*, 1979, pp. 205–220.
- [3] J. Li and G. AlRegib, "Distributed estimation in energy-constrained wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3746–3758, 2009.
- [4] J.-J. Xiao, A. Ribeiro, Z.-Q. Luo, and G. B. Giannakis, "Distributed compression-estimation using wireless sensor networks," *IEEE Signal Processing Magazine*, vol. 23, no. 4, pp. 27–41, 2006.
- [5] A. Ribeiro and G. B. Giannakis, "Bandwidth-constrained distributed estimation for wireless sensor networks-part i: Gaussian case," *IEEE transactions on signal processing*, vol. 54, no. 3, pp. 1131–1143, 2006.

- [6] W. Li, Y. Jia, J. Du, *et al.*, “Diffusion kalman filter for distributed estimation with intermittent observations,” in *2015 American Control Conference (ACC)*. IEEE, 2015, pp. 4455–4460.
- [7] F. Govaers and W. Koch, “Distributed kalman filter fusion at arbitrary instants of time,” in *2010 13th International Conference on Information Fusion*. IEEE, 2010, pp. 1–8.
- [8] M. Reinhardt, B. Noack, and U. D. Hanebeck, “The hypothesizing distributed kalman filter,” in *2012 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, 2012, pp. 305–312.
- [9] S. Kay and N. Vankayalapati, “Improvement of tdoa position fixing using the likelihood curvature,” *IEEE Transactions on Signal Processing*, vol. 61, no. 8, pp. 1910–1914, 2013.
- [10] A. Yeredor and E. Angel, “Joint tdoa and fdoa estimation: A conditional bound and its use for optimally weighted localization,” *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1612–1623, 2011.
- [11] S. M. Kay, “Fundamentals of statistical signal processing, volume i: Estimation theory (v. 1),” *PTR Prentice-Hall, Englewood Cliffs*, 1993.
- [12] E. Aboutanios, “Estimating the parameters of sinusoids and decaying sinusoids in noise,” *IEEE Instrumentation & Measurement Magazine*, vol. 14, no. 2, pp. 8–14, 2011.
- [13] P. Grambsch *et al.*, “Sequential sampling based on the observed fisher information to guarantee the accuracy of the maximum likelihood estimator,” *The Annals of Statistics*, vol. 11, no. 1, pp. 68–77, 1983.
- [14] Y. Yilmaz and X. Wang, “Sequential decentralized parameter estimation under randomly observed fisher information,” *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 1281–1300, 2013.

CHAPTER 4

Future Work

Some of the assumptions and methods used in this research can be extended in the future:

- An efficient way of finding the solution of RITE estimator should be explored since its robustness against additive noise in the spectral estimation problem.
- AR model is the one we used in our study while applying RITE in spectral estimation. However, this estimator is not limited to AR model only. We can try other models as well. For example, the exponential model, which simplifies the solution finding problem to a convex optimization problem.
- As for the data fusion for estimation problem studied in chapter 3, we defined and simulated the scalar case only. It can be potentially extended to the vector case.
- Also, in chapter 3, σ_i^2 's are assumed to be known. In the future, we may further study the case that σ_i^2 's are unknown, which might be more practical in real world applications.

APPENDIX A

Supplementary Material for Chapter 2

A.1 Proof of Theorem 1

Assume the observable random process $\mathbf{x}$ is composed of a random process $\mathbf{s}$ with power equals 1 and noise $\mathbf{w}$. The PSD of $\mathbf{x}$ is $Q(f; \boldsymbol{\theta}^*)$. We propose a PSD model $P(f; \boldsymbol{\theta})$. Asymptotic MLE $\hat{\boldsymbol{\theta}}$ is found by maximizing the log likelihood function (apart from the constant not depending on the PSD)

$$l_M = - \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\ln P(f; \boldsymbol{\theta}) + \frac{I(f)}{P(f; \boldsymbol{\theta})} \right) df$$

Define $\boldsymbol{\theta}_0$ to be the one that maximizes $E_{\boldsymbol{\theta}^*}(l_M)$. We make two assumptions:

- 1, l_M is differentiable w.r.t $\boldsymbol{\theta}$.
- 2, $\{\mathbf{c} \cdot \frac{\partial l_M}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\}$ satisfies the Lyapunov condition for any constant vector $\mathbf{c}$ at any frequency.

If the first assumption is valid, then $\hat{\boldsymbol{\theta}}$ is the solution of

$$\frac{\partial l_M}{\partial \boldsymbol{\theta}} = - \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{I(f)}{P(f; \boldsymbol{\theta})} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} df = 0$$

For large data record, $\frac{\partial l_M}{\partial \boldsymbol{\theta}}$ can be approximated by

$$\frac{\partial l_M}{\partial \boldsymbol{\theta}} \approx - \frac{2}{N} \sum_{i=1}^{\frac{N}{2}} \frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{I(f_i)}{P(f_i; \boldsymbol{\theta})} \frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \approx - \frac{2}{N} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{M_i}}{\partial \boldsymbol{\theta}}$$

where $f_1 = 0$ and $f_{i+1} - f_i = \frac{1}{N}$ for $i = 1, \dots, \frac{N}{2} - 1$. If we expand $\frac{\partial l_M}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ near the value $\boldsymbol{\theta}_0$, we will get

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx - \left(\frac{1}{N} \sum_{i=1}^{\frac{N}{2}} \frac{\partial^2 l_{M_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right)^{-1} \times \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{M_i}}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right)$$

where

$$\frac{\partial l_{M_i}}{\partial \boldsymbol{\theta}} = \frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{I(f_i)}{P(f_i; \boldsymbol{\theta})} \frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

$$[\frac{\partial^2 l_{M_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}]_{ul} = \frac{\partial^2 \ln P(f_i; \boldsymbol{\theta})}{\partial \theta_u \partial \theta_l} \left(1 - \frac{I(f_i)}{P(f_i; \boldsymbol{\theta})} \right) + \frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \theta_l} \frac{I(f_i)}{P(f_i; \boldsymbol{\theta})}$$

The term

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{M_i}}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \frac{1}{\sqrt{N}} \left[\sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{M_i}}{\partial \theta_1}; \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{M_i}}{\partial \theta_2}; \dots; \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{M_i}}{\partial \theta_p} \right]^T |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

is multidimensional normal distributed. The proof is following:

Proof. Let $\mathbf{c} = [c_1 \ c_2 \ \dots \ c_p]^T$ be a $p \times 1$ vector. A linear combination of the components of $\frac{1}{\sqrt{N}} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{M_i}}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ is

$$\begin{aligned} & \mathbf{c}^T \frac{1}{\sqrt{N}} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{M_i}}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^{\frac{N}{2}} \left(c_1 \frac{\partial l_{M_i}}{\partial \theta_1} + c_2 \frac{\partial l_{M_i}}{\partial \theta_2} + \dots + c_p \frac{\partial l_{M_i}}{\partial \theta_p} \right) |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \sum_{i=1}^{\frac{N}{2}} S_i \end{aligned}$$

Define

$$\begin{aligned} B_{\frac{N}{2}} &= \sum_{i=1}^{\frac{N}{2}} E(|S_i - \mu_i|^{2+\epsilon}) \\ \mu_i &= E(S_i) \\ C_{\frac{N}{2}} &= \left(\sum_{i=1}^{\frac{N}{2}} \sigma_{S_i}^2 \right)^{\frac{2+\epsilon}{2}} \\ \sigma_{S_i}^2 &= Var(S_i) \end{aligned}$$

By the second assumption, $\lim_{N \rightarrow \infty} \frac{B_{\frac{N}{2}}}{C_{\frac{N}{2}}} = 0$ for some positive ϵ . It follows by Lyapunov's central limit theorem (CLT) that

$$\sum_{i=1}^{\frac{N}{2}} S_i \sim \mathcal{N}(0, C_{\frac{N}{2}}^2)$$

Since $\sum_{i=1}^{\frac{N}{2}} S_i = \mathbf{c}^T \frac{1}{\sqrt{N}} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{M_i}}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ is normal distributed for any $\mathbf{c}$, it is equivalent to say that $\frac{1}{\sqrt{N}} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{M_i}}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ has multivariate normal distribution. $\square$

Hence

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{M_i}}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \stackrel{a}{\sim} \mathcal{N}(\boldsymbol{\mu}, \mathbf{B}(\boldsymbol{\theta}_0)) \quad (\text{A.1})$$

where

$$\boldsymbol{\mu} = E_{\boldsymbol{\theta}^*} \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{M_i}}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right) = -\frac{\sqrt{N}}{2} E_{\boldsymbol{\theta}^*} \left(\frac{\partial l_M}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right)$$

By definition of $\boldsymbol{\theta}_0$, $E_{\boldsymbol{\theta}^*}(\frac{\partial l_M}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}) = 0$, hence $\boldsymbol{\mu} = \mathbf{0}$.

The ul^{th} element in $\mathbf{B}(\boldsymbol{\theta})$ can be expressed by

$$\begin{aligned}
& [\mathbf{B}(\boldsymbol{\theta})]_{ul} \\
&= E_{\boldsymbol{\theta}^*} \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{M_i}}{\partial \theta_u} \cdot \frac{1}{\sqrt{N}} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{M_i}}{\partial \theta_l} \right) \\
&= \frac{1}{N} E_{\boldsymbol{\theta}^*} \left(\sum_{i=1}^{\frac{N}{2}} \sum_{j=1}^{\frac{N}{2}} \frac{\partial l_{M_i}}{\partial \theta_u} \frac{\partial l_{M_j}}{\partial \theta_l} \right) \\
&= \frac{1}{N} E_{\boldsymbol{\theta}^*} \left(\sum_{i=1}^{\frac{N}{2}} \sum_{j=1}^{\frac{N}{2}} \frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \theta_u} \left(1 - \frac{I(f_i)}{P(f_i; \boldsymbol{\theta})}\right) \times \frac{\partial \ln P(f_j; \boldsymbol{\theta})}{\partial \theta_l} \left(1 - \frac{I(f_j)}{P(f_j; \boldsymbol{\theta})}\right) \right) \\
&= \frac{1}{N} \sum_{i=1}^{\frac{N}{2}} \sum_{j=1}^{\frac{N}{2}} \frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f_j; \boldsymbol{\theta})}{\partial \theta_l} \\
&\quad \times E_{\boldsymbol{\theta}^*} \left(1 - \frac{I(f_i)}{P(f_i; \boldsymbol{\theta})} - \frac{I(f_j)}{P(f_j; \boldsymbol{\theta})} + \frac{I(f_i)I(f_j)}{P(f_i; \boldsymbol{\theta})P(f_j; \boldsymbol{\theta})} \right) \\
&= \frac{1}{N} \sum_{i=1}^{\frac{N}{2}} \sum_{j=1}^{\frac{N}{2}} \left(\frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f_j; \boldsymbol{\theta})}{\partial \theta_l} \right. \\
&\quad \times \left(1 - \frac{Q(f_i; \boldsymbol{\theta}^*)}{P(f_i; \boldsymbol{\theta})} - \frac{Q(f_j; \boldsymbol{\theta}^*)}{P(f_j; \boldsymbol{\theta})} + \frac{Q(f_i; \boldsymbol{\theta}^*)Q(f_j; \boldsymbol{\theta}^*)}{P(f_i; \boldsymbol{\theta})P(f_j; \boldsymbol{\theta})} \right) \Bigg) \\
&\quad + \frac{1}{N} \sum_{i=1}^{\frac{N}{2}} \frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \theta_l} \frac{Q^2(f_i; \boldsymbol{\theta}^*)}{P^2(f_i; \boldsymbol{\theta})} \\
&= N \int_0^{\frac{1}{2}} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_u} \left(1 - \frac{Q(f; \boldsymbol{\theta}^*)}{P(f; \boldsymbol{\theta})}\right) df \times \int_0^{\frac{1}{2}} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_l} \left(1 - \frac{Q(f; \boldsymbol{\theta}^*)}{P(f; \boldsymbol{\theta})}\right) df \\
&\quad + \int_0^{\frac{1}{2}} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_l} \frac{Q^2(f; \boldsymbol{\theta}^*)}{P^2(f; \boldsymbol{\theta})} df \\
&[\mathbf{B}(\boldsymbol{\theta}_0)]_{ul} = \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_l} \frac{Q^2(f; \boldsymbol{\theta}^*)}{P^2(f; \boldsymbol{\theta})} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} df
\end{aligned}$$

By weak law of large number, it can be proved that

$$\frac{1}{N} \sum_{i=1}^{\frac{N}{2}} \left[\frac{\partial^2 l_{M_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right]_{ul} \xrightarrow{P} [\mathbf{A}(\boldsymbol{\theta}_0)]_{ul}$$

where

$$\begin{aligned}
& [\mathbf{A}(\boldsymbol{\theta}_0)]_{ul} \\
&= E_{\boldsymbol{\theta}^*} \left(\frac{1}{N} \sum_{i=1}^{\frac{N}{2}} \left[\frac{\partial^2 l_{M_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right)_{ul} \\
&= \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{\partial^2 \ln P(f; \boldsymbol{\theta})}{\partial \theta_u \partial \theta_l} \left(1 - \frac{Q(f; \boldsymbol{\theta}^*)}{P(f; \boldsymbol{\theta})} \right) + \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_l} \frac{Q(f; \boldsymbol{\theta}^*)}{P(f; \boldsymbol{\theta})} \right) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} df
\end{aligned}$$

Since every element in $\frac{1}{N} \sum_{i=1}^{\frac{N}{2}} \left[\frac{\partial^2 l_{M_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ converges to $[\mathbf{A}(\boldsymbol{\theta}_0)]_{ul}$, then

$$\frac{1}{N} \sum_{i=1}^{\frac{N}{2}} \left[\frac{\partial^2 l_{M_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \xrightarrow{P} \mathbf{A}(\boldsymbol{\theta}_0) \quad (\text{A.2})$$

By (A.1), (A.2), and Slutsky's theorem, it follows that

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{A}(\boldsymbol{\theta}_0)^{-1} \mathbf{B}(\boldsymbol{\theta}_0) \mathbf{A}(\boldsymbol{\theta}_0)^{-T})$$

where

$$\begin{aligned}
[\mathbf{A}(\boldsymbol{\theta}_0)]_{ul} &= \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{\partial^2 \ln P(f; \boldsymbol{\theta})}{\partial \theta_u \partial \theta_l} \left(1 - \frac{Q(f; \boldsymbol{\theta}^*)}{P(f; \boldsymbol{\theta})} \right) \right. \\
&\quad \left. + \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_l} \frac{Q(f; \boldsymbol{\theta}^*)}{P(f; \boldsymbol{\theta})} \right) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} df \\
[\mathbf{B}(\boldsymbol{\theta}_0)]_{ul} &= \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_l} \times \frac{Q^2(f; \boldsymbol{\theta}^*)}{P^2(f; \boldsymbol{\theta})} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} df
\end{aligned}$$

A.2 Proof of Theorem 2

Similarly to asymptotic MLE, RITE estimator $\hat{\boldsymbol{\theta}}$ is asymptotically multivariate distributed when N goes to infinity. Here the likelihood function to be maximized is

$$l_R = \int_{-\frac{1}{2}}^{\frac{1}{2}} I(f) \ln P(f; \boldsymbol{\theta}) df$$

Denote $\boldsymbol{\theta}_0$ as the one that maximizes $E_{\boldsymbol{\theta}^*}(l_R)$. Assume the two assumptions are valid: 1, l_R is differentiable w.r.t $\boldsymbol{\theta}$. 2, $\{\mathbf{c} \cdot \frac{\partial l_R}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\}$ satisfies the Lyapunov condition for any constant vector $\mathbf{c}$ at any frequency.

If the first assumption is valid, If l_R is differentiable w.r.t $\boldsymbol{\theta}$, then $\hat{\boldsymbol{\theta}}$ is the solution of

$$\frac{\partial l_R}{\partial \boldsymbol{\theta}} = \int_{-\frac{1}{2}}^{\frac{1}{2}} I(f) \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} df$$

When N is large enough, $\frac{\partial l_R}{\partial \boldsymbol{\theta}}$ can be approximated by

$$\frac{\partial l_R}{\partial \boldsymbol{\theta}} \approx \frac{2}{N} \sum_{i=1}^{\frac{N}{2}} I(f_i) \frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \approx \frac{2}{N} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{R_i}}{\partial \boldsymbol{\theta}}$$

where $f_1 = 0$ and $f_{i+1} - f_i = \frac{1}{N}$ for $i = 1, \dots, \frac{N}{2} - 1$. If we expand $\frac{\partial l_R}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ near the value $\boldsymbol{\theta}_0$, we will get

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx - \left(\frac{1}{N} \sum_{i=1}^{\frac{N}{2}} \frac{\partial^2 l_{R_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right)^{-1} \times \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{R_i}}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right)$$

where

$$\begin{aligned} \frac{\partial l_{R_i}}{\partial \boldsymbol{\theta}} &= I(f_i) \frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ \left[\frac{\partial^2 l_{R_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right]_{ul} &= I(f_i) \frac{\partial^2 \ln P(f_i; \boldsymbol{\theta})}{\partial \theta_u \partial \theta_l} \end{aligned}$$

The term

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{R_i}}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \frac{1}{\sqrt{N}} \left[\sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{R_i}}{\partial \theta_1}; \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{R_i}}{\partial \theta_2}; \dots; \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{R_i}}{\partial \theta_p} \right]^T |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

is multidimensional normal distributed. The proof is similar as in A.1. Hence

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{R_i}}{\partial \boldsymbol{\theta}} \big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \stackrel{a}{\sim} \mathcal{N}(\boldsymbol{\mu}, \mathbf{B}(\boldsymbol{\theta}_0)) \quad (\text{A.3})$$

where

$$\boldsymbol{\mu} = E_{\boldsymbol{\theta}^*} \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{R_i}}{\partial \boldsymbol{\theta}} \big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right) = \frac{\sqrt{N}}{2} E_{\boldsymbol{\theta}^*} \left(\frac{\partial l_R}{\partial \boldsymbol{\theta}} \big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right)$$

By definition of $\boldsymbol{\theta}_0$, $E_{\boldsymbol{\theta}^*} \left(\frac{\partial l_R}{\partial \boldsymbol{\theta}} \big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right) = \mathbf{0}$, hence $\boldsymbol{\mu} = \mathbf{0}$.

The ul^{th} element in $\mathbf{B}(\boldsymbol{\theta})$ is

$$\begin{aligned} [\mathbf{B}(\boldsymbol{\theta})]_{ul} &= E_{\boldsymbol{\theta}^*} \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{R_i}}{\partial \theta_u} \cdot \frac{1}{\sqrt{N}} \sum_{i=1}^{\frac{N}{2}} \frac{\partial l_{R_i}}{\partial \theta_l} \right) \\ &= \frac{1}{N} E_{\boldsymbol{\theta}^*} \left(\sum_{i=1}^{\frac{N}{2}} \sum_{j=1}^{\frac{N}{2}} \frac{\partial l_{R_i}}{\partial \theta_u} \frac{\partial l_{R_j}}{\partial \theta_l} \right) \\ &= \frac{1}{N} E_{\boldsymbol{\theta}^*} \left(\sum_{i=1}^{\frac{N}{2}} \sum_{j=1}^{\frac{N}{2}} \frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \theta_u} I(f_i) \times \frac{\partial \ln P(f_j; \boldsymbol{\theta})}{\partial \theta_l} I(f_j) \right) \\ &= \frac{1}{N} \sum_{i=1}^{\frac{N}{2}} \sum_{j=1}^{\frac{N}{2}} \frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f_j; \boldsymbol{\theta})}{\partial \theta_l} \times E_{\boldsymbol{\theta}^*} (I(f_i) I(f_j)) \\ &= \frac{1}{N} \sum_{i=1}^{\frac{N}{2}} \sum_{j=1}^{\frac{N}{2}} \frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f_j; \boldsymbol{\theta})}{\partial \theta_l} \times Q(f_i; \boldsymbol{\theta}^*) Q(f_j; \boldsymbol{\theta}^*) \\ &\quad + \frac{1}{N} \sum_{i=1}^{\frac{N}{2}} \frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f_i; \boldsymbol{\theta})}{\partial \theta_l} Q^2(f_i; \boldsymbol{\theta}^*) \\ &= N \int_0^{\frac{1}{2}} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_u} Q(f; \boldsymbol{\theta}^*) df \times \int_0^{\frac{1}{2}} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_l} Q(f; \boldsymbol{\theta}^*) df \\ &\quad + \int_0^{\frac{1}{2}} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_l} Q^2(f; \boldsymbol{\theta}^*) df \\ [\mathbf{B}(\boldsymbol{\theta}_0)]_{ul} &= \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_l} \times Q^2(f; \boldsymbol{\theta}^*) \big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} df \end{aligned}$$

By weak law of large number, it can be proved that

$$\frac{1}{N} \sum_{i=1}^{\frac{N}{2}} \left[\frac{\partial^2 l_{R_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right]_{ul} \xrightarrow{P} [\mathbf{A}(\boldsymbol{\theta}_0)]_{ul}$$

where

$$\begin{aligned} [\mathbf{A}(\boldsymbol{\theta}_0)]_{ul} &= E_{\boldsymbol{\theta}^*} \left(\frac{1}{N} \sum_{i=1}^{\frac{N}{2}} \left[\frac{\partial^2 l_{R_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right)_{ul} \\ &= \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\partial^2 \ln P(f; \boldsymbol{\theta})}{\partial \theta_u \partial \theta_l} Q(f; \boldsymbol{\theta}^*)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} df \end{aligned}$$

Since every element in $\frac{1}{N} \sum_{i=1}^{\frac{N}{2}} \left[\frac{\partial^2 l_{R_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ converges to $[\mathbf{A}(\boldsymbol{\theta}_0)]_{ul}$, then

$$\frac{1}{N} \sum_{i=1}^{\frac{N}{2}} \left[\frac{\partial^2 l_{R_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \xrightarrow{P} \mathbf{A}(\boldsymbol{\theta}_0) \quad (\text{A.4})$$

By (A.3), (A.4), and Slutsky's theorem, we have

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{a}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{A}(\boldsymbol{\theta}_0)^{-1} \mathbf{B}(\boldsymbol{\theta}_0) \mathbf{A}(\boldsymbol{\theta}_0)^{-T})$$

where

$$\begin{aligned} [\mathbf{A}(\boldsymbol{\theta}_0)]_{ul} &= \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\partial^2 \ln P(f; \boldsymbol{\theta})}{\partial \theta_u \partial \theta_l} Q(f; \boldsymbol{\theta}^*)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} df \\ [\mathbf{B}(\boldsymbol{\theta}_0)]_{ul} &= \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_u} \frac{\partial \ln P(f; \boldsymbol{\theta})}{\partial \theta_l} \times Q^2(f; \boldsymbol{\theta}^*)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} df \end{aligned}$$

A.3

Regarding to the univariate and correct model, the variance of RITE is

$$\begin{aligned}\sigma_R^2 &= \frac{2 \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(P(f; \theta_0) \frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 |_{\theta=\theta_0} df}{\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} P(f; \theta_0) \frac{\partial^2 \ln P(f; \theta)}{\partial \theta^2} |_{\theta=\theta_0} df \right)^2} \\ &= \frac{\int_0^{\frac{1}{2}} \left(P(f; \theta_0) \frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 |_{\theta=\theta_0} df}{\left(\int_0^{\frac{1}{2}} P(f; \theta_0) \frac{\partial^2 \ln P(f; \theta)}{\partial \theta^2} |_{\theta=\theta_0} df \right)^2}\end{aligned}$$

and the variance of MLE is

$$\sigma_M^2 = \frac{2}{\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 |_{\theta=\theta_0} df} = \frac{1}{\int_0^{\frac{1}{2}} \left(\frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 |_{\theta=\theta_0} df}$$

Cauchy Schwarz inequality gives that,

$$\int |f(x)|^2 dx \int |g(x)|^2 dx \geq \left| \int f(x) \bar{g}(x) dx \right|^2$$

If we let $f(x) = P(f; \theta_0) \frac{\partial \ln P(f; \theta)}{\partial \theta} |_{\theta=\theta_0}$ and $g(x) = \frac{\partial \ln P(f; \theta)}{\partial \theta} |_{\theta=\theta_0}$, then

$$\begin{aligned}& \int_0^{\frac{1}{2}} \left(P(f; \theta_0) \frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 |_{\theta=\theta_0} df \int_0^{\frac{1}{2}} \left(\frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 |_{\theta=\theta_0} df \\ & \geq \left(\int_0^{\frac{1}{2}} P(f; \theta_0) \left(\frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 |_{\theta=\theta_0} df \right)^2\end{aligned}$$

Since $\int_0^{\frac{1}{2}} P(f; \theta_0) df = 0.5$, it follows that $\int_{-\frac{1}{2}}^{\frac{1}{2}} P''(f; \theta_0) df = 0$ where $P''(f; \theta_0)$ is the second derivative with respect to θ . If we add $-\int_{-\frac{1}{2}}^{\frac{1}{2}} P''(f; \theta_0) df$ to the right side of the inequality, we have

$$\begin{aligned}& \left(\int P(f; \theta_0) \left(\left(\frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 - \frac{P''(f; \theta_0)}{P(f; \theta_0)} \right) |_{\theta=\theta_0} df \right)^2 \\ & = \left(\int_0^{\frac{1}{2}} P(f; \theta_0) \frac{\partial^2 \ln P(f; \theta)}{\partial \theta^2} |_{\theta=\theta_0} df \right)^2\end{aligned}$$

Hence the inequality becomes

$$\begin{aligned} & \int_0^{\frac{1}{2}} \left(P(f; \theta_0) \frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 \Big|_{\theta=\theta_0} df \times \int_0^{\frac{1}{2}} \left(\frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 \Big|_{\theta=\theta_0} df \\ & \geq \left(\int_0^{\frac{1}{2}} P(f; \theta_0) \frac{\partial^2 \ln P(f; \theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} df \right)^2 \end{aligned}$$

or

$$\frac{\int_0^{\frac{1}{2}} \left(P(f; \theta_0) \frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 \Big|_{\theta=\theta_0} df}{\left(\int_0^{\frac{1}{2}} P(f; \theta_0) \frac{\partial^2 \ln P(f; \theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} df \right)^2} \geq \frac{1}{\int_0^{\frac{1}{2}} \left(\frac{\partial \ln P(f; \theta)}{\partial \theta} \right)^2 \Big|_{\theta=\theta_0} df}$$

The left side is the variance of RITE, and the right side is the variance of MLE.

So MLE always has smaller variance than RITE.

A.4 Spectral Estimation for AR Process in Noise Modeled by AR

The AR process $s[n]$ is noise corrupted such that the observed data $x[n]$ becomes

$$x[n] = s[n] + \alpha w[n]$$

The scalar α reflects the SNR. In Section A.4 and A.5, the data length N is 350. $s[n]$ is an AR(4) process with the same parameters as in the main paper,

$$[a[1], a[2], a[3], a[4]] = [-2.7428, 3.7906, -2.6454, 0.93]$$

Here we adopt AR(4) model when comparing spectral estimation by RITE and by Burg method.

A.4.A White Mixture Gaussian Noise

$w[n]$ in this section is a white mixture Gaussian noise with PDF:

$$f(w) = (1 - \epsilon)\mathcal{N}(w; \mu, \sigma^2) + \epsilon\mathcal{N}(w; -\mu, \lambda\sigma^2)$$

where

$$\mathcal{N}(w; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(w - \mu)^2}{2\sigma^2}\right)$$
$$\epsilon = 0.5, \mu = 1, \sigma^2 = 1, \lambda = 100$$

So the variance of the noise is 51.5. Simulation results for mixture Gaussian white noise corrupted AR process are shown in Fig.A1 to A4.

A.4.B White Laplacian Noise

This section presents the simulation results for white Laplacian noise corrupted AR(4) process (shown in Fig.A5 to A8).

A.4.C α -Stable Modeled Impulsive Noise

The α -Stable distribution can model an impulsive noise. It is denoted by $S(\alpha, \beta, \gamma, \delta)$. α is called the characteristic exponent. β is the skewness. γ is

the scale or dispersion parameter. δ is the location. Since the power of an α -Stable process is not defined, the conventional SNR cannot be used. We define the modified SNR to be:

$$MSNR = 10 \log_{10} \frac{r_s[0]}{2\gamma^2}$$

where $r_s[0]$ is the power of the signal $s[n]$. In our case, $r_s[0] = 1$ due to the unit power assumption. The impulsive noise is generated by independent identical distributed α -Stable variable with $\alpha = 1.5$, $\beta = 0$, $\delta = 0$. Results are shown in Fig.A9 to A12. .

A.5 Spectral Estimation for AR process in Noise Modeled by ARMA

Instead of using an AR model for an AR process in noise, we can use an ARMA model for spectral estimation. This model represents the data better. However, unlike AR model, the ARMA parameter estimation via maximum likelihood criteria requires to minimize a nonlinear function. For this reason, it is not computationally efficient. The least squares modified Yule-Walker equations (LSMYWE) is a suboptimal estimator, but can be easily implemented.

A.5.A White Gaussian Noise

If the noise is white noise, then the true model for the noise corrupted AR(p) process is ARMA(p,p). In this section, we use ARMA(4,4) via LSMYWE for spectral estimation of the AR(4) process in WGN. Due to less error in modeling, results are better than using AR(4) model via Burg method (Fig. 4, 5 in main paper), and are shown in Fig.A13, A14.

A.5.B α -Stable Modeled Impulsive Noise

Since the α -Stable Noise is not white noise, it is not clear how to choose a proper order of the MA part. One can use a model order selection criteria. Here we simply use ARMA(4,4). Results are shown in Fig.A15, A16. They are not as good as in A.5.A because the order for the MA part is incorrect. At the same time, we also provide RITE estimations using AR(4) model and LSMYWE as the initial value for *fmincon* (Figs.A17, A18). As we can see, even for SNR=15dB, when LSMYWE PSDs are all flattened, RITE is able to resolve the peaks occasionally. Since the *fmincon* solution of RITE is heavily dependent on the initial value, the flattened RITE PSDs are probably produced by local solution instead of the true RITE. The true RITE may have even more promising results.

A.6 Proof of the Statement in Page 23

The Kullback-Leibler divergence is always non-negative,

$$\begin{aligned} D_{KL}(\bar{I}(f)||P(f)) &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \bar{I}(f) \ln \left(\frac{\bar{I}(f)}{P(f)} \right) df \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \bar{I}(f) \ln \bar{I}(f) df - \int_{-\frac{1}{2}}^{\frac{1}{2}} \bar{I}(f) \ln P(f) df \end{aligned}$$

a result known as Gibbs' Inequality, with $D_{KL}(\bar{I}(f)||P(f))$ zero if and only if $\bar{I}(f) = P(f)$ almost everywhere. Since $\int_{-\frac{1}{2}}^{\frac{1}{2}} \bar{I}(f) \ln \bar{I}(f) df$ involves only $\bar{I}(f)$, minimizing $D_{KL}(\bar{I}(f)||P(f))$ is equivalent to maximizing $\int_{-\frac{1}{2}}^{\frac{1}{2}} \bar{I}(f) \ln P(f) df$. Hence, the function $\int_{-\frac{1}{2}}^{\frac{1}{2}} \bar{I}(f) \ln P(f) df$ achieves its maximum when $P(f)$ is identical to $\bar{I}(f)$.

Figure A1: 100 Overlaid RITE Realizations (White Mixture Gaussian Noise)

(a) SNR=40dB

(b) SNR=35dB

(c) SNR=15dB

Figure A2: Average of RITE Realizations (White Mixture Gaussian Noise)

(a) SNR=40dB

(b) SNR=35dB

(c) SNR=15dB

Figure A3: 100 Overlaid Burg Realizations (White Mixture Gaussian Noise)

(a) SNR=40dB

(b) SNR=35dB

(c) SNR=15dB

Figure A4: Average of Burg Realizations (White Mixture Gaussian Noise)

Figure A5: 100 Overlaid RITE Realizations (White Laplacian Noise)

(a) SNR=40dB

(b) SNR=35dB

(c) SNR=15dB

Figure A6: Average of RITE Realizations (White Laplacian Noise)

(a) SNR=40dB

(b) SNR=35dB

(c) SNR=15dB

Figure A7: 100 Overlaid Burg Realizations (White Laplacian Noise)

(a) SNR=40dB

(b) SNR=35dB

(c) SNR=15dB

Figure A8: Average of Burg Realizations (White Laplacian Noise)

(a) MSNR=40dB

(b) MSNR=35dB

(c) MSNR=15dB

Figure A9: 100 Overlaid RITE Realizations (Impulsive Noise)

(a) MSNR=40dB

(b) MSNR=35dB

(c) MSNR=15dB

Figure A10: Average of RITE Realizations (Impulsive Noise)

(a) MSNR=40dB

(b) MSNR=35dB

(c) MSNR=15dB

Figure A11: 100 Overlaid Burg Realizations (Impulsive Noise)

(a) MSNR=40dB

(b) MSNR=35dB

(c) MSNR=15dB

Figure A12: Average of Burg Realizations (Impulsive Noise)

(a) SNR=40dB

(b) SNR=35dB

(c) SNR=15dB

Figure A13: 100 Overlaid LSMYWE Realizations (WGN)

(a) SNR=40dB

(b) SNR=35dB

(c) SNR=15dB

Figure A14: Average of LSMYWE Realizations (WGN)

(a) MSNR=40dB

(b) MSNR=35dB

(c) MSNR=15dB

Figure A15: 100 Overlaid LSMYWE Realizations (Impulsive Noise)

(a) MSNR=40dB

(b) MSNR=35dB

(c) MSNR=15dB

Figure A16: Average of LSMYWE Realizations (Impulsive Noise)

Figure A17: 100 Overlaid RITE (LSMYWE Initial) Realizations (Impulsive Noise)

(a) MSNR=40dB

(b) MSNR=35dB

(c) MSNR=15dB

Figure A18: Average of RITE (LSMYWE Initial) Realizations (Impulsive Noise)

APPENDIX B

Supplementary Material for Chapter 3

B.1 Analysis of the Estimations at the Third Sensor for Case 1

At the third sensor, when the SNR is below a certain threshold, $\hat{f}_0^{(3)}$ is no longer an asymptotic optimal estimator and its MSE does not follow the CRLB. Outliers occur with a certain probability and are uniformly distributed along $(0, 0.5)$. For large data record, $\hat{f}_0^{(3)} \approx \mathcal{N}(0, I_3^{-1}(f_0))$. 99% of $\hat{f}_0^{(3)}$'s fall in the range $(f_0 - 3\sqrt{I_3^{-1}(f_0)}, f_0 + 3\sqrt{I_3^{-1}(f_0)})$. Since most of the outliers of the final estimator are caused by $\hat{f}_0^{(3)}$, we define the global estimation to be a good one if it falls inside $(f_0 - 3\sqrt{I_3^{-1}(f_0)}, f_0 + 3\sqrt{I_3^{-1}(f_0)})$ and an outlier if it falls outside this range. In total 10^5 realizations for $SNR_3 = -9dB$, $\hat{f}_A$ has 20102 outliers, $\hat{f}_O$ has 14592, and $\hat{f}_F$ has 15194. As we can see from Fig. B2, $\hat{f}_A$ has many outliers distributed from 0.06 to 0.24. Note that there is a valley at $\hat{f}_0 = 0.1$ due to the fact that nearby estimations are all counted as good estimations instead of outliers. The outliers lead to a large increase in MSE. Unlike $\hat{f}_A$, the other two estimators $\hat{f}_F$ and $\hat{f}_O$ have their outliers centered around the true value $f = 0.1$. This is because α_3 and β_3 are much smaller than α_1, α_2 and β_1, β_2 . So the poor estimations of $\hat{f}^{(3)}$ do not affect the overall $\hat{f}_F$ and $\hat{f}_O$ much, but do affect $\hat{f}_A$. Besides, in Fig. B1, the good estimations of $\hat{f}_A$ have larger variance than that of $\hat{f}_O$ and $\hat{f}_F$, which also makes a difference in the MSE.

B.2 Higher Order Taylor Expansion

According to simulation results, 3rd order and 4th order are not as good as the 2nd order Taylor expansion of MLE. Since the higher order expansion generates some wrong estimation results at 0.5 (true value should be 0.1).

The Taylor expansion can approximate a function in a nearby region of a given point, but might be quite far away from the true value in other region. In our case, we want to find the maximum of the overall log likelihood. If we do Taylor expansion of the second order at a given point (MLE), then it guarantees a convex function and has only one maximum (at the given point, MLE). However, if we do a higher order expansion, then the function might be approximated more accurately near the given point, but at the end point 0 or 0.5, the expansion could be far from the true value and may be a maximum. So if we use this maximum, it will give us a wrong solution.

Below are the figures of the true log likelihood and Taylor expansion for one realization. In Fig.3, we can see the higher order can approximate the nearby region more accurately. However in Fig.1 and 2, the 3rd and 4th order expansion both give maximum at 0.5, which results in a wrong estimation. This is an example for one realization. It doesn't happen to all realizations, but when it happens in some realizations, it does decrease the performance of the higher order expansion. So among, 2nd, 3rd, 4th order of expansions, the best expansion order is 2. Or, if we have large enough order of expansion, which means it can represent the true log likelihood function, we might have better performance, but it is not possible in practice.

B.3 More Simulation Examples

This section we use more examples to show the performance of the proposed estimators. Since the observed FIM and FIM perform similarly, we focus on the

difference between the averaged one $\hat{f}_A$ and the FIM one $\hat{f}_F$.

The SNR's from different sensors are $[-6.94; 0.45; -0.86]$ in dB. $\hat{f}_A$ is shown in red and $\hat{f}_F$ is shown in blue. Over all tested frequencies of f_0 (Fig. B6), $\hat{f}_F$ has a smaller MSE than $\hat{f}_A$. The MSE of Both estimators decreases as f_0 moves closer to 0.25 (Fig. B6, B9). As we can see from Fig. B7, B8 $\hat{f}_F$ has less variance and smaller bias compared to $\hat{f}_A$.

Figure B1: The number of good estimations. 10^5 realizations for $N = 100$, $f_0 = 0.1$, $SNR_1 = 1dB$, $SNR_2 = 0dB$, $SNR_3 = -9dB$.

Figure B2: The number of outliers. 10^5 realizations for $N = 100$, $f_0 = 0.1$, $SNR_1 = 1dB$, $SNR_2 = 0dB$, $SNR_3 = -9dB$.

Figure B3: Taylor expansion up to 3rd order

Figure B4: Taylor expansion up to 4th order

Figure B5: Taylor expansion up to 4th order at the region near $f=0.1$

Figure B6: The MSE vs different value of f_0 for $N = 50$, $\phi_1 = 0.314$ (5000 realizations).

(a) Scattered Realizations

(b) Histogram of Estimators

Figure B7: $N = 50$, $\phi_1 = 0.314$, $f_1 = 0.05$ (5000 realizations).

(a) Scattered Realizations

(b) Histogram of Estimators

Figure B8: $N = 50$, $\phi_1 = 0.314$, $f_1 = 0.05$ (5000 realizations).

(a) Scattered Realizations

(b) Histogram of Estimators

Figure B9: $N = 50$, $\phi_1 = 0.314$, $f_1 = 0.05$ (5000 realizations).

BIBLIOGRAPHY

- Aboutanios, E., “Estimating the parameters of sinusoids and decaying sinusoids in noise,” *IEEE Instrumentation & Measurement Magazine*, vol. 14, no. 2, pp. 8–14, 2011.
- Brockwell, P. J. and Davis, R. A., *Time series: theory and methods*. Springer Science & Business Media, 2013.
- Cadzow, J., “High performance spectral estimation—a new arma method,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 5, pp. 524–529, 1980.
- Chen, W. and Stegen, G., “Experiments with maximum entropy power spectra of sinusoids,” *Journal of Geophysical Research*, vol. 79, no. 20, pp. 3019–3022, 1974.
- Chong, C.-Y., “Hierarchical estimation,” in *Proc. MIT/ONR Workshop on C3*, 1979, pp. 205–220.
- Cramér, H., “Mathematical methods of statistics,” vol. 9, 1999.
- Fisher, R. A., “Theory of statistical estimation,” in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 22, no. 5. Cambridge University Press, 1925, pp. 700–725.
- Fortunati, S., Gini, F., Greco, M. S., and Richmond, C. D., “Performance bounds for parameter estimation under misspecified models: Fundamental findings and applications,” *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 142–157, 2017.
- Gersch, W., “Estimation of the autoregressive parameters of a mixed autoregressive moving-average time series,” *IEEE Transactions on Automatic Control*, vol. 15, no. 5, pp. 583–588, 1970.
- Govaers, F. and Koch, W., “Distributed kalman filter fusion at arbitrary instants of time,” in *2010 13th International Conference on Information Fusion*. IEEE, 2010, pp. 1–8.
- Grambsch, P. *et al.*, “Sequential sampling based on the observed fisher information to guarantee the accuracy of the maximum likelihood estimator,” *The Annals of Statistics*, vol. 11, no. 1, pp. 68–77, 1983.
- Grigoriu, M., “A spectral representation based model for monte carlo simulation,” *Probabilistic Engineering Mechanics*, vol. 15, no. 4, pp. 365–370, 2000.

- Huber, P. J. *et al.*, “Robust estimation of a location parameter,” *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.
- Jenkins, G. M. and Watts, D. G., “Spectral analysis,” 1968.
- Kay, S., “Noise compensation for autoregressive spectral estimates,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 3, pp. 292–303, 1980.
- Kay, S. and Makhoul, J., “On the statistics of the estimated reflection coefficients of an autoregressive process,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 31, no. 6, pp. 1447–1455, 1983.
- Kay, S., “Representation and generation of non-gaussian wide-sense stationary random processes with arbitrary psds and a class of pdfs,” *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3448–3458, 2010.
- Kay, S., “Poisson maximum likelihood spectral inference,” 2017, unpublished <http://www.ele.uri.edu/faculty/kay/New%20web/Books.htm>.
- Kay, S. and Vankayalapati, N., “Improvement of tdoa position fixing using the likelihood curvature,” *IEEE Transactions on Signal Processing*, vol. 61, no. 8, pp. 1910–1914, 2013.
- Kay, S. M., *Modern spectral estimation*. Pearson Education India, 1988.
- Kay, S. M., “Fundamentals of statistical signal processing, volume i: Estimation theory (v. 1),” *PTR Prentice-Hall, Englewood Cliffs*, 1993.
- Kingman, J. F. C., *Poisson processes*. Wiley Online Library, 1993.
- Lacoss, R. T., “Data adaptive spectral analysis methods,” *Geophysics*, vol. 36, no. 4, pp. 661–675, 1971.
- Li, J. and AlRegib, G., “Distributed estimation in energy-constrained wireless sensor networks,” *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3746–3758, 2009.
- Li, W., Jia, Y., Du, J., *et al.*, “Diffusion kalman filter for distributed estimation with intermittent observations,” in *2015 American Control Conference (ACC)*. IEEE, 2015, pp. 4455–4460.
- Liu, Y., Li, C., Tang, W. K., and Zhang, Z., “Distributed estimation over complex networks,” *Information Sciences*, vol. 197, pp. 91–104, 2012.
- Marple, L., “Resolution of conventional fourier, autoregressive, and special arma methods of spectrum analysis,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP’77*, vol. 2. IEEE, 1977, pp. 74–77.

- Rao, C. R., “Information and the accuracy attainable in the estimation of statistical parameters,” in *Breakthroughs in statistics*. Springer, 1992, pp. 235–247.
- Reinhardt, M., Noack, B., and Hanebeck, U. D., “The hypothesizing distributed kalman filter,” in *2012 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, 2012, pp. 305–312.
- Ribeiro, A. and Giannakis, G. B., “Bandwidth-constrained distributed estimation for wireless sensor networks-part i: Gaussian case,” *IEEE transactions on signal processing*, vol. 54, no. 3, pp. 1131–1143, 2006.
- Snyder, D. L. and Miller, M. I., *Random point processes in time and space*. Springer Science & Business Media, 2012.
- Stefanski, L. A. and Boos, D. D., “The calculus of m-estimation,” *The American Statistician*, vol. 56, no. 1, pp. 29–38, 2002.
- White, H., “Maximum likelihood estimation of misspecified models,” *Econometrica: Journal of the Econometric Society*, pp. 1–25, 1982.
- Xiao, J.-J., Ribeiro, A., Luo, Z.-Q., and Giannakis, G. B., “Distributed compression-estimation using wireless sensor networks,” *IEEE Signal Processing Magazine*, vol. 23, no. 4, pp. 27–41, 2006.
- Yeredor, A. and Angel, E., “Joint tdoa and fdof estimation: A conditional bound and its use for optimally weighted localization,” *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1612–1623, 2011.
- Yilmaz, Y. and Wang, X., “Sequential decentralized parameter estimation under randomly observed fisher information,” *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 1281–1300, 2013.