Fall 2016

A unified discrepancy-based approach for balancing efficiency and robustness in state-space modeling estimation, selection, and diagnosis

Nan Hu
University of Iowa

A UNIFIED DISCREPANCY-BASED APPROACH FOR BALANCING EFFICIENCY AND ROBUSTNESS IN STATE-SPACE MODELING ESTIMATION, SELECTION, AND DIAGNOSIS

by

Nan Hu

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Applied Mathematical and Computational Sciences
in the Graduate College of
The University of Iowa

December 2016

Thesis Supervisor: Professor Joseph Cavanaugh

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Nan Hu

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy degree in
Applied Mathematical and Computational Sciences
at the December 2016 graduation.

Thesis Committee: _____
Joseph Cavanaugh, Thesis Supervisor

_____
Kung-Sik Chan

_____
Brian Smith

_____
Kai Wang

_____
Eric Foster

To my mom,
for your endless support and love.

ACKNOWLEDGMENTS

From the depth of my heart, I would like to first thank my thesis advisor and great friend, Professor Joseph Cavanaugh, for his encouragement, inspiration, passion, and selfless and enormous mentoring throughout my tenure at Iowa. As my thesis advisor, Joe would always guide me in the right direction and relieve my frustration when I faced research obstacles. Further, as my teacher in several classes, Joe has proven to be a role model of how to interact with students and how to engage students in lecture. As a great friend outside of the classroom, Joe is always a good listener, and I truly enjoy our genuine conversations. I am also thankful for his encouragement whenever I doubt myself. I know that one cannot take having a fantastic Ph.D. advisor for granted, and words cannot express how grateful I am to have worked with Joe! I am confident to say that meeting Joe has been one of the best things that has happened to me at Iowa. Thank you, Joe!

I would also like to thank my committee members, Drs. Kung-Sik Chan, Brian Smith, Kai Wang, and Eric Foster, for sharing their invaluable comments and suggestions. Professor Chan's theoretical time series class piqued my interest in time series analysis, which later became the framework of my Ph.D. thesis. I would also like to thank Professor Smith for advising me on the biostatistics preceptorship project, through which I have further developed my appreciation on the applied perspective in mathematics and biostatistics. I would also express my appreciation to Professor Wang, who made me feel proud to tell people that we graduated from the same alma mater. Further, being exposed to his personality over the years has taught me the proper way to conduct myself in a professional setting. I would also like to thank Eric for giving me suggestions when my research got stuck. I thoroughly appreciate all of the encouragement that he provided. I would like to thank Professor Weimin Han and Professor Laurent Jay as well for their kind

support these years.

I would like to thank the Department of Mathematics for giving me the opportunity to serve as a teaching assistant for five years. I truly enjoyed interacting with students as a teacher. I would like to thank every one of my students for their patience and encouragement, especially in my early stages of teaching. Standing in front of the classroom and sharing my appreciation of mathematics has really been an unforgettable experience.

I have also been blessed to have remarkable friends, both in China and in the United States. Thank you for sharing the ups and downs in my life, and thank you for being the perfect distraction whenever I need to take a break from school.

Finally, I owe much to my mom and dad. Dad, thank you for shaping me into who I am today; your influence has directed me towards following a career as a mathematician, and it is one of my most important decisions. You know you will always have a place in my heart. Mom, you are the most caring and toughest person that I have ever met. Thank you for your unconditional love.

ABSTRACT

Due to its generality and flexibility, the state-space model has become one of the most popular models in modern time domain analysis for the description and prediction of time series data. The model is often used to characterize processes that can be conceptualized as "signal plus noise," where the realized series is viewed as the manifestation of a latent signal that has been corrupted by observation noise.

In the state-space framework, parameter estimation is generally accomplished by maximizing the innovations Gaussian log-likelihood. The maximum likelihood estimator (MLE) is efficient when the normality assumption is satisfied. However, in the presence of contamination, the MLE suffers from a lack of robustness. Basu, Harris, Hjort, and Jones (1998) introduced a discrepancy measure (BHHJ) with a non-negative tuning parameter that regulates the trade-off between robustness and efficiency. In this manuscript, we propose a new parameter estimation procedure based on the BHHJ discrepancy for fitting state-space models. As the tuning parameter is increased, the estimation procedure becomes more robust but less efficient. We investigate the performance of the procedure in an illustrative simulation study. In addition, we propose a numerical method to approximate the asymptotic variance of the estimator, and we provide an approach for choosing an appropriate tuning parameter in practice. We justify these procedures theoretically and investigate their efficacy in simulation studies.

Based on the proposed parameter estimation procedure, we then develop a new model selection criterion in the state-space framework. The traditional Akaike information criterion (AIC), where the goodness-of-fit is assessed by the empirical log-likelihood, is not robust to outliers. Our new criterion is comprised of a goodness-of-fit term based on the empirical BHHJ discrepancy, and a penalty term based on both the tuning parameter and the dimension of the candidate model.

We present a comprehensive simulation study to investigate the performance of the new criterion. In instances where the time series data is contaminated, our proposed model selection criterion is shown to perform favorably relative to AIC.

Lastly, using the BHHJ discrepancy based on the chosen tuning parameter, we propose two versions of an influence diagnostic in the state-space framework. Specifically, our diagnostics help to identify cases that influence the recovery of the latent signal, thereby providing initial guidance and insight for further exploration. We illustrate the behavior of these measures in a simulation study.

PUBLIC ABSTRACT

Time series data are very common in many disciplines, including finance, economics, meteorology, ecology, and epidemiology. Investigators are often interested in building statistical models to understand the mechanisms behind time series data, and to make forecasts to guide future decisions. Due to its generality and flexibility, the state-space model is one of the most popular models in modern time series analysis.

Once a model is formulated, the parameters of the model must be estimated from the data. Two salient issues in parameter estimation include the efficiency and robustness of the estimator. Efficiency refers to the principle of obtaining estimators with low variability based on the size of the sample at hand; robustness pertains to the extent to which the estimator is likely to be influenced by outlying values. We develop a parameter estimation procedure for fitting state-space models that balances efficiency and robustness. In addition, based on the estimation procedure, we present a model selection criterion in the state-space framework that is robust to outliers. Finally, we propose two diagnostics for state-space modeling to facilitate the identification of points that could substantially influence predictors generated by the fitted model.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

1.1 Motivation of the Thesis

In statistical modeling, parameter estimation is an inevitable and sometimes formidable task. Accurate model parameter estimation facilitates the characterization and the subsequent understanding of the mechanism that generates the observed data. Some commonly used point estimation techniques include maximum likelihood, method of moments, and least squares.

Due to its generality and flexibility, the state-space model has become one of the most popular models in modern time domain analysis. The model is often used to characterize processes that can be conceptualized as "signal plus noise," where the realized series is viewed as the manifestation of a latent signal that has been corrupted by observation noise.

In the state-space framework, parameter estimation is generally accomplished by maximizing the Gaussian log-likelihood based on the innovations. The maximum likelihood estimator (MLE) is efficient when the normality assumption is satisfied. However, in the presence of contamination, the MLE suffers from a lack of robustness. Specifically, the parameter estimates may be dramatically impacted by both the percentage of contaminated data points, and the extent to which these points deviate from the temporal pattern established by the remainder of the series. When the estimator is not robust, the data generating mechanism cannot be appropriately characterized. State-space predictions based on the parameter estimates are also affected.

In addition to parameter estimation, investigators also often face the problem of choosing an appropriate model to characterize the data at hand. For state-space models in particular, the selection problem frequently arises when one is uncertain

as to the appropriate degree of complexity for the signal structure. Some popular and well developed model selection criteria include the Akaike information criterion (AIC) (Akaike, 1973, 1974), the corrected Akaike information criterion (AICc) (Hurvich and Tsai, 1989), the Bayesian information criterion (BIC) (Schwarz, 1978), the Hannan-Quinn information criterion (HQC) (Hannan and Quinn, 1979), and the Takeuchi information criterion (TIC) (Takeuchi, 1976). However, all of these criteria are developed based on the use of the MLE. Since the MLE is not robust to outliers, these criteria are also impacted by unusual points in a time series.

Another challenge in statistical modeling involves the identification of cases that exhibit a high degree of influence on key inferential objectives, such as estimation and prediction. In the state-space framework, characterizing the trajectory of latent signals is generally of paramount interest. Thus, identifying cases that impact the recovery of the signals presents an important aim.

Parameter estimation, model selection, and influence diagnosis, as well as many other statistical problems, are often addressed using procedures based on *discrepancy* or *divergence* measures. A discrepancy or divergence is a functional that reflects the disparity between two probability distributions or densities. Perhaps the best known discrepancy measure is the Kullback-Leibler (K-L) discrepancy (Kullback and Leibler, 1951), which is ubiquitous in statistics due to its close connection with the likelihood function. Jeffreys (1946) proposed a symmetric version of the K-L discrepancy, which is often referred to as J-divergence. Csiszár (1963), Morimoto (1963), and Ali and Silvey (1966) simultaneously introduced and studied the ϕ-divergence family, which generalizes the K-L discrepancy and J-divergence. An important sub-family of ϕ-divergences is the class of measures studied by Cressie and Read (1984): the power-divergence family. For more examples of the ϕ-divergence family, see Chapter 1 in Pardo (2005). Of key relevance to the present work, Basu, Harris, Hjort, and Jones (1998) introduced a discrepancy measure (BHHJ) with

a non-negative tuning parameter α. The BHHJ discrepancy is a generalization of the K-L discrepancy; when the tuning parameter α converges to 0^+, the BHHJ discrepancy reduces to the K-L discrepancy.

In parameter estimation, the K-L discrepancy is closely connected to maximum likelihood estimation. In particular, the maximum likelihood estimator (MLE) is consistent for the parameter that minimizes the K-L discrepancy. Basu, Harris, Hjort, and Jones (1998) proposed a parameter estimation procedure based on the BHHJ discrepancy that controls the trade-off between robustness and efficiency. Specifically, as the tuning parameter α approaches 0^+, the estimator converges to the MLE.

Robust methods in statistics have a long history. In fact, work on statistical analyses based on the rejection of outliers can be traced back to Bernoulli (1777). Robust parameter estimation started to blossom in the 1960s. In particular, Huber (1964) presented seminal, pioneering research that formed the foundation for robust estimation theory. The M-estimation method, proposed by Huber (1964), is one of the most widely used robust estimation techniques; the basic idea is to estimate parameters based on appropriate objective functions that downweight the influence of outliers. The L-estimation method (Bickel and Lehmann, 1975) and R-estimation method (Hodges and Lehmann, 1963) serve as alternatives. Robust estimation based on mixture models has also been widely studied (Fujisawa and Eguchi, 2006; Peel and Mclachlan, 2000). Research on robust regression for heavy-tailed distributions can be found in many disciplines (Ibragimov, Ibragimov, and Walden, 2015; Takeuchi, Bengio, and Kanamori, 2002). The text of Hampel, Ronchetti, Rousseeuw, and Stahel (1986) provides a detailed and thorough presentation of classic robust statistical methods. In time series analysis, the M-estimation method can be generalized for traditional time series models; the robust filter algorithm and the robust Durbin-Levinson algorithm are additional robust inferential procedures

for time series applications (Maronna, Martin, and Yohai, 2006, Chapter 8).

The use of discrepancy measures also arises in the development of tools for model selection and diagnosis. Under certain regularity assumptions, the well-known Akaike information criterion (AIC) serves as an asymptotically unbiased estimator of the expectation of the K-L discrepancy between the fitted candidate model at hand and the true model. Mattheou, Lee, and Karagrigoriou (2009) developed a robust model selection criterion based on the BHHJ discrepancy; this criterion was later modified by Mantalos, Mattheou, and Karagrigoriou (2010). The K-L discrepancy has also been utilized to develop measures for diagnosing influential values in a data set. In the state-space framework, Cavanaugh and Johnson (1999) proposed a diagnostic based on the K-L discrepancy for the identification of cases that influence the recovery of the latent signal. Cavanaugh and Oleson (2001) generalized this diagnostic to the broader modeling problem of identifying cases that impact the recovery of missing or unobserved data.

The first part of this thesis provides a link between the BHHJ discrepancy and the estimation of state-space model parameters. Specifically, we propose a new parameter estimation procedure based on the BHHJ discrepancy for fitting state-space models. In the presence of contamination, relative to the MLE, the estimator based on our method is shown to be more robust, at the cost of a marginal loss of efficiency. We investigate the performance of the procedure in an illustrative simulation study. In addition, we propose a numerical method to approximate the asymptotic variance of the estimator, and we provide an approach for choosing an appropriate tuning parameter in practice. We justify these procedures theoretically and investigate their efficacy in simulation studies.

In the second part of the thesis, based on our robust parameter estimation procedure, we develop a model selection criterion in the state-space framework. Traditional AIC, where the goodness-of-fit is assessed by the empirical log-likelihood,

is not robust to outliers. Our new criterion is comprised of a goodness-of-fit term based on the empirical BHHJ discrepancy, and a penalty term based on both the tuning parameter and the dimension of the candidate model. We present a comprehensive simulation study to investigate the performance of the new criterion. In instances where the time series data is contaminated, our proposed model selection criterion is shown to perform favorably relative to AIC.

In the third part of the thesis, using the BHHJ discrepancy based on the chosen tuning parameter, we propose two versions of an influence diagnostic in the state-space framework. Specifically, our diagnostics help to identify cases that influence the recovery of the latent signal, thereby providing initial guidance and insight for further exploration. We illustrate the behavior of these measures in a simulation study.

1.2 Overview of the Thesis

The remainder of the thesis is organized as follows. In Chapter 2, we provide some background knowledge of the state-space model and the BHHJ discrepancy. In Chapter 3, we propose a new parameter estimation procedure based on the BHHJ discrepancy for fitting state-space models. We investigate the performance of the procedure in an illustrative simulation study. In Chapter 4, we discuss how to approximate the variance of our parameter estimator, and provide an approach for choosing an appropriate tuning parameter α in practice. These procedures are justified theoretically and investigated in simulation studies. We also present two applications to illustrate the use of our proposed estimation method. In Chapter 5, we develop and investigate a robust model selection criterion in the BHHJ discrepancy framework. Finally, in Chapter 6, we propose two versions of an influence diagnostic for state-space modeling. Chapter 7 concludes.

CHAPTER 2
BACKGROUND

In this chapter, the state-space model and the BHHJ discrepancy are presented and some of their properties are also briefly introduced. These two statistical constructs provide the foundation for the development of our new estimation procedure, model selection criterion, and influence diagnostics.

2.1 State-Space Model

Introduced by the work of Kalman (1960) and Kalman and Bucy (1961), the *state-space model* or the *dynamic linear model* is one of the most popular models for the analysis of time series. The ubiquity of the state-space framework is largely due to its generality and flexibility. Indeed, many well-known time series models, such as autoregressive (AR) models and autoregressive moving-average (ARMA) models, arise as special cases of the state-space model. In the name of the framework, "state" refers to the latent process or signal, and "space" refers to the characterization of possible and probable values of the state. The state-space model has been applied in many disciplines including finance (Zeng and Wu, 2013), economics (Shumway and Stoffer, 1982), meteorology (Bengtsson and Cavanaugh, 2008; Tandeo, Ailliot, and Autret, 2011), ecology (Buckland, Newman, Thomas, and Koesters, 2004), and epidemiology (Yang, Cavanaugh, and Zamba, 2015).

The state-space model is defined by the following two equations:

$$\boldsymbol{y}_t = A_t \boldsymbol{x}_t + \boldsymbol{v}_t, \tag{2.1}$$

$$\boldsymbol{x}_t = \Phi \boldsymbol{x}_{t-1} + \boldsymbol{\omega}_t, \tag{2.2}$$

where $\boldsymbol{y}_t$ is a $q \times 1$ observed vector, and $\boldsymbol{x}_t$ is a $p \times 1$ latent or unobserved state vector, defined for time points $t = 1, 2, ..., N$.

The first equation, called the *observation equation*, relates the observed vector $\boldsymbol{y}_t$ to the state vector $\boldsymbol{x}_t$. Specifically, the observed data $\boldsymbol{y}_t$ is expressed as a linear transformation of the unobserved state $\boldsymbol{x}_t$ with added noise $\boldsymbol{v}_t$. The $q \times p$ matrix A_t is referred to as the observation matrix, and is usually specified and fixed for each t. The $\{\boldsymbol{v}_t\}$ disturbances are $q \times 1$, and assumed to be independent and identically distributed (i.i.d.) zero-mean normal vectors with $q \times q$ covariance matrix R.

The second equation, called the *state equation*, is a vector autoregression of order one. Here, the state vector $\boldsymbol{x}_t$ is related to the previous state vector $\boldsymbol{x}_{t-1}$ through a $p \times p$ transition matrix Φ, for time points $t = 1, 2, ..., N$. The $\{\boldsymbol{\omega}_t\}$ are $p \times 1$, and assumed to be i.i.d. zero-mean normal vectors with $p \times p$ covariance matrix Q. The $\{\boldsymbol{v}_t\}$ from the observation equation and the $\{\boldsymbol{\omega}_t\}$ from the state equation are assumed to be independent.

The structure of the state-space model presumes that the observed data are comprised of two components: the true "signal" and measurement or observation "noise". For this reason, the model is often referred to as the "signal plus noise" model. More general formulations are available that could allow the inclusion of covariate series, exogenous variables in the state or the observation equations, and correlated noise processes. For details regarding such model formulations, we refer the reader to Shumway and Stoffer (2010). In this work, we only focus on the previous model structure. However, our proposed methodologies could be naturally adapted to more general state-space settings.

We denote the collection of parameters as $\boldsymbol{\Theta} = \{\boldsymbol{\mu}_0, \Sigma_0, \Phi, R, Q\}$, where $\boldsymbol{\mu}_0$ and Σ_0 are the initial mean and covariance matrix of the signal $\boldsymbol{x}_0$; Φ, R, and Q are as stated previously. At times in our development, $\boldsymbol{\Theta}$ will be assumed to have the structure of a vector. If $\boldsymbol{\Theta}$ is known, the Kalman filter, one-step predictor, and Kalman smoother are often used to recover the signal $\boldsymbol{x}_t$. These constructs are respectively denoted as $\boldsymbol{x}_t^t$, $\boldsymbol{x}_t^{t-1}$, and $\boldsymbol{x}_t^N$. The subscript represents the time index

of the current state process; the superscript represents the time index of the last observation used to recover the signal. The signal estimators are obtained from calculating the mean of the conditional density functions of $\boldsymbol{x}_t$. Specifically,

$$\boldsymbol{x}_t^t = E[\boldsymbol{x}_t|\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_t], \tag{2.3}$$

$$\boldsymbol{x}_t^{t-1} = E[\boldsymbol{x}_t|\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{t-1}], \tag{2.4}$$

$$\boldsymbol{x}_t^N = E[\boldsymbol{x}_t|\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N]. \tag{2.5}$$

The corresponding covariance matrices of the estimators are defined as

$$P_t^t = E[(\boldsymbol{x}_t - \boldsymbol{x}_t^t)(\boldsymbol{x}_t - \boldsymbol{x}_t^t)^T|\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_t], \tag{2.6}$$

$$P_t^{t-1} = E[(\boldsymbol{x}_t - \boldsymbol{x}_t^{t-1})(\boldsymbol{x}_t - \boldsymbol{x}_t^{t-1})^T|\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{t-1}], \tag{2.7}$$

$$P_t^N = E[(\boldsymbol{x}_t - \boldsymbol{x}_t^N)(\boldsymbol{x}_t - \boldsymbol{x}_t^N)^T|\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N]. \tag{2.8}$$

In addition, it can be shown that the conditional density functions of the signal $\boldsymbol{x}_t$ are normal distributions. Specifically,

$f(\boldsymbol{x}_t|\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_t)$ corresponds to $\mathcal{N}(\boldsymbol{x}_t^t, P_t^t)$;

$f(\boldsymbol{x}_t|\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{t-1})$ corresponds to $\mathcal{N}(\boldsymbol{x}_t^t, P_t^{t-1})$;

$f(\boldsymbol{x}_t|\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N)$ corresponds to $\mathcal{N}(\boldsymbol{x}_t^t, P_t^N)$.

The recursive algorithm to obtain the signal estimators $\boldsymbol{x}_t^t$, $\boldsymbol{x}_t^{t-1}$, $\boldsymbol{x}_t^N$, along the corresponding covariance matrices P_t^t, P_t^{t-1}, P_t^N, can be found in Shumway and Stoffer (2010).

The Kalman smoother incorporates information based on past, present, and future observations. Therefore, qualitatively, a plot of the overall estimated signal produced using this approach is indeed smoother than a corresponding plot produced using either the Kalman filter or the one-step predictor. Because the one-step predictor only incorporates information based on the past, the pattern for this estimator often lags the pattern for the true underlying state. However, the

one-step predictor is the most meaningful state estimation approach for forecasting future values of the signal.

In practice, $\boldsymbol{\Theta}$ is usually unknown and needs to be estimated. The estimation is typically accomplished by maximizing the likelihood function of the mutually independent innovations $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, ..., \boldsymbol{\epsilon}_N$, defined by

$$\boldsymbol{\epsilon}_t = \boldsymbol{y}_t - A_t \boldsymbol{x}_t^{t-1}, \tag{2.9}$$

where $t = 1, 2, ..., N$. It can be easily shown that

$$\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \Sigma_t = A_t P_t^{t-1} A_t^T + R). \tag{2.10}$$

Ignoring a constant, the negative of the log-likelihood function can be written as

$$-\log L(\boldsymbol{\Theta}) = \frac{1}{2}\sum_{t=1}^{N} \log |\Sigma_t(\boldsymbol{\Theta})| + \frac{1}{2}\sum_{t=1}^{N} \boldsymbol{\epsilon}_t(\boldsymbol{\Theta})^T \Sigma_t(\boldsymbol{\Theta})^{-1} \boldsymbol{\epsilon}_t(\boldsymbol{\Theta}). \tag{2.11}$$

Since the negative of the log-likelihood function is highly non-linear with respect to $\boldsymbol{\Theta}$, an analytical solution for the parameter vector that minimizes $-\log L(\boldsymbol{\Theta})$ is not available. Two popular numerical algorithms are often used to find a numerical solution to the likelihood equations: the Newton-Raphson algorithm (Ypma, 1995) and the expectation maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977).

The problem of missing values often arises in time series analysis. One attractive feature of the state-space model is its ability to accommodate missing data. Shumway and Stoffer (1982) developed the missing data modifications in the state-space setting for the implementation of the EM algorithm.

2.2 BHHJ Discrepancy

We first introduce the general definition of discrepancy in statistics. Let f and g denote two probability densities belonging to a class of densities denoted by

$\mathcal{M}$. A *discrepancy* Δ is a functional from $\mathcal{M} \times \mathcal{M}$ to $\mathbb{R}$ that satisfies the property $\Delta(g, f) \geq \Delta(g, g)$ (Linhart and Zucchini, 1986).

A discrepancy reflects the proximity between two distributions. In particular, as the two density functions f and g become "more similar," the value of the discrepancy becomes smaller. Note that a discrepancy is not necessarily a metric, since a metric requires the additional properties of non-negativity, symmetry and the triangle inequality.

Since the methodologies we develop in this thesis are built upon the BHHJ discrepancy, we next present the form of this measure. Let $f(\boldsymbol{z})$ and $g(\boldsymbol{z})$ be two densities belonging to a class $\mathcal{M}$, for a continuous random vector $\boldsymbol{z}$. For any fixed $\alpha \in \mathbb{R}^+$, the BHHJ discrepancy between g and f is defined as

$$\Delta^{\alpha}_{BHHJ}(g, f) = \int \left\{ f^{1+a}(\boldsymbol{z}) - \left(1 + \frac{1}{\alpha}\right) g(\boldsymbol{z}) f^{\alpha}(\boldsymbol{z}) + \frac{1}{\alpha} g^{1+\alpha}(\boldsymbol{z}) \right\} d\boldsymbol{z}. \tag{2.12}$$

When $\alpha \to 0^+$, it can be shown that the BHHJ discrepancy reduces to the K-L discrepancy between g and f, which is defined as

$$\Delta_{K\text{-}L}(g, f) = \int g(\boldsymbol{z}) \log \left\{ \frac{g(\boldsymbol{z})}{f(\boldsymbol{z})} \right\} d\boldsymbol{z}. \tag{2.13}$$

In other words,

$$\lim_{\alpha \to 0^+} \Delta^{\alpha}_{BHHJ}(g, f) = \Delta_{K\text{-}L}(g, f). \tag{2.14}$$

When $\alpha = 1$, the BHHJ discrepancy is the squared L_2 distance between g and f.

It is fairly straightforward to verify that the BHHJ discrepancy is indeed a discrepancy. Again, one only needs to show that $\Delta^{\alpha}_{BHHJ}(g, f) \geq \Delta^{\alpha}_{BHHJ}(g, g)$, for any $\alpha \in \mathbb{R}^+$ and any $f, g \in \mathcal{M}$. For a detailed proof, we refer the reader to Theorem 1 in Basu, Harris, Hjort, and Jones (1998).

CHAPTER 3
DISCREPANCY-BASED PARAMETER ESTIMATION FOR BALANCING EFFICIENCY AND ROBUSTNESS IN FITTING STATE-SPACE MODELS

In this chapter, we develop a new parameter estimation procedure that balances efficiency and robustness in fitting state-space models. We discuss in detail how the estimation procedure, originally proposed for the simple i.i.d. setting, may be adapted to the state-space framework. In addition, we feature a two-part simulation study to illustrate the trade-off between efficiency and robustness of the proposed estimator.

3.1 Proposed Parameter Estimation Procedural Development

This section presents a detailed development of the new parameter estimation method in the state-space modeling framework. In the i.i.d. setting, Basu, Harris, Hjort, and Jones (1998) originally proposed an estimation procedure based on an empirical variant of the discrepancy that they had introduced. They showed that as the tuning parameter α is increased, their estimator becomes more robust but less efficient. In our work, we adapt their estimation procedure to the state-space setting.

We first revisit the BHHJ discrepancy defined in equation (2.12). Following the same notation as in Basu, Harris, Hjort, and Jones (1998), we assume f corresponds to a parametric density, denoted as $f(\boldsymbol{z}|\boldsymbol{\Theta})$, that represents a parametric candidate model; we assume $g(\boldsymbol{z})$ corresponds to a density that represents the true data generating model.

For the state-space model, one major problem with applying the estimation procedure proposed by Basu, Harris, Hjort, and Jones (1998) is that the observations $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N$ are not i.i.d. To address this problem, we turn our focus to the

innovations $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, ..., \boldsymbol{\epsilon}_N$ instead of the observations.

In the state equation (2.2), assuming that the eigenvalues of the transition matrix Φ are all less than one in absolute value suffices to ensure that the variances of the innovations will stabilize as the time index increases (Jazwinski, 1970; Anderson and Moore, 1979). Since the expectation of the innovations is $\mathbf{0}$, the marginal distributions of the innovations will also stabilize. Under the eigenvalue assumption, although the observations are not i.i.d., we can loosely treat the innovations as i.i.d. if the length of the time series is reasonably large. For simplicity, we retain our eigenvalue assumption throughout this thesis. However, the assumption to ensure stability of the innovations may be weakened; see Harvey (1991).

Next, we introduce some additional notation for the development of our estimation procedure. For a fixed collection of parameters $\boldsymbol{\Theta}$, we denote the density functions of the innovations as $f_1(\cdot|\boldsymbol{\Theta}), f_2(\cdot|\boldsymbol{\Theta}), ..., f_N(\cdot|\boldsymbol{\Theta})$ and denote their corresponding cumulative distribution functions as $F_1(\cdot|\boldsymbol{\Theta}), F_2(\cdot|\boldsymbol{\Theta}), ..., F_N(\cdot|\boldsymbol{\Theta})$. Based on the eigenvalue assumption, we may conclude that the innovations $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, ..., \boldsymbol{\epsilon}_N$ converge in distribution to some random vector $\boldsymbol{\epsilon}$. In other words,

$$F(\boldsymbol{x}|\boldsymbol{\Theta}) = \lim_{t\to\infty} F_t(\boldsymbol{x}|\boldsymbol{\Theta}), \tag{3.1}$$

for all $\boldsymbol{x}$ at which F is continuous, where F is the cumulative distribution function of $\boldsymbol{\epsilon}$. We denote the density function of $\boldsymbol{\epsilon}$ as $f(\cdot|\boldsymbol{\Theta})$. Recall that $f_1, f_2, ..., f_N, f, F_1, F_2, ..., F_N, F$ are all parametric, under the assumption of a particular state-space candidate model. We then assume that there exists a true density function for $\boldsymbol{\epsilon}$, denoted as $g(\cdot)$.

We may now establish the new estimation procedure based on the BHHJ discrepancy in the state-space framework. In the formulation of the BHHJ discrepancy, we choose the stabilized innovation $\boldsymbol{\epsilon}$ as the random variable of interest. For any $\alpha \in \mathbb{R}^+$, the BHHJ discrepancy between the true density function for $\boldsymbol{\epsilon}$, $g(\cdot)$, and

the model-based parametric density function of $\boldsymbol{\epsilon}$, $f(\cdot|\boldsymbol{\Theta})$ is

$$\Delta^{\alpha}_{BHHJ}(g, f) = \int \left\{ f^{1+\alpha}(\boldsymbol{z}|\boldsymbol{\Theta}) - \left(1 + \frac{1}{\alpha}\right) g(\boldsymbol{z}) f^{\alpha}(\boldsymbol{z}|\boldsymbol{\Theta}) + \frac{1}{\alpha} g^{1+\alpha}(\boldsymbol{z}) \right\} d\boldsymbol{z}. \quad (3.2)$$

Ideally, we would like to find the parametric model (and thus the collection of parameters) that is "closest" to the generating model, in the sense of minimizing the BHHJ discrepancy. In other words, we seek to find the minimizer $\overline{\boldsymbol{\Theta}}$ defined as

$$\overline{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} \Delta^{\alpha}_{BHHJ}(g, f). \quad (3.3)$$

However, finding $\overline{\boldsymbol{\Theta}}$ is not possible in practice, since we do not have access to the true density function g. (Indeed, if we did know g, there would be no need to postulate a candidate model, let alone estimate parameters.) The minimizer $\overline{\boldsymbol{\Theta}}$ defined in equation (3.3) is often called the *pseudo true parameter*.

In discrepancy-based parameter estimation, the true model g is conventionally replaced with the empirical density function; the resulting measure is known as the *empirical discrepancy*. Given the observations $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N\}$, instead of minimizing the BHHJ discrepancy, we minimize the BHHJ empirical discrepancy, which may be defined as

$$\Delta^{\alpha}_{BHHJ}(\boldsymbol{Y}, \boldsymbol{\Theta}) = \frac{1}{N} \sum_{t=1}^{N} \int f_t^{1+\alpha}(\boldsymbol{z}|\boldsymbol{\Theta}) d\boldsymbol{z} - \left(1 + \frac{1}{\alpha}\right) \left\{ \frac{1}{N} \sum_{t=1}^{N} f_t^{\alpha}(\boldsymbol{\epsilon}_t|\boldsymbol{\Theta}) \right\}. \quad (3.4)$$

Since we loosely treat $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, ..., \boldsymbol{\epsilon}_N$ as i.i.d., the first term in $\Delta^{\alpha}_{BHHJ}(\boldsymbol{Y}, \boldsymbol{\Theta})$, the average of N integrals, serves as an approximation for $\int f^{1+\alpha}(\boldsymbol{z}|\boldsymbol{\Theta}) d\boldsymbol{z}$. Also,

$$-\left(1 + \frac{1}{\alpha}\right) \left\{ \frac{1}{N} \sum_{t=1}^{N} f_t^{\alpha}(\boldsymbol{\epsilon}_t|\boldsymbol{\Theta}) \right\} \quad (3.5)$$

provides an approximation for

$$\int -\left(1 + \frac{1}{\alpha}\right) g(\boldsymbol{z}) f^{\alpha}(\boldsymbol{z}|\boldsymbol{\Theta}) d\boldsymbol{z}. \quad (3.6)$$

Lastly, in the definition of the BHHJ empirical discrepancy, note that we discard the last term in equation (3.2), since this term does not involve parameters.

For $t \in \{1, 2, ..., N\}$, evaluating $\int f_t^{1+\alpha}(\boldsymbol{z})d\boldsymbol{z}$ in $\Delta^{\alpha}_{BHHJ}(\boldsymbol{Y}, \boldsymbol{\Theta})$ could be potentially problematic. However, if f_t is assumed to be a multivariate normal distribution function, we can obtain an explicit form of the integral based on the following theorem.

Theorem 1. *If $f(\boldsymbol{x})$ is an n dimensional multivariate normal distribution function with $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, then for any $\alpha \in \mathbb{R}^+$,*

$$\int f^{1+\alpha}(\boldsymbol{x})d\boldsymbol{x} = \left(\frac{1}{1+\alpha}\right)^{\frac{n}{2}} (2\pi)^{-\frac{n\alpha}{2}} |\Sigma|^{-\frac{\alpha}{2}}. \tag{3.7}$$

Proof. See Appendix. □

Applying Theorem 1, and using the expression (2.10) for the density function of $\boldsymbol{\epsilon}_t$, we rewrite the BHHJ empirical discrepancy in equation (3.4) as

$$\begin{aligned}
\Delta^{\alpha}_{BHHJ}(\boldsymbol{Y}, \boldsymbol{\Theta}) &= \frac{1}{N}\sum_{t=1}^{N}\left(\frac{1}{1+\alpha}\right)^{\frac{q}{2}} (2\pi)^{-\frac{q\alpha}{2}} |\Sigma_t|^{-\frac{\alpha}{2}} \\
&\quad - \left(1+\frac{1}{\alpha}\right)\frac{1}{N}\sum_{t=1}^{N}\left[\frac{1}{(2\pi)^{\frac{q}{2}}|\Sigma_t|^{\frac{1}{2}}}\right]^{\alpha} \exp\left(-\frac{1}{2}\alpha\, \boldsymbol{\epsilon}_t^T \Sigma_t^{-1}\boldsymbol{\epsilon}_t\right) \\
&= \frac{1}{N}\sum_{t=1}^{N}\left(\frac{1}{1+\alpha}\right)^{\frac{q}{2}} (2\pi)^{-\frac{q\alpha}{2}} |\Sigma_t|^{-\frac{\alpha}{2}} \\
&\quad - \left(1+\frac{1}{\alpha}\right)\frac{1}{N}\sum_{t=1}^{N}\left[\frac{1}{(2\pi)^{\frac{q}{2}}|A_t P_t^{t-1} A_t^T + R|^{\frac{1}{2}}}\right]^{\alpha} \times \\
&\quad \exp\left[-\frac{1}{2}\alpha \boldsymbol{\epsilon}_t^T \left(A_t P_t^{t-1} A_t^T + R\right)^{-1}\boldsymbol{\epsilon}_t\right] \\
&= \frac{1}{N}\sum_{t=1}^{N}\left(\frac{1}{1+\alpha}\right)^{\frac{q}{2}} (2\pi)^{-\frac{q\alpha}{2}} |A_t P_t^{t-1} A_t^T + R|^{-\frac{\alpha}{2}} \\
&\quad - \left(1+\frac{1}{\alpha}\right)\frac{1}{N}\sum_{t=1}^{N}\left[\frac{1}{(2\pi)^{\frac{q}{2}}|A_t P_t^{t-1} A_t^T + R|^{\frac{1}{2}}}\right]^{\alpha} \times \\
&\quad \exp\left[-\frac{1}{2}\alpha \left(\boldsymbol{y}_t - A_t \boldsymbol{x}_t^{t-1}\right)^T \left(A_t P_t^{t-1} A_t^T + R\right)^{-1} \left(\boldsymbol{y}_t - A_t \boldsymbol{x}_t^{t-1}\right)\right].
\end{aligned} \tag{3.8}$$

With an explicit form of the BHHJ empirical discrepancy, we can then find

the *BHHJ minimum discrepancy estimator* (MDE), defined as

$$\widehat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} \Delta^{\alpha}_{BHHJ}(\boldsymbol{Y}, \boldsymbol{\Theta}). \tag{3.9}$$

Since the expression for the BHHJ empirical discrepancy is highly non-linear in the parameters $\boldsymbol{\Theta}$, obtaining an analytic expression of $\widehat{\boldsymbol{\Theta}}$ is not possible. The traditional approach of estimating the parameters in state-space models, i.e., finding the MLE, presents the same challenge. The two commonly used numerical approaches for computing the MLE are the Newton-Raphson algorithm and the EM algorithm. However, since the EM algorithm is specifically developed for maximizing the log-likelihood function, it is not applicable in our setting. We thus choose to employ the Newton-Raphson algorithm to obtain $\widehat{\boldsymbol{\Theta}}$. In each iteration of the algorithm, the intermediate parameter estimates are used in the implementation of the Kalman filter to obtain the value of the objective function. A detailed description of how to apply the Newton-Raphson algorithm in the state-space setting can be found in Section 6.3 in Shumway and Stoffer (2010). The same method is relevant in the present setting, with the negative of the log-likelihood simply replaced by the empirical BHHJ discrepancy. We again emphasize that as α goes to 0^+, the MDE $\widehat{\boldsymbol{\Theta}}$ converges to the MLE.

Next, we illustrate how the tuning parameter α regulates the balance between efficiency and robustness in the BHHJ MDE. The asymptotic properties of the estimator will be discussed in the next chapter.

3.2 Efficiency vs. Robustness

In this section, we start by using data from a simulated example to show how the MLE suffers from a lack of robustness when a state space time series is corrupted by influential or outlying values. On the other hand, our proposed estimator, the BHHJ MDE based on a positive tuning parameter α, is shown to be reasonably

robust.

Through a subsequent two-part simulation study, we then show that as the tuning parameter α is increased, the BHHJ MDE typically becomes more robust. However, as α is increased, the BHHJ MDE also becomes less efficient. Following the conventions in Basu, Harris, Hjort, and Jones (1998), we are mostly interested in relatively small α values, ranging from 0 to 1, since the efficiency loss is usually excessive when $\alpha > 1$.

In the following example, we generate a time series of length 100 ($N = 100$) from the following state-space model:

$$y_t = x_t + v_t, \tag{3.10}$$

$$x_t = \phi x_{t-1} + \omega_t, \tag{3.11}$$

where $\phi = 0.8$, $v_t \sim N(0, R = \sigma_v^2 = 1)$, and $\omega_t \sim N(0, Q = \sigma_\omega^2 = 1)$. Notice that the eigenvalue assumption is satisfied ($0.8 < 1$), and that the length of the time series is sufficiently large, thereby ensuring the validity of our proposed estimation procedure.

Our goal is to estimate $\mathbf{\Theta} = (\mu_0, \Sigma_0, \phi, \sigma_v, \sigma_\omega)^T$. Since the absolute value of ϕ is less than 1, the state process defined in equation (3.11) is stationary. We fix the initial state mean μ_0 as 0. The initial state variance Σ_0 can be expressed as $\sigma_\omega^2/(1-\phi^2)$, based on the stationarity of the state process. Essentially, we are only estimating $\mathbf{\Theta} = (\phi, \sigma_v, \sigma_\omega)^T$.

In applying the Newton-Raphson algorithm to the state-space setting, we obtain an MLE of $\mathbf{\Theta} = (\phi, \sigma_v, \sigma_\omega)^T = (0.81, 0.87, 0.85)^T$. The MLE is reasonably accurate due to the following: (i) the length of the time series is reasonably large, and (ii) there is no data contamination in the series.

Next, we randomly shift three observations in the time series by adding a constant perturbation with a magnitude of 8. Figure 3.1 shows the original simulated

series and the perturbed series.

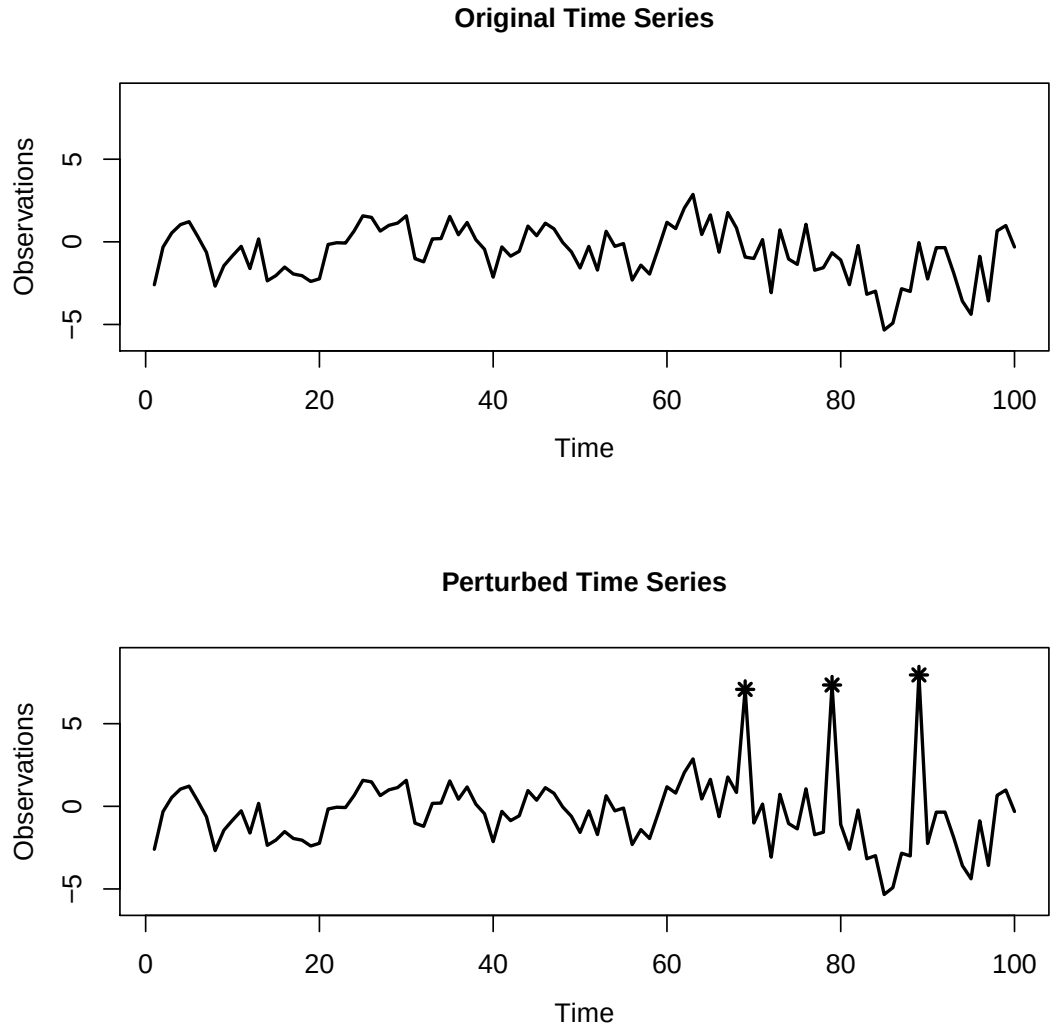

Figure 3.1: Top: original simulated time series. Bottom: perturbed time series; observations $69, 79$ and 89 are shifted upwards from their original values by an additive factor of 8. The three perturbed points are marked with "$*$".

As we see from the bottom plot in Figure 3.1, the three anomalous observations clearly stand out from the other observations. These three outliers influence the MLE dramatically; in fact, under this contamination, the MLE of $\boldsymbol{\Theta} = (\phi, \sigma_v, \sigma_\omega)^T$ is $(0.64, 1.66, 1.04)^T$. Compared to the MLE based on the original series, there is a

noticeable increase in the estimate of σ_v based on the perturbed series – the three outliers exaggerate the variability of the observation noise v_t.

Applying our proposed estimation method to the perturbed series with α values ranging from 0 to 0.5, we obtain more robust estimates. By "robust," we mean that the MDEs are close to the MLEs based on the original series:

$$(\hat{\phi}, \hat{\sigma}_v, \hat{\sigma}_\omega)^T = (0.81, 0.87, 0.85)^T.$$

Table 3.1 features the parameter estimates based on different α values. Note that the BHHJ MDEs for ϕ and σ_ω do not vary appreciably as α changes, and all of the estimates are relatively close to the corresponding MLEs based on the original series. Thus, we mainly focus on the BHHJ MDE for the standard deviation of the observation noise, σ_v. As we see from the table, as α increases, the estimators become more robust. Note that when $\alpha = 0.001$, the BHHJ MDE is very close to the MLE (i.e., the MDE corresponding to $\alpha = 0$).

Table 3.1: BHHJ MDE illustration based on simulated data.

	Parameter Estimate		
α Value	$\hat{\phi}$	$\hat{\sigma}_v$	$\hat{\sigma}_\omega$
0 (MLE)	0.64	1.66	1.04
0.001	0.64	1.65	1.05
0.1	0.78	1.43	0.79
0.2	0.81	1.28	0.73
0.3	0.79	1.19	0.77
0.4	0.76	1.07	0.89
0.5	0.75	0.91	1.02

With any non-robust estimator, in the presence of contamination, the data

generating mechanism cannot be appropriately characterized. Furthermore, any inferential procedure that is based upon the non-robust estimator will also be affected. In the state-space setting, prediction of the unobserved states is often an important inferential objective. One major reason to prefer robust estimators is that they will naturally lead to robust prediction. Next, we illustrate that our robust BHHJ MDE leads to robust state prediction for the simulated contaminated data, whereas the MLE based on the perturbed series results in substantially altered state prediction.

Figure 3.2 shows the one-step state predictors based on three different sets of parameter estimates. For the purpose of illustration, we choose $\alpha = 0.5$ for the computation of the BHHJ MDE. The black line represents the perturbed observations with the three anomalous data marked with "$*$". The blue line represents a plot of the one-step predictors using the MLE based on the original series; the red line represents a plot of the one-step predictors using the MLE based on the perturbed series; the green line represents a plot of the one-step predictors using the BHHJ MDE with $\alpha = 0.5$ based on the perturbed series. We clearly see that the one-step state predictors based on the MLE are greatly affected by the three outliers. In particular, the outliers cause the one-step predictors to be smoother, due to the inflated variance estimate of the observation noise. On the other hand, the one-step predictors based on the BHHJ MDE are barely affected by the outliers; in fact, the difference between the BHHJ MDE-based one-step predictors using the perturbed series and the MLE-based one-step predictors using the original series is almost unnoticeable. This example illustrates that robust BHHJ MDE leads to robust time series prediction.

Next, we run a two-part Monte Carlo simulation study to respectively show that (i) as the tuning parameter α increases, the BHHJ MDE becomes more robust, and (ii) as the tuning parameter α increases, the BHHJ MDE becomes less efficient.

Simulation Study, Part I: We conduct the first part of the simulation study

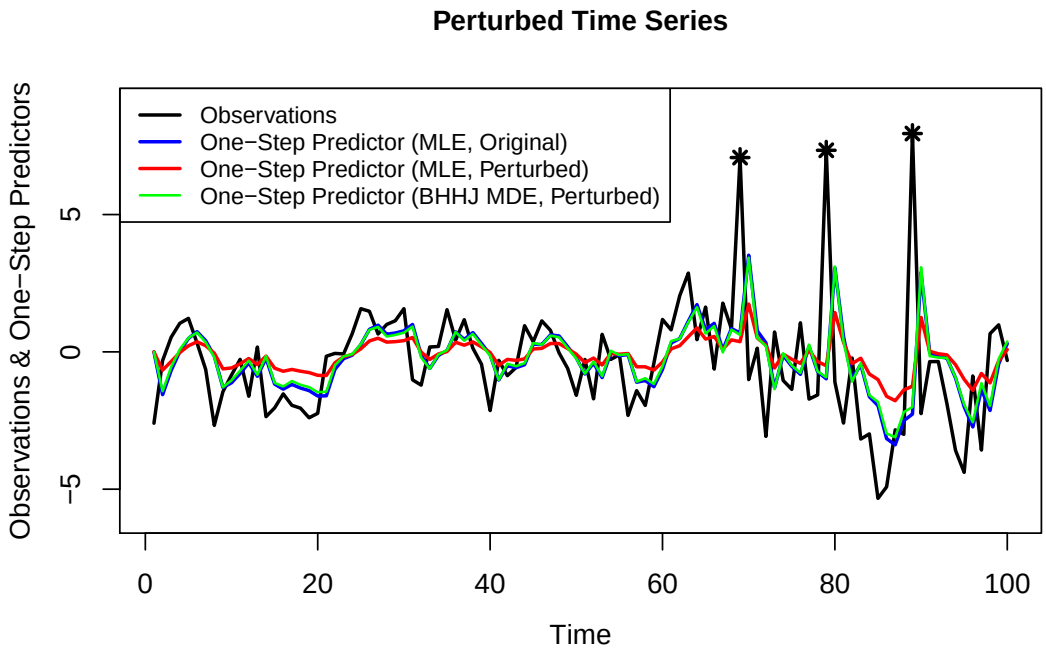

Figure 3.2: Plot of one-step predictors. Note that the predictors using the BHHJ MDE with $\alpha = 0.5$ based on the perturbed series closely follow the predictors using the MLE based on the unperturbed series (green and blue, respectively).

to investigate the robustness of the BHHJ MDE when the time series is corrupted by randomly perturbing certain observations. In each trial, the state-space time series is generated by equations (3.10) and (3.11). The length of each time series is 100. For each series, 3 random observations are perturbed based on an additive shift with a magnitude of 8. The number of iterations in this study is $10,000$.

Table 3.2 shows the empirical Monte Carlo mean of the BHHJ MDE for α ranging from 0 (corresponding to the MLE) to 1. Note that the Monte Carlo mean of the AR parameter estimate and the mean of the state noise standard deviation estimate do not vary much as α increases, whereas the Monte Carlo mean of the observation noise standard deviation estimate varies dramatically as α increases. We thus evaluate the robustness of the BHHJ MDE by focusing on the Monte Carlo mean of the standard deviation estimate for the observation noise. Table 3.2

shows that as α increases, the BHHJ MDE becomes more robust to outliers, since the Monte Carlo mean of the standard deviation estimate in the observation noise gets closer to the true observation noise standard deviation, $\sigma_v = 1$.

Table 3.2: BHHJ MDE robustness illustration.

	Empirical Monte Carlo Mean		
α Value	mean($\hat{\phi}$)	mean($\hat{\sigma}_v$)	mean($\hat{\sigma}_\omega$)
0 (MLE)	0.76	1.86	1.07
0.1	0.75	1.66	0.94
0.2	0.74	1.45	0.86
0.3	0.74	1.34	0.83
0.4	0.74	1.28	0.84
0.5	0.74	1.24	0.85
0.6	0.74	1.20	0.87
0.7	0.73	1.17	0.89
0.8	0.73	1.14	0.90
0.9	0.73	1.12	0.92
1.0	0.73	1.10	0.93

Simulation Study, Part II: In the second part of the simulation study, we investigate the efficiency loss in the BHHJ MDE as α increases. We again use the generating model defined by equations (3.10) and (3.11). In the i.i.d. setting, in estimating parameters from certain popular density families, Basu, Harris, Hjort, and Jones (1998) showed that the BHHJ MDE becomes less efficient as α increases. In investigating the efficiency loss property, they assumed that the generating model g belongs to the parametric family $\{f(\cdot|\boldsymbol{\Theta})\}$, which essentially implies there are no

outliers. Thus, in this part of the simulation study, we do not introduce perturbations.

Table 3.3 shows the empirical Monte Carlo variance of the BHHJ MDE for α ranging from 0 to 1. For each of the three parameter estimates $(\hat{\phi}, \hat{\sigma}_v, \hat{\sigma}_\omega)$, the variances of the BHHJ MDE increase monotonically as α increases. This shows that the BHHJ MDE becomes less efficient as α increases. Note that when $\alpha = 0$, the BHHJ MDE reduces to the MLE. Since the MLE is asymptotically efficient under correct model specification, the fact that the MLE has the smallest empirical variance should not be surprising.

The two parts of this simulation study show that the tuning parameter α in the BHHJ MDE regulates the robustness and efficiency of the estimator. By increasing α, we generally obtain more robust estimators; however, we pay the price of losing efficiency in the estimator. In Chapter 4, we propose an approach to automatically choose an appropriate α value in practice.

Table 3.3: BHHJ MDE efficiency loss illustration.

	Empirical Monte Carlo Variance		
α Value	var($\hat{\phi}$)	var($\hat{\sigma}_v$)	var($\hat{\sigma}_\omega$)
0 (MLE)	0.0106	0.0606	0.0611
0.1	0.0113	0.0727	0.0647
0.2	0.0116	0.0755	0.0667
0.3	0.0120	0.0790	0.0688
0.4	0.0124	0.0836	0.0718
0.5	0.0130	0.0879	0.0754
0.6	0.0135	0.0928	0.0788
0.7	0.0139	0.0967	0.0822
0.8	0.0144	0.1008	0.0868
0.9	0.0148	0.1054	0.0901
1.0	0.0153	0.1095	0.0937

CHAPTER 4
ESTIMATION ISSUES

Since the asymptotic variance of a parameter estimator is generally of interest for large-sample inferential procedures, in the first part of this chapter, we propose a numerical method to estimate the asymptotic variance of the BHHJ MDE in the state-space setting.

In the presence of contamination, sacrificing some efficiency to obtain a more robust parameter estimator is a reasonable trade off. However, for the BHHJ MDE, choosing an appropriate tuning parameter α that balances robustness and efficiency is not a trivial task. The second part of this chapter provides a data-driven approach for automatically choosing a tuning parameter α in practice.

At the end of this chapter, through two applications based on real data, we illustrate in detail how to apply the BHHJ estimation procedure, including how to obtain BHHJ MDE with the chosen α and its asymptotic variance.

4.1 Asymptotic Variance of the BHHJ MDE

In this section, we propose a numerical method to estimate the asymptotic variance of the BHHJ MDE in the state-space modeling framework. The asymptotic variance of any parameter estimator plays an important role in large-sample statistical inference. In the relatively simplistic setting of fitting models for univariate applications where the outcomes belong to certain distributional families (e.g., normal distribution, Student's t-distribution, gamma distribution), Warwick (2002) proposed a data-based method to obtain the asymptotic variance of the BHHJ MDE and derived explicit variance forms. Due to the complexity of the state-space framework, however, the method proposed by Warwick (2002) is not applicable in fitting state-space models. Therefore, we propose an alternate numerical method to

estimate the asymptotic variance of the BHHJ MDE, specifically designed for the state-space setting.

We first state a theorem from Basu, Harris, Hjort, and Jones (1998) regarding the asymptotic properties of the BHHJ MDE. The theorem is based on a set of regularity assumptions that include boundary conditions and functional smoothness. For detailed information, we refer the reader to conditions (A1)-(A5) in the technical report of Basu, Harris, Hjort, and Jones (1998).

Theorem 2. *Under certain regularity conditions, as* $N \to \infty$*, the BHHJ MDE* $\widehat{\boldsymbol{\Theta}}$ *has the following asymptotic properties:*

(i) $\widehat{\boldsymbol{\Theta}}$ *is consistent for* $\overline{\boldsymbol{\Theta}}$ *(defined in equation (3.3)), and*

(ii) $N^{1/2}(\widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{\Theta}})$ *is asymptotically multivariate normal with mean* $\mathbf{0}$ *and covariance matrix* $J^{-1}KJ^{-1}$*, where* $J = J(\overline{\boldsymbol{\Theta}})$ *and* $K = K(\overline{\boldsymbol{\Theta}})$ *are given by*

$$J = \int u(\boldsymbol{z}|\overline{\boldsymbol{\Theta}})u^T(\boldsymbol{z}|\overline{\boldsymbol{\Theta}})f^{1+\alpha}(\boldsymbol{z}|\overline{\boldsymbol{\Theta}})d\boldsymbol{z} + \int \{i(\boldsymbol{z}|\overline{\boldsymbol{\Theta}}) - \alpha u(\boldsymbol{z}|\overline{\boldsymbol{\Theta}})u^T(\boldsymbol{z}|\overline{\boldsymbol{\Theta}})\}\{g(\boldsymbol{z}) - f(\boldsymbol{z}|\overline{\boldsymbol{\Theta}})\}f^{\alpha}(\boldsymbol{z}|\overline{\boldsymbol{\Theta}})d\boldsymbol{z}, \tag{4.1}$$

$$K = \int u(\boldsymbol{z}|\overline{\boldsymbol{\Theta}})u^T(\boldsymbol{z}|\overline{\boldsymbol{\Theta}})f^{2\alpha}(\boldsymbol{z}|\overline{\boldsymbol{\Theta}})g(\boldsymbol{z})d\boldsymbol{z} - \boldsymbol{\xi}\boldsymbol{\xi}^T \tag{4.2}$$

with $\boldsymbol{\xi} = \int u(\boldsymbol{z}|\overline{\boldsymbol{\Theta}})f^{\alpha}(\boldsymbol{z}|\overline{\boldsymbol{\Theta}})g(\boldsymbol{z})d\boldsymbol{z}$*,* $u(\boldsymbol{z}|\boldsymbol{\Theta}) = \partial \log f(\boldsymbol{z}|\boldsymbol{\Theta})/\partial\boldsymbol{\Theta}$*, and* $i(\boldsymbol{z}|\boldsymbol{\Theta}) = -\partial^2 \log f(\boldsymbol{z}|\boldsymbol{\Theta})/\partial\boldsymbol{\Theta}^2$*.*

Proof. See Appendix A.1 in Warwick (2002). □

According to Theorem 2, the evaluation of the asymptotic variance of the BHHJ MDE requires knowledge of the true density function g and the pseudo true parameter $\overline{\boldsymbol{\Theta}}$. Warwick (2002) proposed a data-based method to approximate the asymptotic variance of the BHHJ MDE; the basic idea is to replace the true density g with the empirical density, and to replace the true parameter $\overline{\boldsymbol{\Theta}}$ with the BHHJ MDE $\widehat{\boldsymbol{\Theta}}$. Note that the evaluation of the integrals in equations (4.1) and (4.2) requires explicit forms of the score function $u(\boldsymbol{z})$ and the information function

$i(\boldsymbol{z})$. Unfortunately, when fitting state-space models, explicit analytic expressions do not exist for either the score function or the information function. Granted, we could numerically approximate the score function or the information function at any given $\boldsymbol{z}$ and any given $\boldsymbol{\Theta}$ in the state-space framework; however, the evaluation of the integrals in equations (4.1) and (4.2) would still be very challenging.

To avoid the evaluation of integrals in equations (4.1) and (4.2), we propose a numerical method to approximate the asymptotic variance of the BHHJ MDE, specifically designed for state-space modeling. We start by revisiting the BHHJ empirical discrepancy defined in equation (3.4). For simplicity, we denote

$$\delta(\boldsymbol{\Theta}) := \left(\frac{1}{1+\alpha}\right) \Delta^{\alpha}_{BHHJ}(\boldsymbol{Y}, \boldsymbol{\Theta}), \tag{4.3}$$

emphasizing the role of the parameter vector $\boldsymbol{\Theta}$. We multiply $\Delta^{\alpha}_{BHHJ}(\boldsymbol{Y}, \boldsymbol{\Theta})$ by the constant $\left(\frac{1}{1+\alpha}\right)$ so that our approach will parallel to the development in Warwick (2002).

Expanding $\delta(\widehat{\boldsymbol{\Theta}})$ in a Taylor series around $\overline{\boldsymbol{\Theta}}$, we have

$$\delta(\widehat{\boldsymbol{\Theta}}) = \delta(\overline{\boldsymbol{\Theta}}) + \delta^{'}(\overline{\boldsymbol{\Theta}})(\widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{\Theta}}) + \frac{1}{2}(\widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{\Theta}})^{T}\delta^{''}(\overline{\boldsymbol{\Theta}})(\widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{\Theta}}) + \cdots . \tag{4.4}$$

Taking the first derivative on both sides yields

$$\delta^{'}(\widehat{\boldsymbol{\Theta}}) \approx \delta^{'}(\overline{\boldsymbol{\Theta}}) + (\widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{\Theta}})^{T}\delta^{''}(\overline{\boldsymbol{\Theta}}). \tag{4.5}$$

Note that $\delta^{'}(\widehat{\boldsymbol{\Theta}}) = \mathbf{0}$ since $\widehat{\boldsymbol{\Theta}}$ is the minimizer of function δ. So

$$\sqrt{N}(\widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{\Theta}}) \approx \sqrt{N}\left[-\delta^{''}(\overline{\boldsymbol{\Theta}})\right]^{-1}\left[\delta^{'}(\overline{\boldsymbol{\Theta}})\right]^{T}. \tag{4.6}$$

The asymptotic variance of $\sqrt{N}(\widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{\Theta}})$ can then be expressed as

$$\mathrm{var}_g\left(\sqrt{N}(\widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{\Theta}})\right) \approx N\left[-E_g\left(\delta^{''}(\overline{\boldsymbol{\Theta}})\right)\right]^{-1}\mathrm{var}_g\left(\delta^{'}(\overline{\boldsymbol{\Theta}})\right)\left[-E_g\left(\delta^{''}(\overline{\boldsymbol{\Theta}})\right)\right]^{-1}. \tag{4.7}$$

Therefore, to estimate $\mathrm{var}_g\left(\sqrt{N}(\widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{\Theta}})\right)$, we need to devise approaches to estimate the two terms $E_g\left(\delta^{''}(\overline{\boldsymbol{\Theta}})\right)$ and $\mathrm{var}_g\left(\delta^{'}(\overline{\boldsymbol{\Theta}})\right)$. Next, we propose two separate

approaches to respectively estimate the two terms.

1) For $E_g\left(\delta^{''}(\overline{\mathbf{\Theta}})\right)$, we can use $\delta^{''}(\widehat{\mathbf{\Theta}})$ as an estimator, where the Hessian matrix evaluated at $\widehat{\mathbf{\Theta}}$ is approximated via the finite difference numerical method.

2) For $\text{var}_g\left(\delta^{'}(\overline{\mathbf{\Theta}})\right)$, we first define

$$\delta_t(\mathbf{\Theta}) = \frac{1}{1+\alpha}\left\{\frac{1}{N}\int f_t^{1+\alpha}(\boldsymbol{z}|\mathbf{\Theta})d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}f_t^{\alpha}(\boldsymbol{\epsilon}_t|\mathbf{\Theta})\right\}\right\}. \tag{4.8}$$

It is easy to see that $\delta(\mathbf{\Theta}) = \sum_{t=1}^{N}\delta_t(\mathbf{\Theta})$, and thus $\delta^{'}(\mathbf{\Theta}) = \sum_{t=1}^{N}\delta_t^{'}(\mathbf{\Theta})$. Since we treat $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, ..., \boldsymbol{\epsilon}_N$ as i.i.d., then $\delta_1^{'}(\mathbf{\Theta}), \delta_2^{'}(\mathbf{\Theta}), ..., \delta_N^{'}(\mathbf{\Theta})$ are also i.i.d. To proceed with our development, we first present the following theorem.

Theorem 3. *Under the assumption of* $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, ..., \boldsymbol{\epsilon}_N$ *being i.i.d.,* $E_g\left(\delta_t^{'}(\overline{\mathbf{\Theta}})\right) = \mathbf{0}$*, for all* $t = 1, 2, ..., N$*.*

Proof. See Appendix. □

We rewrite $\text{var}_g\left(\delta^{'}(\overline{\mathbf{\Theta}})\right)$ as

$$\begin{aligned}
\text{var}_g\left(\delta^{'}(\overline{\mathbf{\Theta}})\right) &= \text{var}_g\left(\sum_{t=1}^{N}\delta_t^{'}(\overline{\mathbf{\Theta}})\right) \\
&= \sum_{t=1}^{N}\text{var}_g\left(\delta_t^{'}(\overline{\mathbf{\Theta}})\right) \\
&= \sum_{t=1}^{N}\left[E_g\left(\delta_t^{'}(\overline{\mathbf{\Theta}})^T\delta_t^{'}(\overline{\mathbf{\Theta}})\right) - E_g\left(\delta_t^{'}(\overline{\mathbf{\Theta}})\right)E_g\left(\delta_t^{'}(\overline{\mathbf{\Theta}})^T\right)\right] \\
&= \sum_{t=1}^{N}E_g\left(\delta_t^{'}(\overline{\mathbf{\Theta}})^T\delta_t^{'}(\overline{\mathbf{\Theta}})\right),
\end{aligned} \tag{4.9}$$

where the last step in equation (4.9) makes use of Theorem 3. For each $t \in \{1, 2, ..., N\}$, we use $\delta_t^{'}(\widehat{\mathbf{\Theta}})^T\delta_t^{'}(\widehat{\mathbf{\Theta}})$ to estimate $E_g\left(\delta_t^{'}(\overline{\mathbf{\Theta}})^T\delta_t^{'}(\overline{\mathbf{\Theta}})\right)$. Again, the gradient function evaluated at $\widehat{\mathbf{\Theta}}$ is approximated using the finite difference numerical method.

To summarize the preceding development, our proposed estimator for

$$\text{var}_g\left(\sqrt{N}(\widehat{\mathbf{\Theta}} - \overline{\mathbf{\Theta}})\right) \tag{4.10}$$

is given by

$$\left[-\delta^{''}(\widehat{\boldsymbol{\Theta}})\right]^{-1}\left[N\sum_{t=1}^{N}\delta_t^{'}(\widehat{\boldsymbol{\Theta}})^T\delta_t^{'}(\widehat{\boldsymbol{\Theta}})\right]\left[-\delta^{''}(\widehat{\boldsymbol{\Theta}})\right]^{-1}. \tag{4.11}$$

Note that we move the constant N into the middle term to be consistent with the form of the asymptotic variance in Theorem 2. Specifically in our approach,

$$\widehat{J}(\widehat{\boldsymbol{\Theta}}) := \left[-\delta^{''}(\widehat{\boldsymbol{\Theta}})\right]^{-1} \tag{4.12}$$

provides an estimator of $J(\overline{\boldsymbol{\Theta}})$, and

$$\widehat{K}(\widehat{\boldsymbol{\Theta}}) := N\sum_{t=1}^{N}\delta_t^{'}(\widehat{\boldsymbol{\Theta}})^T\delta_t^{'}(\widehat{\boldsymbol{\Theta}}) \tag{4.13}$$

provides an estimator of $K(\overline{\boldsymbol{\Theta}})$.

We point out that our proposed asymptotic variance estimation method for the BHHJ MDE is more computationally expensive than the variance estimation method proposed by Warwick (2002). The finite difference numerical calculation of both the Hessian matrix $\delta^{''}(\widehat{\boldsymbol{\Theta}})$ and the gradient vector $\delta_t^{'}(\widehat{\boldsymbol{\Theta}})$ is fairly computationally burdensome. However, due to the complex nature of the state-space model, the evaluation of the variance estimate in Warwick (2002) is not possible in the present setting. Our proposed asymptotic variance estimator for the BHHJ MDE is specifically designed for state-space modeling. Next, we show that our estimator, as defined in equation (4.11), is reasonably accurate in larger sample settings.

We present a simulation study to illustrate the efficacy of our asymptotic variance estimator. Our study shows that our method produces variance estimates that are on par with the Monte Carlo empirical variance.

The generating model for each replication is

$$y_t = x_t + v_t, \tag{4.14}$$

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \omega_t, \tag{4.15}$$

where $\phi_1 = 0.99$, $\phi_2 = -0.8$, $v_t \sim N(0, \sigma_v^2 = 1)$, and $\omega_t \sim N(0, \sigma_\omega^2 = 1)$. We can easily express the generating model in the general state-space form specified by equations (2.1) and (2.2). Specifically, we define the state vector by $\boldsymbol{x}_t = (x_t, x_{t-1})^T$ and the transition matrix by

$$\Phi = \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix}. \tag{4.16}$$

Letting $\boldsymbol{\omega}_t = (\omega_t, 0)^T$ denote the 2×1 state noise process, the state equation is

$$\boldsymbol{x}_t = \Phi \boldsymbol{x}_{t-1} + \boldsymbol{\omega}_t, \tag{4.17}$$

with $Q = \begin{bmatrix} \sigma_\omega^2 & 0 \\ 0 & 0 \end{bmatrix}$. The observation equation can be written as

$$y_t = [1,\ 0]\boldsymbol{x}_t + v_t, \tag{4.18}$$

with $R = \sigma_v^2$.

We use a more complex generating model here compared to (3.10) and (3.11) so that we may assess the variance accuracy of more parameter estimators. The initial mean of the state $\boldsymbol{\mu}_0$ is fixed as $\mathbf{0}$; the diagonal elements of the initial state variance matrix are fixed as 10, and the off-diagonal elements are fixed as 0. We emphasize that the effect of the starting values $\boldsymbol{\mu}_0$ and Σ_0 on the estimated parameters is negligible (Harvey, 1991, pages 118 to 125).

Since standard deviations are generally more directly applicable in statistical inference procedures than variances, we focus on the former in compiling our results. We present two tables in which we compare the empirical Monte Carlo means of our proposed asymptotic standard deviations with the empirical Monte Carlo standard deviations of the parameter estimates. Table 4.1 features the empirical Monte Carlo means of our proposed standard deviation estimates for the two AR parameter estimators, along with the means of our standard deviation estimates

for both the observation noise and the state noise standard deviation estimators. Table 4.2 features the corresponding empirical Monte Carlo standard deviations of the parameter estimates. We can then compare each element in Table 4.1 with the corresponding element in Table 4.2. In both tables, the tuning parameter α ranges from 0 to 1.

Table 4.1: Empirical Monte Carlo mean of our proposed asymptotic standard deviations.

	Empirical Monte Carlo Mean			
α Value	$\text{sd}(\hat{\phi}_1)$	$\text{sd}(\hat{\phi}_2)$	$\text{sd}(\hat{\sigma}_v)$	$\text{sd}(\hat{\sigma}_\omega)$
0 (MLE)	0.0949	0.0930	0.1627	0.1940
0.1	0.0968	0.0979	0.1692	0.2024
0.2	0.0995	0.0967	0.1770	0.2086
0.3	0.1050	0.1018	0.1892	0.2214
0.4	0.1088	0.1054	0.2042	0.2391
0.5	0.1112	0.1068	0.2186	0.2458
0.6	0.1138	0.1077	0.2427	0.2427
0.7	0.1143	0.1084	0.2748	0.2457
0.8	0.1184	0.1122	0.2396	0.2558
0.9	0.1247	0.1177	0.2634	0.2726
1.0	0.1297	0.1188	0.2968	0.2815

Table 4.2: Empirical Monte Carlo standard deviation of the parameter estimates.

	Empirical Monte Carlo Standard Deviation			
α Value	$\text{sd}(\hat{\phi}_1)$	$\text{sd}(\hat{\phi}_2)$	$\text{sd}(\hat{\sigma}_v)$	$\text{sd}(\hat{\sigma}_\omega)$
0 (MLE)	0.1014	0.0977	0.1637	0.2319
0.1	0.1027	0.0985	0.1648	0.2308
0.2	0.1056	0.1002	0.1721	0.2489
0.3	0.1090	0.1035	0.1783	0.2500
0.4	0.1224	0.1175	0.2099	0.2559
0.5	0.1274	0.1222	0.2188	0.2599
0.6	0.1326	0.1267	0.2283	0.2646
0.7	0.1375	0.1307	0.2426	0.2699
0.8	0.1410	0.1333	0.2513	0.2745
0.9	0.1441	0.1356	0.2556	0.2791
1.0	0.1462	0.1380	0.2590	0.2869

In general, the corresponding entries in the two tables are fairly comparable, thereby establishing the efficacy of our method for the setting at hand. In addition, the results show that as the tuning parameter α is increased, the standard deviation (and thus the variance) increases.

Again, we acknowledge that our proposed method for estimating the asymptotic variance is more computationally intensive than the method proposed by Warwick (2002). In cases where the integrals in equations (4.1) and (4.2) have explicit forms, we recommend the procedure of Warwick (2002). Specifically, Warwick (2002) provided examples of how to obtain the BHHJ MDE asymptotic variance when estimating parameters from certain univariate distributions. However, when one uses the BHHJ MDE in fitting state-space models, our proposed method serves

as an effective approach for approximating the large-sample variance of the MDE.

4.2 Determination of α

In Chapter 3, we proposed an estimation method that balances robustness and efficiency in fitting state-space models. In the previous section, we also proposed an approach for estimating the asymptotic variance of the BHHJ MDE. The tuning parameter α regulates how robust one expects the BHHJ MDE to be and how much efficiency loss one might tolerate. In this section, we propose an approach for automatically choosing an appropriate tuning parameter α in practice. We illustrate the validity of the procedure through simulation studies.

Before introducing our approach for choosing α, we first acknowledge that there is no universally accepted method for choosing the value of a tuning parameter. Basu, Harris, Hjort, and Jones (1998) gave some general directions on how to determine a suitable value for α. The following quotes are particularly relevant:

1) "One way of selecting it is to fix the efficiency loss, at the ideal parametric model employed, at some low level, like five or ten percent..."

2) "Other ways could in some practical applications involve prior notions of the extent of contamination of the model..."

Approach 1) could be useful if the analytic relationship between α and the efficiency loss is known. For example, when we estimate the mean of a normal distribution using the appropriate parametric model, the asymptotic variance of $N^{1/2}$ (N is the size of the sample) times the BHHJ MDE of the mean is given by

$$\left(1 + \frac{\alpha^2}{1+2\alpha}\right)^{3/2} \sigma^2, \tag{4.19}$$

where σ^2 is the variance of the normal distribution (Basu, Harris, Hjort, and Jones, 1998). In this case, we could analytically solve for α, given our tolerance level of efficiency loss.

Unfortunately, for state-space models, there is no explicit expression for the efficiency loss in terms of α. In fact, although we see the trend in efficiency loss in Table 3.3, 4.1, and 4.2, we do not know the exact relationship between α and the efficiency loss. Of course, for a particular state-space model, one could simulate the efficiency loss based on a relatively fine grid for α. However, in practice, choosing a reference state-space model also presents a difficult challenge.

Approach 2), could potentially make use of the breakdown point of the BHHJ MDE. A detailed explanation of the breakdown point can be found in Hampel, Ronchetti, Rousseuw, and Stahel (1986). Intuitively, the breakdown point is defined as the maximum proportion of anomalous data in the sample that an estimator can tolerate before the estimator becomes arbitrarily large in absolute value. When estimating the mean and the variance from a univariate normal distribution, Basu, Harris, Hjort, and Jones (1998) established in their technical report that the breakdown point of the BHHJ MDE is $\alpha/(1+\alpha)^{3/2}$. If in such a setting, one has the preliminary knowledge of the percentage of perturbation in the data, then perhaps an appropriate α could be chosen based the value of that corresponds to the breakdown point. However, obtaining precise information as to the proportion of perturbed data is often not possible. Influence diagnostics, such as the measures that will be introduced in Chapter 6, could be used as a crude determination for the proportion of contamination. However, for fitting state-space models, the breakdown point of the BHHJ MDE is still unsolved.

A common way of choosing a tuning parameter in statistics is to use the *mean squared error* (MSE). The MSE provides the foundation for the criterion we develop for selecting an α. The MSE is a comprehensive measure of accuracy that takes into account both the robustness and the efficiency of the estimator. Specifically, if $\widehat{\boldsymbol{\Theta}}$ is an estimator of $\boldsymbol{\Theta}$, then

$$\mathrm{MSE}(\widehat{\boldsymbol{\Theta}}) = E(\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|^2)$$

$$= \text{tr}(\text{var}(\widehat{\boldsymbol{\Theta}})) + \|\text{bias}(\widehat{\boldsymbol{\Theta}})\|^2, \tag{4.20}$$

where "tr" stands for the trace of a matrix. The MSE is comprised of two components: $\text{tr}(\text{var}(\widehat{\boldsymbol{\Theta}}))$ reflects the efficiency of the estimator, and $\|\text{bias}(\widehat{\boldsymbol{\Theta}})\|^2$ reflects the robustness of the estimator. In Section 4.1, we discussed how to estimate the asymptotic variance of the BHHJ MDE in fitting state-space models. However, $\|\text{bias}(\widehat{\boldsymbol{\Theta}})\|^2$ requires the knowledge of the true parameter $\boldsymbol{\Theta}$. Hong and Kim (2001) proposed a method for choosing an α based on $\text{tr}(\text{var}(\widehat{\boldsymbol{\Theta}}))$ only. Note that $\text{tr}(\text{var}(\widehat{\boldsymbol{\Theta}}))$ monotonically increases only under correct model specification; when there is contamination in the data, $\text{tr}(\text{var}(\widehat{\boldsymbol{\Theta}}))$ usually achieves the minimum at a positive α. Their method, however, does not incorporate information that reflects robustness. Warwick (2005) suggested using the BHHJ MDE with $\alpha = 1$ as a surrogate for the true parameter $\boldsymbol{\Theta}$, but also acknowledged that the method did not always work well.

We propose a method for choosing an α that is specifically designed for the state-space framework. In the development of our approach, rather than focusing on the original parameters from the state-space model, we turn our focus to the mean and variance parameters of the standardized innovations. For simplicity, we assume that observations for the state-space series are scalars throughout the procedural development. Under no data contamination and under a correct model specification, the standardized innovations are distributed as $N(0, 1)$ (Shumway and Stoffer, 2010, Section 6.7); the standard normal distribution thus provides us with the true parameter values in evaluating the MSE, since the true parameter vector $\boldsymbol{\Theta}$ is associated with the ideal model specification.

Assume the length of the time series is N, and recall that the collection of parameters in the state-space model is $\boldsymbol{\Theta} = \{\boldsymbol{\mu}_0, \Sigma_0, \Phi, R, Q\}$. Denote the BHHJ MDE of $\boldsymbol{\Theta}$ as $\widehat{\boldsymbol{\Theta}} = \left\{\widehat{\boldsymbol{\mu}}_0, \widehat{\Sigma}_0, \widehat{\Phi}, \widehat{R}, \widehat{Q}\right\}$.

Next, we outline our procedure for choosing α.

1. Calculate the BHHJ MDE of $\boldsymbol{\Theta}$ with $\alpha = 1$, and obtain the corresponding

innovations $\epsilon_1, \epsilon_2, ..., \epsilon_N$ through the Kalman filter. Then standardize the innovations by defining $e_t = \Sigma_t^{-1/2}\epsilon_t$, where $\Sigma_t = A_t P_t^{t-1}(\widehat{\boldsymbol{\Theta}})A_t^T + \widehat{R}(\widehat{\boldsymbol{\Theta}})$, for $t = 1, 2, ..., N$.

2. The task is then switched to estimating the mean and the variance of the standardized innovations, assuming they are from a normal distribution $N(\mu, \sigma^2)$. Under no data contamination, the true parameter values are $\mu = 0$ and $\sigma = 1$. For a fixed α, calculate the estimated asymptotic variances of the BHHJ MDE of the mean μ and the standard deviation σ, respectively, using the method proposed by Warwick (2002). Denote the estimated asymptotic variances of the BHHJ MDE of μ and σ as $\text{var}(\widehat{\mu})$ and $\text{var}(\widehat{\sigma})$, respectively.

3. Repeat Step 2 for α ranging from 0 to 1 on a relatively fine grid.

4. Choose the α with the smallest $\text{MSE}((\widehat{\mu}, \widehat{\sigma})^T) = \text{var}(\widehat{\mu}) + \text{var}(\widehat{\sigma}) + (\widehat{\mu} - 0)^2 + (\widehat{\sigma} - 1)^2$.

In Step 1, the reason why we use the BHHJ MDE with $\alpha = 1$ to obtain the innovations is that we aim to accommodate any contamination in the series. Since any substantial perturbations in the observations will be reflected in the innovations, we are compelled to use a "large" value of α to ensure that the MDE is sufficiently robust.

In Step 2, the way we estimate μ and σ is to minimize the empirical BHHJ discrepancy, defined as

$$\Delta^{\alpha}_{BHHJ}(\boldsymbol{E}; \mu, \sigma) = \int f^{1+\alpha}(x|\mu, \sigma)dx - \left(1 + \frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(e_t|\mu, \sigma)\right\}, \quad (4.21)$$

where $\boldsymbol{E} = \{e_1, e_2, ..., e_N\}$, and $f(\cdot|\mu, \sigma)$ is the normal density function with mean μ and standard deviation σ. The explicit form of $\Delta^{\alpha}_{BHHJ}(\boldsymbol{E}; \mu, \sigma)$ is obtained from the following theorem.

Theorem 4. *If* $\boldsymbol{E} = \{e_1, e_2, ..., e_N\}$ *are i.i.d. observations from* $N(\mu, \sigma^2)$*, then the empirical BHHJ discrepancy defined in equation (4.21) can be expressed as*

$$\begin{aligned}\Delta^{\alpha}_{BHHJ}(\boldsymbol{E}; \mu, \sigma) =& (1+\alpha)^{-\frac{1}{2}}(2\pi)^{-\frac{\alpha}{2}}\sigma^{-\alpha} \\ &- \left(\frac{\alpha+1}{\alpha}\right)(2\pi)^{-\frac{\alpha}{2}}\sigma^{-\alpha}\left\{\frac{1}{N}\sum_{t=1}^{N}\exp\left[-\frac{\alpha(e_t-\mu)^2}{2\sigma^2}\right]\right\}. \quad (4.22)\end{aligned}$$

Proof. The proof is fairly straightforward via the application of Theorem 1. □

In Step 4, the true mean parameter μ and the true standard deviation parameter σ, 0 and 1 respectively, are known. Specifically, under a correct model specification, $\{e_t\}$ are i.i.d. and $e_t \sim N(0,1)$, for $t = 1, 2, ..., N$. Here, the "true" parameter corresponds to the parameter value assuming a correct model specification, which precludes data contamination. The squared bias gauges the robustness of the MDE. For the purpose of comparison, in the following discussion, we denote the objective function in Step 4 as $\text{MSE}^{\mu,\sigma}_{\text{std}}$, where the subscript "std" stands for the standardized innovations, and the superscript "μ, σ" denotes the parameters of interest.

Since our proposed approach in choosing an α requires the estimation of the asymptotic variances of the BHHJ MDE of μ and σ, we provide the following theorem. The estimated asymptotic variances are calculated using the method proposed by Warwick (2002). Since Warwick (2002) did not present the explicit variance forms for $\widehat{\mu}$ and $\widehat{\sigma}$ in the case where both μ and σ are unknown parameters from a univariate normal distribution, we present the explicit variance forms in the following theorem.

Theorem 5. *Assume* $\boldsymbol{E} = \{e_1, e_2, ..., e_N\}$ *are i.i.d. observations from* $N(\mu, \sigma^2)$*. Denote the BHHJ MDE with* α *for* μ *and* σ *as* $\widehat{\mu}$ *and* $\widehat{\sigma}$*, respectively. Based on the method in Warwick (2002), the estimated asymptotic variance of* $\sqrt{N}$ *times the*

BHHJ MDE $\widehat{\mu}$ *is* $\widehat{J}_\mu^{-1}\widehat{K}_\mu\widehat{J}_\mu^{-1}$, *where*

$$\widehat{J}_\mu = (2\pi)^{-\frac{\alpha}{2}}\widehat{\sigma}^{-(2+\alpha)}\Bigg\{(\alpha+1)^{-\frac{1}{2}}\widehat{\mu}^2\widehat{\sigma}^{-2} + \frac{1}{N}\sum_{t=1}^{N}\left[(1-\alpha\widehat{\sigma}^{-2}(e_t-\widehat{\mu})^2)\exp\left(-\frac{\alpha(e_t-\widehat{\mu})^2}{2\widehat{\sigma}^2}\right)\right]\Bigg\}, \tag{4.23}$$

and

$$\begin{aligned}\widehat{K}_\mu =& \frac{1}{N}\sum_{t=1}^{N}\left[(2\pi)^{-\alpha}\widehat{\sigma}^{-2\alpha-4}\exp\left(-\frac{\alpha(e_t-\widehat{\mu})^2}{\widehat{\sigma}^2}\right)(e_t-\widehat{\mu})^2\right] \\ & - \frac{1}{N^2}\left\{\sum_{t=1}^{N}\left[(2\pi)^{-\frac{\alpha}{2}}\widehat{\sigma}^{-(\alpha+2)}\exp\left(-\frac{\alpha(e_t-\widehat{\mu})^2}{2\widehat{\sigma}^2}\right)(e_t-\widehat{\mu})\right]\right\}^2. \end{aligned} \tag{4.24}$$

The estimated asymptotic variance of $\sqrt{N}$ *times the BHHJ MDE* $\widehat{\sigma}$ *is* $\widehat{J}_\sigma^{-1}\widehat{K}_\sigma\widehat{J}_\sigma^{-1}$, *where*

$$\begin{aligned}\widehat{J}_\sigma =& (1+\alpha)^{\frac{1}{2}}(2\pi)^{-\frac{\alpha}{2}}\widehat{\sigma}^{-(\alpha+2)} \\ & \left(1-2\widehat{\sigma}^{-2}(\alpha+1)^{-1}(\widehat{\mu}^2+\widehat{\sigma}^2)+\widehat{\sigma}^{-4}(\alpha+1)^{-2}(\widehat{\mu}^4+6\widehat{\mu}^2\widehat{\sigma}^2+3\widehat{\sigma}^4)\right) \\ & -(2\pi)^{-\frac{\alpha}{2}}\widehat{\sigma}^{-(\alpha+2)}(1+\alpha)^{-\frac{1}{2}}\left(-1+3\widehat{\sigma}^{-2}(\alpha+1)^{-1}(\widehat{\mu}^2+\widehat{\sigma}^2)\right) \\ & +\frac{1}{N}\sum_{t=1}^{N}\Bigg[\left((-\widehat{\sigma}^{-2}+3\widehat{\sigma}^{-4}(e_t-\widehat{\mu})^2)-\alpha(-\widehat{\sigma}^{-1}+\widehat{\sigma}^{-3}(e_t-\widehat{\mu})^2)^2\right) \\ & \qquad (2\pi)^{-\frac{\alpha}{2}}\widehat{\sigma}^{-\alpha}\exp\left(-\frac{\alpha(e_t-\widehat{\mu})^2}{2\widehat{\sigma}^2}\right)\Bigg], \end{aligned} \tag{4.25}$$

and

$$\begin{aligned}\widehat{K}_\sigma =& \frac{1}{N}\sum_{t=1}^{N}\left[(2\pi)^{-\alpha}\widehat{\sigma}^{-2\alpha}\exp\left(-\frac{\alpha(e_t-\widehat{\mu})^2}{\widehat{\sigma}^2}\right)\left(-\frac{1}{\widehat{\sigma}}+\frac{(e_t-\widehat{\mu})^2}{\widehat{\sigma}^3}\right)^2\right] \\ & -\frac{1}{N^2}\left[\sum_{t=1}^{N}(2\pi)^{-\frac{\alpha}{2}}\widehat{\sigma}^{-\alpha}\exp\left(-\frac{\alpha(e_t-\widehat{\mu})^2}{2\widehat{\sigma}^2}\right)\left(-\frac{1}{\widehat{\sigma}}+\frac{(e_t-\widehat{\mu})^2}{\widehat{\sigma}^3}\right)\right]^2. \end{aligned} \tag{4.26}$$

Proof. See Appendix. □

We also provide alternate approaches for defining the objective function for choosing an α. Two variants defined in the following discussion are essentially the approaches proposed by Hong and Kim (2001) and Warwick (2002). Of note, none

of the variants standardize the innovations obtained from using the BHHJ MDE with $\alpha = 1$.

- In the first variant, we estimate the mean μ and the standard deviation σ of the innovations. The true σ value is replaced by the BHHJ MDE with $\alpha = 1$: $\widehat{\sigma}_{\alpha=1}$. The objective function is defined as

$$\text{MSE}((\widehat{\mu}, \widehat{\sigma})^T) := \text{var}(\widehat{\mu}) + \text{var}(\widehat{\sigma}) + (\widehat{\mu} - 0)^2 + (\widehat{\sigma} - \widehat{\sigma}_{\alpha=1})^2. \tag{4.27}$$

  This is essentially the approach proposed by Warwick (2002). We denote the relevant objective function as $\text{MSE}^{\mu,\sigma}_{\alpha=1}$.

- The second variant only evaluates the MSE of the BHHJ MDE for the mean μ of the innovations. The objective function is defined as

$$\text{MSE}(\widehat{\mu}) := \text{var}(\widehat{\mu}) + (\widehat{\mu} - 0)^2. \tag{4.28}$$

  We denote the relevant objective function as MSE^{μ}.

- The third variant only considers the asymptotic variances of $\widehat{\mu}$ and $\widehat{\sigma}$. Specifically, the objective function is defined as

$$\text{var}(\widehat{\mu}; \widehat{\sigma}) := \text{var}(\widehat{\mu}) + \text{var}(\widehat{\sigma}). \tag{4.29}$$

  This is essentially the approach proposed by Hong and Kim (2001). We denote the relevant objective function as $\text{var}^{\mu,\sigma}$.

- The fourth variant may only be employed in simulation studies. The true μ value is still chosen as 0, but the true σ value is replaced by the MLE based on the unperturbed data. We denote the objective function for this idealized method as $\text{MSE}^{\mu,\sigma}_{\text{true}}$.

We now have five approaches for choosing an α in fitting state-space models, including our original approach and the four variants described in the preceding. We

will explore and illustrate the behavior of all these approaches through simulation studies.

We point out that there are two factors that could contribute to the severity of any contamination in the data: the magnitude of the perturbation (i.e., the extent to which the anomalous points deviate from the values arising under the ideal data generating mechanism) and the percentage of the perturbation (i.e., the percentage of such anomalous points). In the first simulation study, we fix the percentage of the perturbation and explore the behavior of the choice of α as the magnitude of the perturbation increases. In the second simulation study, we fix the magnitude of the perturbation, and explore the behavior of the choice of α as the percentage of the perturbation increases.

In each Monte Carlo iteration, we first generate data from equations (3.10) and (3.11), with the length of the time series being 100. We then randomly perturb values in the time series according to the desired magnitude and percentage. The number of replications for both simulation studies is 500. We then calculate the mean of the choices of α over the 500 replications.

The results are summarized in Table 4.3 and Table 4.4. Table 4.3 features the average choice of α for different magnitudes of perturbation, based on the five proposed approaches. Table 4.3 features the average choice of α for different percentages of perturbation, based on the five approaches.

As we can see from Table 4.3 and Table 4.4, the choices of α using $\text{MSE}^{\mu,\sigma}_{\text{std}}$ and $\text{MSE}^{\mu,\sigma}_{\alpha=1}$ are very similar. Essentially, by using the objective function $\text{MSE}^{\mu,\sigma}_{\text{std}}$ to choose an α, we are trusting the robustness of the BHHJ MDE of $\boldsymbol{\Theta} = \{\boldsymbol{\mu}_0, \Sigma_0, \Phi, R, Q\}$ with $\alpha = 1$; by using the objective function $\text{MSE}^{\mu,\sigma}_{\alpha=1}$, we are trusting the robustness of the BHHJ MDE of σ with $\alpha = 1$. Since these two criteria are developed based on the same robustness principle, the similarity of choices for α should not come as surprising. We note that the criterion $\text{MSE}^{\mu,\sigma}_{\text{true}}$ should be more accurate

Table 4.3: Change in the magnitude of the perturbation.

	Average Choice of α				
Perturbation	$\text{MSE}^{\mu,\sigma}_{\text{std}}$	$\text{MSE}^{\mu,\sigma}_{\alpha=1}$	MSE^{μ}	$\text{var}^{\mu,\sigma}$	$\text{MSE}^{\mu,\sigma}_{\text{true}}$
No Perturbation	0.2940	0.2931	0.1794	0.0517	0.1639
5% of 3	0.3899	0.4385	0.2927	0.1279	0.5139
5% of 5	0.4776	0.5141	0.2638	0.3298	0.6786
5% of 8	0.5312	0.5568	0.1801	0.3568	0.6702
5% of 15	0.4607	0.4651	0.3040	0.2693	0.4457
5% of 20	0.4288	0.4278	0.3084	0.2159	0.3388

Table 4.4: Change in the percentage of the perturbation.

	Average Choice of α				
Perturbation	$\text{MSE}^{\mu,\sigma}_{\text{std}}$	$\text{MSE}^{\mu,\sigma}_{\alpha=1}$	MSE^{μ}	$\text{var}^{\mu,\sigma}$	$\text{MSE}^{\mu,\sigma}_{\text{true}}$
No Perturbation	0.2940	0.2931	0.1794	0.0517	0.1639
3% of 8	0.5056	0.5311	0.1790	0.2934	0.5863
8% of 8	0.4611	0.4996	0.2148	0.4413	0.6741
10% of 8	0.4475	0.4804	0.2541	0.4956	0.6512
15% of 8	0.5369	0.5664	0.4514	0.6297	0.7296
20% of 8	0.7218	0.7727	0.7155	0.6785	0.8811
30% of 8	0.8834	0.9445	0.9213	0.4841	0.9776

in choosing an α; however, the criterion is only accessible in simulation studies. In fact, the patterns of the choices of α between $\text{MSE}^{\mu,\sigma}_{\text{std}}$, $\text{MSE}^{\mu,\sigma}_{\alpha=1}$, and $\text{MSE}^{\mu,\sigma}_{\text{true}}$ are very similar in general.

We can also see from the two tables that the choice of α using MSE^{μ} follows a similar pattern as with the other MSE criteria. However, MSE^{μ} tends to choose

slightly smaller α values compared to other MSE criteria; this agrees with the simulation results in Warwick (2002).

The criterion $\text{var}^{\mu,\sigma}$, which only considers the efficiency of the estimator, performs quite well in choosing a very small α when there is no perturbation in the data. This is expected because we know the BHHJ MDE is the most efficient when $\alpha = 0$, under no data contamination; this principle is illustrated in Section 3.2. Yet while this criterion will generally choose a very small α when there is no perturbation, the criterion does not take into consideration of the robustness of the estimator. Ignoring robustness could be problematic in choosing an α in the presence of data contamination.

Our intuition leads us to believe that as the degree of contamination increases, the choice of α should also increase. Table 4.4 confirms our intuition – all of the MSE criteria typically choose a larger α as the percentage of perturbation increases. In particular, when the percentage of the perturbation is 30%, across all of the criteria, the choice of α is very close to 1. According to Table 4.3, however, for most of the criteria, as the magnitude of the perturbation increases, the choice of α increases to a plateau, and then starts to decrease. Warwick (2002) also uncovered a similar phenomenon in simulation studies. Although the phenomenon may initially seem counterintuitive, in certain settings, it can be reconciled using the form of the influence function. For a detailed explanation, see page 89 of Warwick (2002).

Of the five approaches we present for choosing value of α, we lean towards the method based on $\text{MSE}^{\mu,\sigma}_{\text{std}}$. This approach is arguably the most defensible of those considered, since the objective function incorporates both the variance and the bias, and for the bias, utilizes reference parameter values based on the true mean and standard deviation of the standardized innovations under a properly specified state-space model.

We point out that any objective function based on the MSE is essentially a loss

function, refecting both the variance and the bias of an estimator. However, there is no consensus on defining an objective function for choosing an α. For example, if the robustness of the estimator is the main concern, then one might want to place a larger weight on the bias component in defining the objective function; if on the other hand, efficiency is more important to the investigator, then the variance component could be more dominant in the objective function. In this section, we have provided an approach along with some variants for choosing an α, based purely on the data. However, the problem of choosing an α is still relatively open. Ultimately, the investigator should employ a method for the determination of α that is somewhat governed by the context of the application.

4.3 Applications

In this section, we illustrate how to apply our methodologies to two real time series. In each application, we will use our proposed method to choose an α automatically, to obtain the BHHJ MDE, and to calculate the corresponding asymptotic standard errors of the BHHJ MDE.

4.3.1 Birth Rate Application

The first application serves as an illustrative example where one major outlier could be easily detected and corrected, due to the nature of the phenomenon that gives rise to the time series. We can thus evaluate the robustness of our estimation procedure by comparing the BHHJ MDE based on the original time series and the MLE based on the series with the one major outlier corrected. Further, we consider a subset of the time series with fewer outliers, aiming to further validate the robustness of our estimation procedure.

The time series features the daily average numbers of newborns in the United States from January 1, 1969, to December 31, 1988. Every data point represents the

average number of newborns on a specific day over the 20 year period. Figure 4.1 shows the pattern in the number of newborns throughout the year. The data point corresponding to February 29 is an obvious outlier, since this date only exists in leap years. We use the symbol "*" to mark this point. Other outliers of a lesser magnitude seem to be present over certain holidays. For example, the birth rate seems to be lower on July 4 and on certain dates during the Christmas season.

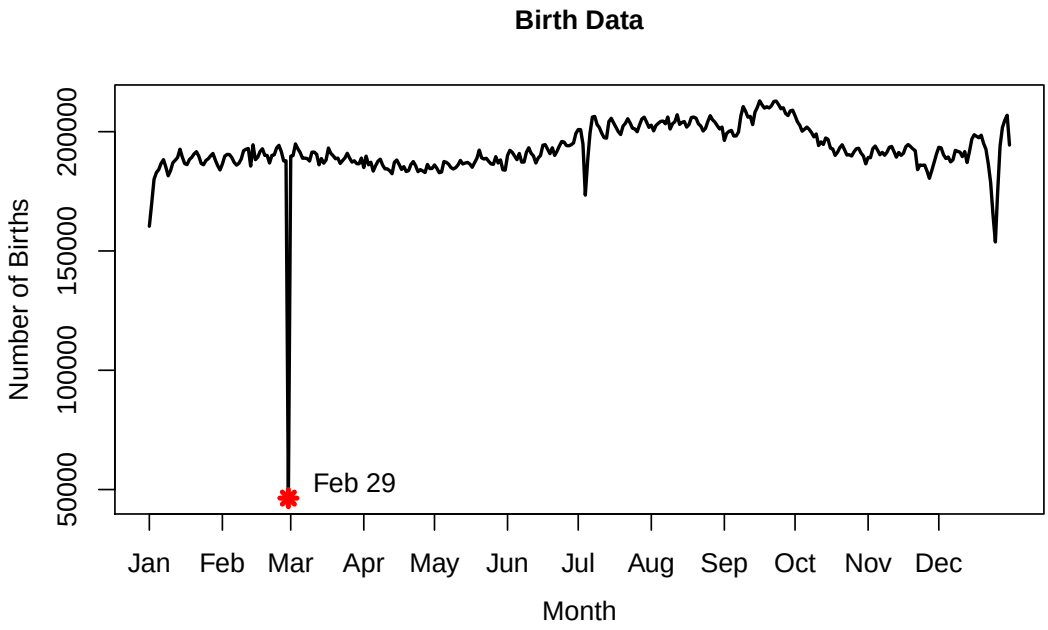

Figure 4.1: Daily average U.S. birth rate data from 1969 to 1988. February 29 only exists in leap years.

To simplify the analysis, we subtract the mean value of the observations from each observation, and for convenience of scaling, we divide each difference by a factor of 1000. We then use an autoregressive model of order 1 (AR(1)) to characterize the state process. The AR order is suggested by both the autocorrelation function (ACF) and the partial autocorrelation function (PACF) based on the raw observations. (We note that the assumption of stationarity is supported by the augmented

Dickey-Fuller test). The state-space model is thereby formulated as

$$y_t = x_t + v_t, \tag{4.30}$$

$$x_t = \phi x_{t-1} + \omega_t, \tag{4.31}$$

where y_t is the centered observation, and x_t is the latent state; $v_t \sim N(0, \sigma_v^2)$, and $\omega_t \sim N(0, \sigma_\omega^2)$, for $t = 1, 2, ..., 366$. Here, $t = 1$ corresponds to January 1, and $t = 366$ corresponds to December 31.

In what follows, we outline the stages of the initial part of the example. We first use the traditional MLE method to obtain the parameter estimates for $\boldsymbol{\Theta} = (\phi, \sigma_v, \sigma_\omega)^T$ and their corresponding standard errors. As we will see in the following, the MLE is very unreliable due to the outliers in the series; in particular the outlier corresponding to February 29. We then determine an appropriate α value for the BHHJ MDE based on the proposed method in Section 4.2. Next, we calculate the BHHJ MDE with the chosen α, and calculate the corresponding asymptotic standard errors based on the numerical method introduced in Section 4.1. We then plot the one-step predictors for the birth rate time series based on both the BHHJ MDE and MLE. Lastly, we fix the major outlier corresponding to February 29 by multiplying the value by 4, and obtain the MLE based on the corrected series. By comparing the MLE based on the corrected series and the BHHJ MDE based on the original series, we aim to validate the robustness of our estimation procedure. In all estimation procedures, we fix the initial state mean μ_0 as 0 and the initial state variance Σ_0 as 10.

To begin, we calculate the MLE and its corresponding asymptotic standard errors (see Table 4.5). From Table 4.5, we see that the standard deviation estimate for the observation noise is extremely large, compared to the standard deviation estimate for the state noise. The major outlier corresponding to February 29 invariably leads to the overestimation of the variability in the observation noise. As

we will see later in this section, after we correct the February 29 outlier, the MLE changes substantially.

Table 4.5: MLE and standard errors based on the original series.

	$\hat{\phi}$	$\hat{\sigma}_v$	$\hat{\sigma}_\omega$
MLE	0.9826	8.4953	1.3425
Standard Error	0.0128	0.3626	0.3865

Next, we apply the BHHJ estimation method and determine an appropriate α value for the BHHJ MDE. Table 4.6 shows the choices of α using four of the methods introduced in Section 4.2. For the purpose of illustration, we choose $\alpha = 0.320$, based on using $\text{MSE}^{\mu,\sigma}_{\text{std}}$. We emphasize that any alternate α value based on Table 4.6 would produce comparable statistical results.

Table 4.6: Choice of α using different methods based on the original series.

$\text{MSE}^{\mu,\sigma}_{\text{std}}$	$\text{MSE}^{\mu,\sigma}_{\alpha=1}$	MSE^{μ}	$\text{var}^{\mu,\sigma}$
0.320	0.315	0.345	0.260

Using the data-driven choice of $\alpha = 0.320$, the corresponding BHHJ MDE and its standard errors are shown in Table 4.7.

Table 4.7: BHHJ MDE with $\alpha = 0.320$ and standard errors based on the original series.

	$\hat{\phi}$	$\hat{\sigma}_v$	$\hat{\sigma}_\omega$
BHHJ MDE ($\alpha = 0.320$)	0.9435	0.0008	2.3762
Standard Error	0.01680	0.0005	0.1091

Based on the BHHJ MDE with $\alpha = 0.320$, we plot the one-step predictors for the average birth rate series from 1969 to 1988. Note that the one-step predictors tend to have a lagging effect – each predictor only uses the information up until the current time index, thus creating a "shift-to-right" effect relative to the original observations. Any unusual fluctuation in the series will be reflected in the one-step predictor at the subsequent time point. We also plot the one-step predictors based on the MLE. As we see from Figure 4.2, the one-step predictors based on the MLE are much smoother than the predictors based on the BHHJ MDE, because the MLE attributes a large proportion of the overall variability in the series to observation noise. To better present the predictors, we re-scale the range of the observations; thus, the February 29 outlier is not shown in the plot.

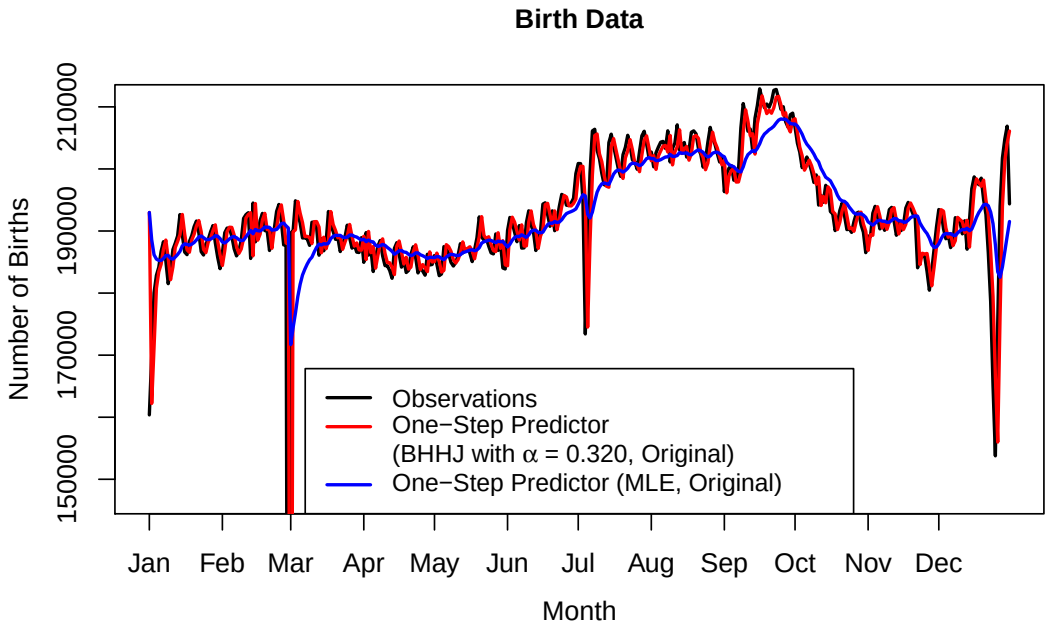

Figure 4.2: Plot of one-step predictors of the U.S. birth rate series from 1969 to 1988, based on the BHHJ MDE with $\alpha = 0.320$ and based on the MLE.

Of course, an obvious way to correct the major outlier corresponding to February 29 is simply to multiply the mean value by 4, since this date only exists once every 4 years. Based on the corrected series, the MLE and its corresponding standard errors are shown in Table 4.8.

Table 4.8: MLE and standard errors based on the corrected series.

	$\hat{\phi}$	$\hat{\sigma}_v$	$\hat{\sigma}_\omega$
MLE	0.9028	0.00001	3.5325
Standard Error	0.0220	0.3219	0.1352

Once we correct the data point for February 29, the MLE based on the resulting series is substantially different from the MLE based on the original series. This shows that the MLE approach is not robust to outliers. In particular, the MLE of the standard deviation of the observation noise is greatly attenuated after the February 29 outlier is corrected. On the other hand, the MLE based on the corrected series and the BHHJ MDE with $\alpha = 0.320$ based on the original series are quite similar, especially for the AR parameter estimates and the standard deviation estimates for the observation noise. Note that both the BHHJ MDE based on the original series and the MLE based on the corrected series assign a very small weight to the observation noise, as reflected by the standard deviation estimates. This is not too surprising. Since every observation is obtained from averaging 20 data points, intuitively, the observation noise should be negligible relative to the state noise. We plot the one-step predictors using different estimates based on the original and corrected series in Figure 4.3.

As we can see from Figure 4.3, the plot of the one-step predictors based on the BHHJ MDE with $\alpha = 0.320$ using the original series is very close to the plot of the

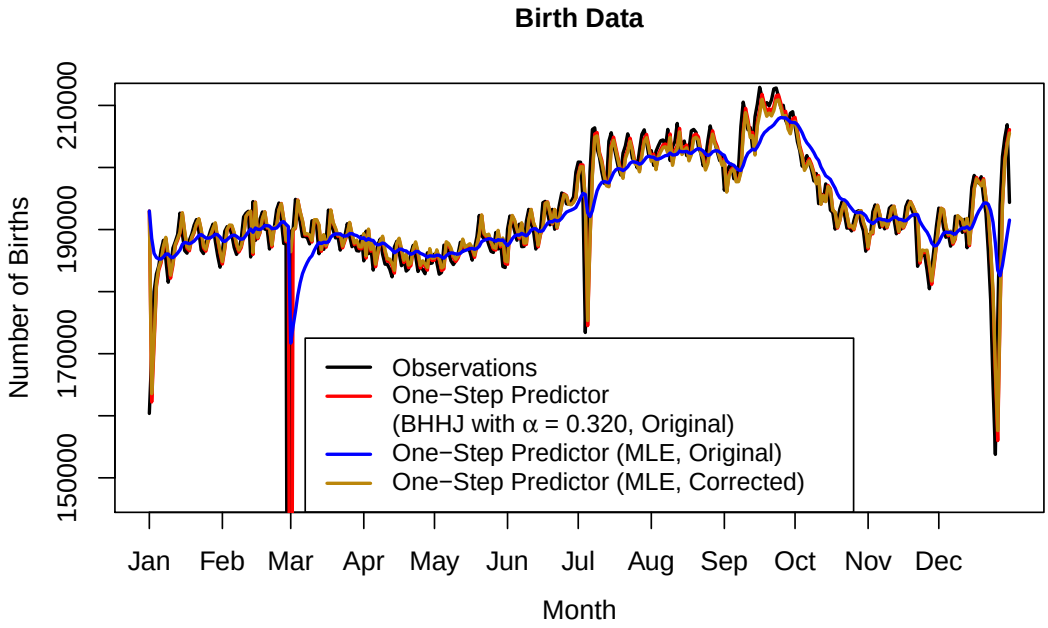

Figure 4.3: Plot of one-step predictors of the U.S. birth rate series from 1969 to 1988. Note that the predictors using the BHHJ MDE with $\alpha = 0.320$ based on the original data closely follow the predictors using the MLE based on the corrected data (red and brown, respectively).

one-step predictors based on the MLE using the corrected series. This illustrates that the BHHJ MDE with the automatically chosen tuning parameter $\alpha = 0.320$ is fairly robust.

In perusing the preceding results, one might note that the BHHJ MDE for the standard deviation of the state noise ($\hat{\sigma}_\omega = 2.3762$) based on the original series is noticeably smaller than the MLE ($\hat{\sigma}_\omega = 3.5325$) based on the corrected series. This could be explained by the existence of several outliers in the time series in addition to the February 29 outlier. For example, the data points corresponding to January 1, January 2, July 4, December 24, and December 25 seem to have a significantly lower magnitude compared to the other data points in the time series (see Figure 4.4). Therefore, the MLE based on the series with only the February 29

outlier corrected is still not too trustworthy. In fact, the birth rate is typically much lower during major holidays in the United States. As posited by Rindfuss and Ladinsky (1976), "the probable explanation is the use of elective induction to bring about the onset of labor coupled with the preference of obstetricians to spend Sundays and holidays with their families." Since it is not obvious how the outliers corresponding to the holidays should be adjusted, and since most major holidays in the United States are in January and December, we consider an analysis based on a truncated version of the series, where the observations from January and December are removed. Thus, the partial series features the average daily birth rate from February to November (10 months). We then compare the BHHJ MDE based on the partial series featuring the February 29 outlier and the MLE based on the partial series with the February 29 outlier corrected, hoping to further validate the robustness of the BHHJ MDE.

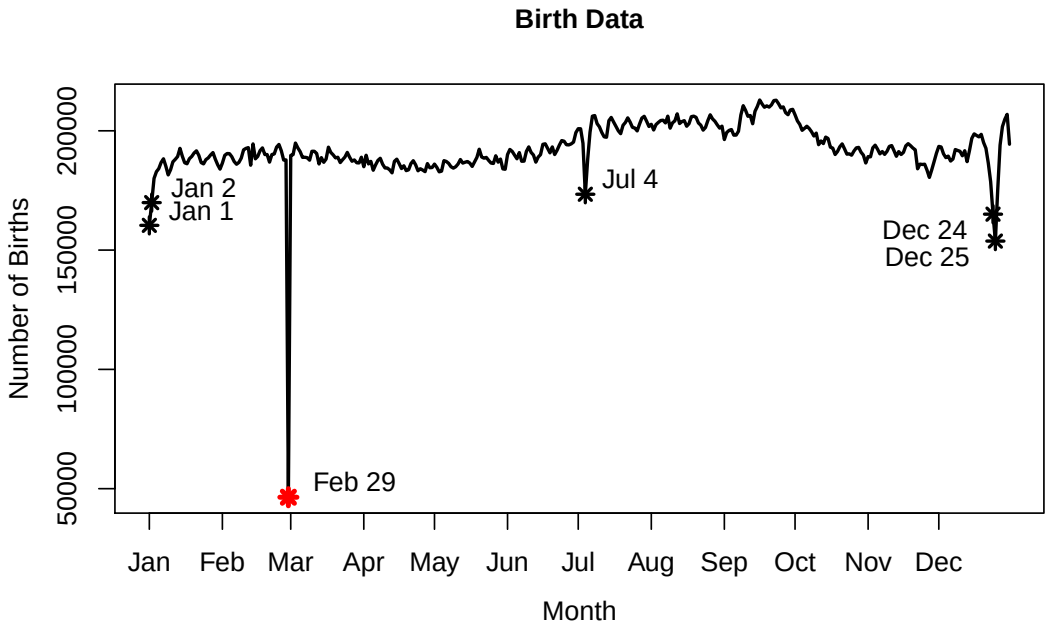

Figure 4.4: Daily average U.S. birth rate data from 1969 to 1988. Some obvious outliers are noted.

Based on the 10-month partial series featuring the February 29 outlier, we calculate the MLE along with the standard errors; the results are shown in Table 4.9. Following the same procedure described in the preceding, we determine an α value for the BHHJ MDE based on $\text{MSE}_{\text{std}}^{\mu,\sigma}$ and calculate the corresponding estimate. The results, including the BHHJ MDE along with the standard errors, are shown in Table 4.10 and Table 4.11. Finally, we correct the February 29 outlier and compute the MLE along with the standard errors; the results are featured in Table 4.12.

Table 4.9: MLE and standard errors based on the partial series.

	$\hat{\phi}$	$\hat{\sigma}_v$	$\hat{\sigma}_\omega$
MLE	0.9868	8.5708	1.3024
Standard Error	0.0114	0.3770	0.3102

Table 4.10: Choice of α using different methods based on the partial series.

$\text{MSE}_{\text{std}}^{\mu,\sigma}$	$\text{MSE}_{\alpha=1}^{\mu,\sigma}$	MSE^{μ}	$\text{var}^{\mu,\sigma}$
0.220	0.260	0.115	0.235

Table 4.11: BHHJ MDE with $\alpha = 0.220$ and standard errors based on the partial series.

	$\hat{\phi}$	$\hat{\sigma}_v$	$\hat{\sigma}_\omega$
BHHJ MDE ($\alpha = 0.220$)	0.9522	0.0033	2.2994
Standard Error	0.0160	0.0208	0.1114

Table 4.12: MLE and standard errors based on the partial corrected series.

	$\hat{\phi}$	$\hat{\sigma}_v$	$\hat{\sigma}_\omega$
MLE	0.9324	0.0001	2.8491
Standard Error	0.0207	0.6336	0.1159

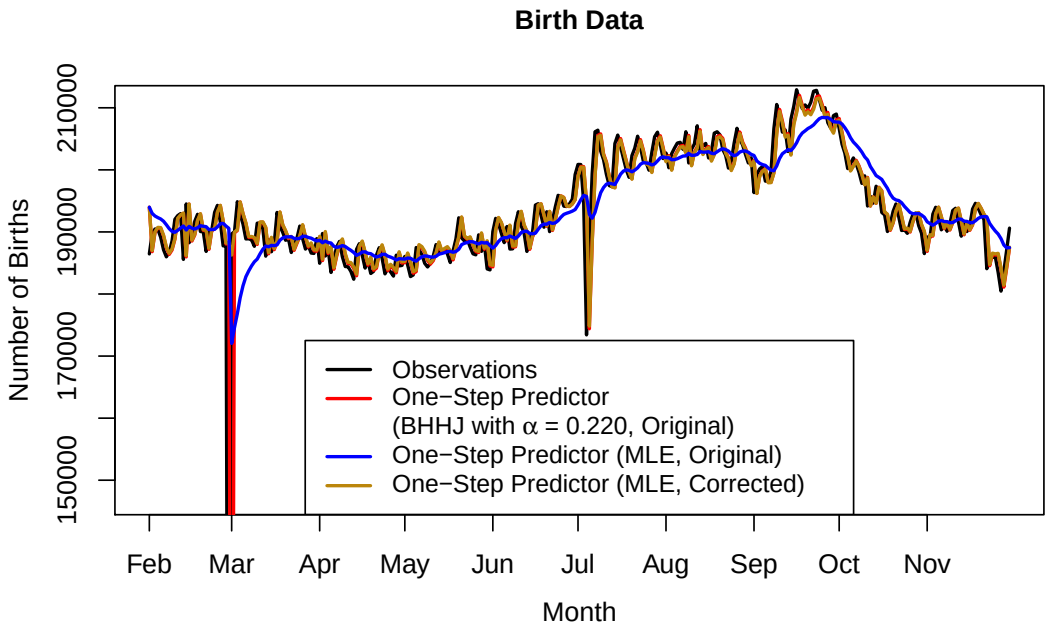

Figure 4.5: Plot of one-step predictors of the partial U.S. birth rate series from 1969 to 1988. Note that the predictors using the BHHJ MDE with $\alpha = 0.220$ based on the partial series nearly overlap with the predictors using the MLE based on the partial corrected series (red and brown, respectively).

The BHHJ MDE with the chosen $\alpha = 0.220$ based on the partial series (Table 4.11) is very close to the MLE (Table 4.12) based on the corrected partial series. The latter estimate could serve as the "gold standard", since most of the major outliers are absent in the corrected partial time series. The fact that our BHHJ MDE is close to the "gold standard" further validates the robustness of our estimation

method.

Again, we also plot the one-step predictors for the partial time series, based on different estimators. As expected, the one-step predictors based on the BHHJ MDE with $\alpha = 0.220$ using the partial series nearly overlap with the one-step predictors based on the MLE using the partial corrected series (see Figure 4.5).

4.3.2 Cardiovascular Disease Application

The second application is based on a time series that represents the incidence of cardiovascular mortality in the Los Angeles area during the 1970s. The series, featured in Figure 4.6, is comprised of 180 observations covering a time span of roughly 3 years; each reading represents the average mortality taken over a 6-day period. Yearly cycles are evident in the series. We will also use this time series for applications to illustrate the methodologies presented in Chapter 5 and Chapter 6.

Cavanaugh and Oleson (2001) suggested modeling the series using an ordinary AR(2) process. Using a proposed diagnostic to assess the influence of each observation on the forecasting of future values, they diagnosed three anomalous observations. In the present application, we consider modeling the series as an AR(2) state process observed with noise. To formulate this model in the state-space framework and to simplify the analysis, we subtract the mean value of the observations from each observation. The state-space model is then given by

$$y_t = [1,\ 0]\boldsymbol{x}_t + v_t, \tag{4.32}$$

$$\boldsymbol{x}_t = \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix} \boldsymbol{x}_{t-1} + \begin{bmatrix} \omega_t \\ 0 \end{bmatrix}, \tag{4.33}$$

where y_t is the centered cardiovascular mortality mean count, $\boldsymbol{x}_t = (x_t, x_{t-1})^T$ is the state variable, $v_t \sim N(0, R = \sigma_v^2)$, and $\begin{bmatrix} \omega_t \\ 0 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, Q = \begin{bmatrix} \sigma_w^2 & 0 \\ 0 & 0 \end{bmatrix} \right)$, for

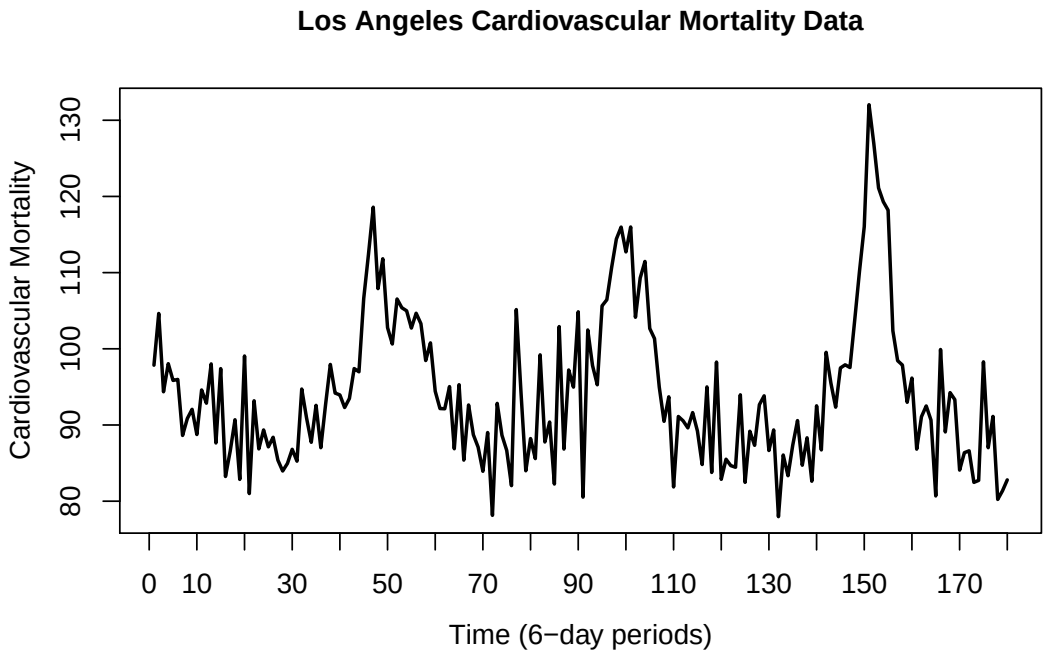

Figure 4.6: Three-year segment of the 1970's Los Angeles cardiovascular mortality incidence series.

$t = 1, 2, ..., 180$. We denote the vector of parameters as $\boldsymbol{\Theta} = (\phi_1, \phi_2, \sigma_v, \sigma_\omega)^T$. Again, the initial mean of the state $\boldsymbol{\mu}_0$ is fixed as $\mathbf{0}$; the diagonal elements in the initial state variance matrix are fixed as 10, and the off-diagonal elements are fixed as 0.

We first determine a suitable α value for the BHHJ MDE. From a visual inspection of Figure 4.6, the degree of contamination in the series does not appear to be pronounced. In other words, there seems to be only a small percentage of perturbed values exhibiting relatively small shifts in magnitude from the mean level. Before applying our data-driven method for choosing an α, intuitively, we expect that a relatively small α should be selected. Table 4.13 shows the choices of α based on four of the methods introduced in Section 4.2. As expected, all four methods choose relatively small α values. In particular, the method based on MSE^{μ} chooses $\alpha = 0.010$, which is very close to 0. Again, for the purpose of illustration,

we choose $\alpha = 0.180$ as governed by $\text{MSE}^{\mu,\sigma}_{\text{std}}$ to calculate the BHHJ MDE and its corresponding asymptotic standard errors.

Table 4.13: Choice of α using different methods.

$\text{MSE}^{\mu,\sigma}_{\text{std}}$	$\text{MSE}^{\mu,\sigma}_{\alpha=1}$	MSE^{μ}	$\text{var}^{\mu,\sigma}$
0.180	0.230	0.010	0.235

Based on the chosen $\alpha = 0.180$, the BHHJ MDE of $\boldsymbol{\Theta}$ and its standard errors are shown in Table 4.14. From inspecting the results, the estimated standard deviation of the observation noise v_t is extremely small, compared to that of the state noise ω_t. Cavanaugh and Oleson (2001) used a pure AR(2) process to model this time series, which essentially assumes all of the variability in the observed process arises from the state process. The fact that our BHHJ MDE puts most of the weight on the state process, to some extent, confirms the validity of the model employed in Cavanaugh and Oleson (2001).

Table 4.14: BHHJ MDE with $\alpha = 0.180$ and standard errors.

	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\sigma}_v$	$\hat{\sigma}_\omega$
BHHJ MDE ($\alpha = 0.180$)	0.3575	0.4935	0.2412	6.013
Standard Error	0.0056	0.0036	0.0744	0.1331

Based on the BHHJ MDE with $\alpha = 0.180$, we plot the one-step predictors for the Los Angeles cardiovascular mortality incidence series.

Lastly, we point out that since the contamination in cardiovascular mortality series is not severe, using the MLE method to estimate the parameters is probably not problematic. In fact, the criterion MSE^{μ} suggests using the BHHJ MDE with $\alpha = 0.010$, which is very close to the MLE. We also plotted the one-step predictors

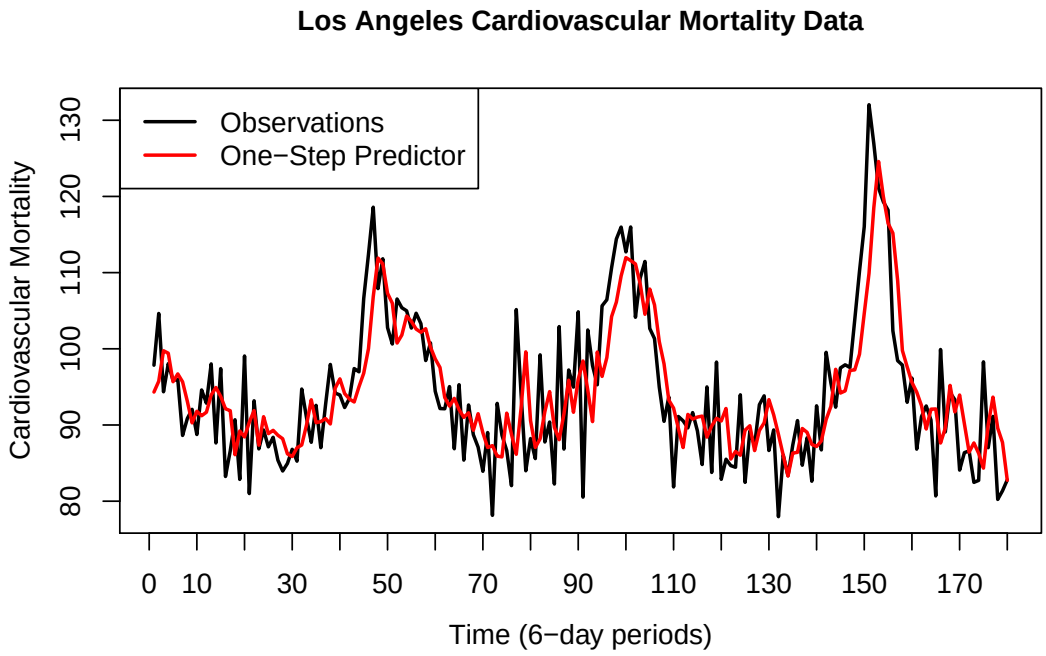

Figure 4.7: One-step predictors of the Los Angeles cardiovascular mortality incidence series, based on the BHHJ MDE with $\alpha = 0.180$.

based on the MLE; the difference between the predictors based on the MLE and the predictors based on the BHHJ MDE with $\alpha = 0.180$ is imperceptible. Thus, when we know the severity of the perturbation in the data is small, using the BHHJ method with the automatically chosen small α is almost equivalent to the MLE method. However, in most cases, assessing the extent of contamination in a series is a challenging task. (In Chapter 6, we propose an influence diagnostic in the state-space models to facilitate this task.) Further, in the presence of severe contamination, the BHHJ MDE with the automatically chosen α downweights the influence of the perturbation much more effectively than the MLE, as we observed in the simulated example in Section 3.2 and the birth rate application in Section 4.3.1.

In this chapter, we have proposed a numerical method to estimate the asymptotic variance of the BHHJ MDE, and an approach for automatically choosing an

appropriate α in practice. We also have illustrated how to apply our methodologies using two real time series: one featuring the monthly mean number of newborns in the United States, and the other one featuring the incidence of Los Angeles cardiovascular mortality. The procedure for applying our estimation method may be summarized as follows:

1) we first construct a state-space model for the given time series;

2) we then use our data-driven approach to automatically choose an α for the BHHJ MDE in fitting the state-space model;

3) based on the chosen α, we calculate the BHHJ MDE, along with the corresponding asymptotic standard errors;

4) if prediction of the states is an inferential objective, based on the BHHJ MDE, we obtain the state predictors.

CHAPTER 5
MODEL SELECTION CRITERION FOR STATE-SPACE MODELS BASED ON THE BHHJ DISCREPANCY

Based on the robust parameter estimator BHHJ MDE, we develop a model selection criterion for state-space models. In instances where the time series data is contaminated, our proposed selection criterion is shown to perform favorably relative to the traditional Akaike information criterion.

5.1 Procedural Development

Recall that under certain regularity conditions, we can treat the innovations of the state-space model $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, ..., \boldsymbol{\epsilon}_N$ as approximately i.i.d. To construct our model selection criterion, we choose the stabilized innovation $\boldsymbol{\epsilon}$ as the random variable of interest. Details on the distributional convergence property of the innovations can be found in Chapter 3.

Next, we introduce some notation necessary for the development of our proposed model selection criterion.

Let $\mathcal{F}$ denote the collection of density functions for $\boldsymbol{\epsilon}$. Let $f(\cdot|\boldsymbol{\Theta}) \in \mathcal{F}$ denote the parametric density function under the candidate or approximating model, and let $g(\cdot) \in \mathcal{F}$ denote the true or generating model for $\boldsymbol{\epsilon}$.

Let $\mathcal{F}(k) = \{f(\cdot|\boldsymbol{\Theta}_k)|\boldsymbol{\Theta}_k \in \Omega(k)\}$ denote a k-dimensional parametric class, i.e., a class of density functions in which the parameter space $\Omega(k)$ consists of k-dimensional vectors whose components are functionally independent. Further, for a fixed non-negative value of α, let $\widehat{\boldsymbol{\Theta}}_k$ denote the BHHJ MDE obtained by minimizing the BHHJ empirical discrepancy over $\Omega(k)$, and let $f(\cdot|\widehat{\boldsymbol{\Theta}}_k)$ represent the resulting empirical likelihood. Our goal is to search among the collection of classes $\mathcal{F} = \{\mathcal{F}(k_1), \mathcal{F}(k_2), ..., \mathcal{F}(k_L)\}$ for the fitted model $f(\cdot|\widehat{\boldsymbol{\Theta}}_k)$, $k \in \{k_1, k_2, ..., k_L\}$, that serves as the "best" approximation to $g(\cdot)$. Again, the set of parameters for the

state-space model is $\boldsymbol{\Theta} = \{\boldsymbol{\mu}_0, \Sigma_0, \Phi, R, Q\}$.

For the purpose of simplicity, we will slightly alter the definition of the BHHJ empirical discrepancy $\Delta^{\alpha}_{BHHJ}(\boldsymbol{Y}, \boldsymbol{\Theta})$. In this chapter, we redefine the measure as

$$\Delta^{\alpha}_{BHHJ}(\boldsymbol{Y}, \boldsymbol{\Theta}) = \int f^{1+\alpha}(\boldsymbol{z}|\boldsymbol{\Theta})d\boldsymbol{z} - \left(1 + \frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\boldsymbol{\Theta})\right\}. \tag{5.1}$$

Note that we change f_t to f for $t = 1, 2, ..., N$ since we are treating $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, ..., \boldsymbol{\epsilon}_N$ as i.i.d.

For any fixed non-negative α, let us first consider a measure of the following form:

$$\Delta^{\alpha}_{BHHJ}(g, \boldsymbol{\Theta}) = E_g\left\{\Delta^{\alpha}_{BHHJ}(\boldsymbol{Y}, \boldsymbol{\Theta})\right\}. \tag{5.2}$$

Intuitively, $\Delta^{\alpha}_{BHHJ}(g, \boldsymbol{\Theta})$ reflects the disparity between a certain parametric model and the generating model.

Now for each k-dimensional parametric class $\mathcal{F}(k)$, we choose the density corresponding to the fitted model, $f(\cdot|\widehat{\boldsymbol{\Theta}}_k)$, as the representative for the class. We define the *overall discrepancy* as

$$\Delta^{\alpha}_{BHHJ}(g, \widehat{\boldsymbol{\Theta}}_k) = E_g\left\{\Delta^{\alpha}_{BHHJ}(\boldsymbol{Y}, \boldsymbol{\Theta})\right\}\big|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}_k}, \tag{5.3}$$

Explicitly in terms of the densities, we can rewrite $\Delta^{\alpha}_{BHHJ}(g, \widehat{\boldsymbol{\Theta}}_k)$ as

$$\begin{aligned}
\Delta^{\alpha}_{BHHJ}(g, \widehat{\boldsymbol{\Theta}}_k) &= E_g\left\{\int f^{1+\alpha}(\boldsymbol{z}|\boldsymbol{\Theta})d\boldsymbol{z} - \left(1 + \frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\boldsymbol{\Theta})\right\}\right\}\Bigg|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}_k} \\
&= \int f^{1+\alpha}(\boldsymbol{z}|\widehat{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1 + \frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} E_g\left\{f^{\alpha}(\boldsymbol{\epsilon_t}|\boldsymbol{\Theta})\right\}\right\}\Bigg|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}_k} \\
&= \int f^{1+\alpha}(\boldsymbol{z}|\widehat{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1 + \frac{1}{\alpha}\right) E_g\left\{f^{\alpha}(\boldsymbol{z}|\boldsymbol{\Theta})\right\}\big|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}_k} \\
&= \int \left\{f^{1+\alpha}(\boldsymbol{z}|\widehat{\boldsymbol{\Theta}}_k) - \left(1 + \frac{1}{\alpha}\right) g(\boldsymbol{z}) f^{\alpha}(\boldsymbol{z}|\widehat{\boldsymbol{\Theta}}_k)\right\} d\boldsymbol{z}.
\end{aligned} \tag{5.4}$$

Note that $\Delta^{\alpha}_{BHHJ}(g, \widehat{\boldsymbol{\Theta}}_k)$ as represented in (5.4) is in fact the first two terms of $\Delta^{\alpha}_{BHHJ}(g, f)$, if we replace the parameter vector in f with $\widehat{\boldsymbol{\Theta}}_k$. Recall that

$\Delta^{\alpha}_{BHHJ}(g, f)$ is defined as

$$\Delta^{\alpha}_{BHHJ}(g, f) = \int \left\{ f^{1+\alpha}(\boldsymbol{z}|\boldsymbol{\Theta}) - \left(1 + \frac{1}{\alpha}\right) g(\boldsymbol{z}) f^{\alpha}(\boldsymbol{z}|\boldsymbol{\Theta}) + \frac{1}{\alpha} g^{1+\alpha}(\boldsymbol{z}) \right\} d\boldsymbol{z}. \quad (5.5)$$

The *expected overall discrepancy* averages the overall discrepancy over the sampling distribution of $\widehat{\boldsymbol{\Theta}}_k$:

$$D^{\alpha}_{BHHJ}(g, k) = E_g \left\{ \Delta^{\alpha}_{BHHJ}(g, \widehat{\boldsymbol{\Theta}}_k) \right\}. \quad (5.6)$$

Intuitively, the expected overall discrepancy $D^{\alpha}_{BHHJ}(g, k)$ reflects how well, on average, fitted approximating models having the same structure as $f(\cdot|\widehat{\boldsymbol{\Theta}}_k)$ predict "new" data generated under the true model g. A model selection criterion is often formulated by constructing an asymptotically unbiased estimator of the expected overall discrepancy. To construct our new model selection criterion, our next task is to find an asymptotically unbiased estimator of $D^{\alpha}_{BHHJ}(g, k)$.

As a side note, $\Delta^{\alpha}_{BHHJ}(g, \widehat{\boldsymbol{\Theta}}_k)$ is clearly unbiased for $D^{\alpha}_{BHHJ}(g, k)$. However, computing $\Delta^{\alpha}_{BHHJ}(g, \widehat{\boldsymbol{\Theta}}_k)$ requires the knowledge of the generating model g, and thus $\Delta^{\alpha}_{BHHJ}(g, \widehat{\boldsymbol{\Theta}}_k)$ cannot be used as a model selection criterion. We need to develop an asymptotically unbiased estimator purely based on the data $\boldsymbol{Y}$ and the form of the candidate model.

A natural estimator of the expected overall discrepancy is the *estimated discrepancy*, $\Delta^{\alpha}_{BHHJ}(\boldsymbol{Y}, \widehat{\boldsymbol{\Theta}}_k)$. The estimated discrepancy $\Delta^{\alpha}_{BHHJ}(\boldsymbol{Y}, \widehat{\boldsymbol{\Theta}}_k)$ reflects how well the fitted approximating model $f(\cdot|\widehat{\boldsymbol{\Theta}}_k)$ predicts the data at hand, $\boldsymbol{Y}$. Therefore, $\Delta^{\alpha}_{BHHJ}(\boldsymbol{Y}, \widehat{\boldsymbol{\Theta}}_k)$ yields an overly optimistic assessment of how effectively the fitted model predicts new data. Consequently, $\Delta^{\alpha}_{BHHJ}(\boldsymbol{Y}, \widehat{\boldsymbol{\Theta}}_k)$ serves as a negatively biased estimator of the expected overall discrepancy $D^{\alpha}_{BHHJ}(g, k)$; correcting for this bias leads to the penalty term of the selection criterion. For the remainder of this section, we will focus on developing this penalty term.

We rewrite $D^{\alpha}_{BHHJ}(g,k)$ as

$$\begin{aligned}
D^{\alpha}_{BHHJ}(g,k) &= E_g\left\{\Delta^{\alpha}_{BHHJ}(g,\widehat{\boldsymbol{\Theta}}_k)\right\} \\
&= E_g\left\{\Delta^{\alpha}_{BHHJ}(\boldsymbol{Y},\widehat{\boldsymbol{\Theta}}_k)\right\} \\
&\quad + E_g\left\{\int f^{1+\alpha}(\boldsymbol{z}|\overline{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\overline{\boldsymbol{\Theta}}_k)\right\}\right\} \\
&\quad - E_g\left\{\Delta^{\alpha}_{BHHJ}(\boldsymbol{Y},\widehat{\boldsymbol{\Theta}}_k)\right\} + E_g\left\{\Delta^{\alpha}_{BHHJ}(g,\widehat{\boldsymbol{\Theta}}_k)\right\} \\
&\quad - E_g\left\{\int f^{1+\alpha}(\boldsymbol{z}|\overline{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\overline{\boldsymbol{\Theta}}_k)\right\}\right\} \\
&= E_g\left\{\int f^{1+\alpha}(\boldsymbol{z}|\widehat{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\widehat{\boldsymbol{\Theta}}_k)\right\}\right\} \\
&\quad + E_g\left\{\int f^{1+\alpha}(\boldsymbol{z}|\overline{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\overline{\boldsymbol{\Theta}}_k)\right\}\right\} \\
&\quad - E_g\left\{\int f^{1+\alpha}(\boldsymbol{z}|\widehat{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\widehat{\boldsymbol{\Theta}}_k)\right\}\right\} \\
&\quad + E_g\left\{\Delta^{\alpha}_{BHHJ}(g,\widehat{\boldsymbol{\Theta}}_k)\right\} \\
&\quad - E_g\left\{\int f^{1+\alpha}(\boldsymbol{z}|\overline{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\overline{\boldsymbol{\Theta}}_k)\right\}\right\}, \quad (5.7)
\end{aligned}$$

where the pseudo true parameter $\overline{\boldsymbol{\Theta}}_k$ is defined as

$$\begin{aligned}
\overline{\boldsymbol{\Theta}}_k &= \arg\min_{\boldsymbol{\Theta}\in\Omega_k} \Delta^{\alpha}_{BHHJ}(g,f) \\
&= \arg\min_{\boldsymbol{\Theta}\in\Omega_k} \int f^{1+\alpha}(\boldsymbol{z}|\boldsymbol{\Theta}) - \left(1+\frac{1}{\alpha}\right) g(\boldsymbol{z}) f^{\alpha}(\boldsymbol{z}|\boldsymbol{\Theta})d\boldsymbol{z}. \quad (5.8)
\end{aligned}$$

The terms in (5.7) are fairly lengthy, and thus for the sake of simplicity, we make the following denotations:

$$\begin{aligned}
\textcircled{1} &:= E_g\left\{\int f^{1+\alpha}(\boldsymbol{z}|\widehat{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\widehat{\boldsymbol{\Theta}}_k)\right\}\right\}, \\
\textcircled{2} &:= E_g\left\{\int f^{1+\alpha}(\boldsymbol{z}|\overline{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\overline{\boldsymbol{\Theta}}_k)\right\}\right\} \\
&\quad - E_g\left\{\int f^{1+\alpha}(\boldsymbol{z}|\widehat{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\widehat{\boldsymbol{\Theta}}_k)\right\}\right\}, \text{ and}
\end{aligned}$$

$$
\begin{aligned}
\textcircled{3} &:= E_g\left\{\Delta^{\alpha}_{BHHJ}(g,\widehat{\boldsymbol{\Theta}}_k)\right\} \\
&\quad - E_g\left\{\int f^{1+\alpha}(\boldsymbol{z}|\overline{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\overline{\boldsymbol{\Theta}}_k)\right\}\right\}.
\end{aligned}
$$

Clearly, we have $D^{\alpha}_{BHHJ}(g,k) = \textcircled{1} + \textcircled{2} + \textcircled{3}$. In particular, $\textcircled{1}$ is the expectation of the estimated discrepancy. Efron (1983, 1986) referred to the quantity $\textcircled{2} + \textcircled{3}$ as the expected optimism in judging the fit of a model using the same data as that which was used to construct the fit.

We now present two important lemmas regarding $\textcircled{2}$ and $\textcircled{3}$. The proofs can be found in Mattheou, Lee, and Karagrigoriou (2009), with some minor changes required for the present setting. Their proofs assume that the observed data are i.i.d. and that the true model belongs to every parametric candidate class. In our state-space modeling framework, we treat the innovations as i.i.d., and thus our procedural development is based on the innovations. Also, since we do not assume that the true model g belongs to every candidate class $\mathcal{F}(k)$, the true parameter $\boldsymbol{\Theta}_0$ must be replaced with the corresponding pseudo true parameter $\overline{\boldsymbol{\Theta}}_k$.

Lemma 1. *Expression* $\textcircled{2}$ *can be written as*

$$
\begin{aligned}
\textcircled{2} &= E_g\left\{\int f^{1+\alpha}(\boldsymbol{z}|\overline{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\overline{\boldsymbol{\Theta}}_k)\right\}\right\} \\
&\quad - E_g\left\{\int f^{1+\alpha}(\boldsymbol{z}|\widehat{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\widehat{\boldsymbol{\Theta}}_k)\right\}\right\} \\
&= \frac{1}{2}(1+\alpha)\, E_g\left\{\left(\overline{\boldsymbol{\Theta}}_k - \widehat{\boldsymbol{\Theta}}_k\right)^T J\left(\overline{\boldsymbol{\Theta}}_k\right)\left(\overline{\boldsymbol{\Theta}}_k - \widehat{\boldsymbol{\Theta}}_k\right)\right\} + o(1), \qquad (5.9)
\end{aligned}
$$

where $J\left(\overline{\boldsymbol{\Theta}}_k\right)$ *is defined in equation (4.1).*

Proof. See Theorem 2.3 in Mattheou, Lee, and Karagrigoriou (2009). □

Lemma 2. *Expression* $\textcircled{3}$ *can be written as*

$$
\textcircled{3} := E_g\left\{\Delta^{\alpha}_{BHHJ}(g,\widehat{\boldsymbol{\Theta}}_k)\right\}
$$

$$
\begin{aligned}
&- E_g\left\{\int f^{1+\alpha}(\boldsymbol{z}|\overline{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\overline{\boldsymbol{\Theta}}_k)\right\} \\
&= \frac{1}{2}(1+\alpha)\, E_g\left\{\left(\overline{\boldsymbol{\Theta}}_k - \widehat{\boldsymbol{\Theta}}_k\right)^T J\left(\overline{\boldsymbol{\Theta}}_k\right)\left(\overline{\boldsymbol{\Theta}}_k - \widehat{\boldsymbol{\Theta}}_k\right)\right\} + o(1), \qquad (5.10)
\end{aligned}
$$

where $J\left(\overline{\boldsymbol{\Theta}}_k\right)$ *is defined in equation (4.1).*

Proof. See Theorem 2.1 in Mattheou, Lee, and Karagrigoriou (2009). □

Based on Lemma 1 and Lemma 2, we are positioned to construct an asymptotically unbiased estimator of the expected overall discrepancy $D^{\alpha}_{BHHJ}(g,k)$.

Theorem 6. *An asymptotically unbiased estimator of* N *times the expected overall discrepancy* $D^{\alpha}_{BHHJ}(g,k)$ *is given by*

$$
N\left\{\int f^{1+\alpha}(\boldsymbol{z}|\widehat{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\widehat{\boldsymbol{\Theta}}_k)\right\}\right\} + (1+\alpha)tr\left(\widehat{J}^{-1}(\widehat{\boldsymbol{\Theta}}_k)\widehat{K}(\widehat{\boldsymbol{\Theta}}_k)\right), \qquad (5.11)
$$

where $\widehat{J}$ *and* $\widehat{K}$ *are defined in equations (4.12) and (4.13), respectively.*

Proof. See Appendix. □

Thus, we have developed a model selection criterion based on the BHHJ discrepancy. The first part of the criterion

$$
N\left\{\int f^{1+\alpha}(\boldsymbol{z}|\widehat{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\widehat{\boldsymbol{\Theta}}_k)\right\}\right\} \qquad (5.12)
$$

is often referred to as a "*goodness-of-fit*" term. For any fixed positive α, one needs to calculate the expression (5.11) for $k = k_1, k_2, ..., k_L$ and choose the model with the smallest criterion value. Note that the estimate $\widehat{\boldsymbol{\Theta}}_k$ in (5.11) is the BHHJ MDE, and thus the evaluation of the criterion requires the use of the estimation method developed in Chapter 3. Since the form of (5.11) is similar to the Takeuchi information criterion (TIC) (Takeuchi, 1976), and since the tuning parameter α is involved, we name (5.11) as TIC_{α}.

Note that the bias correction term $\text{tr}\left(\widehat{J}^{-1}(\widehat{\boldsymbol{\Theta}}_k)\widehat{K}(\widehat{\boldsymbol{\Theta}}_k)\right)$ of TIC_{α} varies from

sample to sample. Although TIC_α serves as an asymptotically unbiased estimator, the accuracy of an estimator is not only dictated by bias, but also by variability. Since the trace component $\text{tr}\left(\widehat{J}^{-1}(\widehat{\boldsymbol{\Theta}}_k)\widehat{K}(\widehat{\boldsymbol{\Theta}}_k)\right)$ is stochastic, the variability in the penalty term is potentially large (Kitagawa, 1987). Furthermore, the trace does not have an explicit form; instead, we need to use the numerical method introduced in Section 4.1 to obtain an approximation. The numerical calculation of $\widehat{J}(\widehat{\boldsymbol{\Theta}}_k)$ and $\widehat{K}(\widehat{\boldsymbol{\Theta}}_k)$ is in fact fairly expensive. In cases where the number of candidate models is large, choosing a model could be time consuming. Computational considerations aside, as we will see in the following, the performance of the model selection criterion TIC_α is not satisfactory for any of the α values we tried, and is not superior to AIC, regardless of whether contamination is present.

We present a two-part Monte Carlo simulation study to compare the performance of the criterion TIC_α with the traditional model selection criteria AIC and TIC. The first part of simulation study does not introduce perturbation, whereas the second part does. Note that TIC is equivalent to $\text{TIC}_{\alpha=0}$. The number of replications in both parts of the simulation study is 100. We investigate the performance of TIC_α for a range of different α values.

The generating model is

$$y_t = x_t + v_t, \tag{5.13}$$

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \omega_t, \tag{5.14}$$

where $\phi_1 = 0.99$, $\phi_2 = -0.8$, $v_t \sim N(0, \sigma_v^2 = 1)$, and $\omega_t \sim N(0, \sigma_\omega^2 = 1)$. From equation (5.14), the state in the generating model is characterized by an AR(2) process. For both parts of the simulation study, the candidate models for the state process are specified as AR(1), AR(2), AR(3), and AR(4). We will assess the effectiveness of each model selection criterion by its overall rate of choosing the correct model order for the state process.

For comparison, we next list the explicit forms of AIC, TIC, and TIC_α. To notationally differentiate the MLE from the BHHJ MDE $\widehat{\boldsymbol{\Theta}}_k$, we use $\widehat{\boldsymbol{\Theta}}_{k,MLE}$ in the definition of both AIC and TIC. Also, we use $\widehat{J}_{MLE}$ and $\widehat{K}_{MLE}$ in the definition of TIC to differentiate these matrices from $\widehat{J}$ and $\widehat{K}$ in TIC_α. Specifically, $\widehat{J}_{MLE}$ and $\widehat{K}_{MLE}$ can be obtained from redefining $\delta(\boldsymbol{\Theta})$ in Section 4.1 as

$$\delta(\boldsymbol{\Theta}) = -\sum_{t=1}^{N} \log f(\boldsymbol{\epsilon_t}|\boldsymbol{\Theta}). \tag{5.15}$$

The derivation of $\widehat{J}_{MLE}$ and $\widehat{K}_{MLE}$ then follows the development of $\widehat{J}$ and $\widehat{K}$ in Section 4.1.

$$\text{AIC} = -\left\{\sum_{t=1}^{N} \log f(\boldsymbol{\epsilon_t}|\widehat{\boldsymbol{\Theta}}_{k,MLE})\right\} + k, \tag{5.16}$$

$$\text{TIC} = -\left\{\sum_{t=1}^{N} \log f(\boldsymbol{\epsilon_t}|\widehat{\boldsymbol{\Theta}}_{k,MLE})\right\} + \text{tr}\left(\widehat{J}^{-1}_{MLE}(\widehat{\boldsymbol{\Theta}}_{k,MLE})\widehat{K}_{MLE}(\widehat{\boldsymbol{\Theta}}_{k,MLE})\right), \tag{5.17}$$

$$\begin{aligned}\text{RIC}_\alpha =& N\left\{\int f^{1+\alpha}(\boldsymbol{z}|\widehat{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\widehat{\boldsymbol{\Theta}}_k)\right\}\right\} \\ &+ (1+\alpha)\text{tr}\left(\widehat{J}^{-1}(\widehat{\boldsymbol{\Theta}}_k)\widehat{K}(\widehat{\boldsymbol{\Theta}}_k)\right).\end{aligned} \tag{5.18}$$

Simulation Study, Part I: In this part of the simulation study, we do not introduce any perturbations to the generated data. Table 5.1 displays the distributions of the choices of the AR order for the state process based on the selection criteria AIC, TIC, and TIC_α. A good model selection criterion should have a relatively large proportion of choices for the correct order 2, since the generating model is AR(2).

We can see from Table 5.1 that the performance of TIC_α and TIC is not superior to AIC when the data is not corrupted.

Simulation Study, Part II: In this part of the simulation study, we randomly perturb 10% of the observations in each replicated sample by introducing an

Table 5.1: Results for simulation study, part I. Generating model for the state process: AR(2).

	AR Order			
Selection Criterion	1	2	3	4
AIC	0	68	19	13
TIC	0	44	37	19
$\text{TIC}_{\alpha=0.001}$	0	38	38	24
$\text{TIC}_{\alpha=0.05}$	0	46	33	21
$\text{TIC}_{\alpha=0.1}$	0	47	32	21
$\text{TIC}_{\alpha=0.3}$	0	45	36	19
$\text{TIC}_{\alpha=0.5}$	0	47	31	22
$\text{TIC}_{\alpha=0.7}$	0	50	27	23
$\text{TIC}_{\alpha=0.8}$	0	54	26	20
$\text{TIC}_{\alpha=1.0}$	0	51	27	22

additive shift with a magnitude of 8. Although the generated data is corrupted, an effective model selection criterion should still have the capacity to identify the original state generating model AR(2). Such a criterion would then be considered robust.

Again, Table 5.2 shows that TIC and TIC_α are still not superior to AIC, even in the presence of contamination. Intuitively, when a series is corrupted by anomalous values, the goodness-of-fit term based on the robust BHHJ MDE should be more reliable than the goodness-of-fit term based on the non-robust MLE, suggesting that TIC_α should outperform AIC. However, the simulation results do not coincide with our intuition. This could be explained by the large variability of the bias correction term for TIC_α.

Table 5.2: Results for simulation study, part II. Generating model for the state process: AR(2). In each sample, 10% of the observations are additively perturbed by a magnitude shift of 8.

	AR Order			
Selection Criterion	1	2	3	4
AIC	12	55	16	17
TIC	17	33	18	32
$\text{TIC}_{\alpha=0.001}$	13	24	27	36
$\text{TIC}_{\alpha=0.05}$	14	31	26	29
$\text{TIC}_{\alpha=0.1}$	19	30	22	29
$\text{TIC}_{\alpha=0.3}$	10	35	26	29
$\text{TIC}_{\alpha=0.5}$	10	29	27	34
$\text{TIC}_{\alpha=0.7}$	11	35	24	30
$\text{TIC}_{\alpha=0.8}$	11	38	25	26
$\text{TIC}_{\alpha=1.0}$	10	41	26	23

Next, we will modify the model selection criterion TIC_α based on some additional assumptions, hoping to obtain a more reliable selection tool. We present the following theorem.

Theorem 7. *Consider the following two assumptions:*

(1) the generating model g belongs to every candidate class $\mathcal{F}(k)$ for $k = k_1, k_2, ..., k_L$, and

(2) the tuning parameter α is relatively close to 0^+.

Under these assumptions, in large-sample settings, an approximately unbiased estimator of N times the expected overall discrepancy $D^\alpha_{BHHJ}(g, k)$ is given by

$$N\left\{\int f^{1+\alpha}(\boldsymbol{z}|\widehat{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^\alpha(\boldsymbol{\epsilon_t}|\widehat{\boldsymbol{\Theta}}_k)\right\}\right\} + (1+\alpha)k, \qquad (5.19)$$

where k is the dimension of the parameter vector.

Proof. See Appendix. □

We name the model selection criterion (5.19) as the *robust information criterion* (RIC_α):

$$\text{RIC}_\alpha = N\left\{\int f^{1+\alpha}(\boldsymbol{z}|\widehat{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^{\alpha}(\boldsymbol{\epsilon_t}|\widehat{\boldsymbol{\Theta}}_k)\right\}\right\} + (1+\alpha)k. \quad (5.20)$$

Note that RIC_α is developed under more stringent assumptions than TIC_α. First, RIC_α is derived under the assumption that the true or generating model belongs to every candidate class. One might argue that the assumption is too strong and thus unrealistic, especially in the presence of contamination in the data. However, Kitagawa (1987) argued that in settings where this assumption is grossly violated for some of the models in the candidate collection, the goodness-of-fit term should eliminate from consideration those models that are far removed from the generating model. If the remaining models are relatively close to the generating model, the assumption should approximately hold for these models, and the bias correction should be reasonably accurate.

In addition, when the level of contamination in the data is severe, a relatively large α value should be chosen to obtain a robust estimator. However, if a large α is selected, then based on assumption (2) in Theorem 7, RIC_α would apparently not serve as an approximately unbiased estimator of the expected overall discrepancy. Yet again, since the accuracy of an estimator is dictated by both bias and variability, perhaps the bias of RIC_α would be offset by the lack of variability in its penalty term. Later in this section, we present a comprehensive simulation study to evaluate the performance of RIC_α and to help substantiate the aforementioned arguments.

To further illuminate the development of RIC_α, we present an additional perspective to motivate the proposed penalization. Toma (2014) developed a model

selection criterion based on an underlying discrepancy of the general form

$$d(g, f) = \int \varphi\left(\frac{f(\boldsymbol{z})}{g(\boldsymbol{z})}\right) g(\boldsymbol{z}) d\boldsymbol{z}, \tag{5.21}$$

where φ satisfies certain regularity requirements. The author proved that, if the true model g belongs to every candidate class $\mathcal{F}(k)$ for $k = k_1, k_2, ..., k_L$, then an asymptotically unbiased estimator of N times the expected overall discrepancy is provided by

$$N\ d(\boldsymbol{Y}, \widehat{\boldsymbol{\Theta}}_k) + \varphi^{''}(1)k, \tag{5.22}$$

where the goodness-of-fit term $d(\boldsymbol{Y}, \widehat{\boldsymbol{\Theta}}_k)$ corresponds to an appropriately defined empirical discrepancy. For further details, we refer the reader to Toma (2014).

We now rewrite the BHHJ discrepancy with α as

$$\begin{aligned}
\Delta^{\alpha}_{BHHJ}(g, f) &= \int \left\{ f^{1+a}(\boldsymbol{z}) - \left(1 + \frac{1}{\alpha}\right) g(\boldsymbol{z}) f^{\alpha}(\boldsymbol{z}) + \frac{1}{\alpha} g^{1+\alpha}(\boldsymbol{z}) \right\} d\boldsymbol{z} \\
&= \int \left[\frac{\alpha \left(\frac{f(\boldsymbol{z})}{g(\boldsymbol{z})}\right)^{1+\alpha} - (1+\alpha)\left(\frac{f(\boldsymbol{z})}{g(\boldsymbol{z})}\right)^{\alpha} + 1}{\alpha} \right] g^{1+\alpha}(\boldsymbol{z}) d\boldsymbol{z} \\
&= \int \varphi\left(\frac{f(\boldsymbol{z})}{g(\boldsymbol{z})}\right) g^{1+\alpha}(\boldsymbol{z}) d\boldsymbol{z},
\end{aligned} \tag{5.23}$$

where

$$\varphi(x) = \frac{\alpha x^{1+\alpha} - (1+\alpha)x^{\alpha} + 1}{\alpha}. \tag{5.24}$$

For α close to 0^+, the form of the BHHJ discrepancy $\Delta^{\alpha}_{BHHJ}(g, f)$ corresponds to the general discrepancy form $d(g, f)$. One can easily check that for the BHHJ discrepancy,

$$\varphi^{''}(x) = \frac{\alpha^2(1+\alpha)x^{\alpha-1} - \alpha(\alpha^2-1)x^{\alpha-2}}{\alpha}, \tag{5.25}$$

and thus $\varphi^{''}(1) = 1 + \alpha$. Therefore, for α close to 0^+, in large-sample settings, an approximately unbiased estimator of N times the expected overall discrepancy is

provided by

$$N\ d(\boldsymbol{Y}, \widehat{\boldsymbol{\Theta}}_k) + \varphi''(1)k = N\ d(\boldsymbol{Y}, \widehat{\boldsymbol{\Theta}}_k) + (1+\alpha)k, \tag{5.26}$$

which is exactly the form of RIC_α.

Lastly, our proposed model selection criterion RIC_α is very similar to the divergence information criterion (DIC) developed by Mattheou, Lee, and Karagrigoriou (2009). Their criterion is given by

$$\text{DIC} = N\ \Delta^\alpha_{BHHJ}(\boldsymbol{Y}, \widehat{\boldsymbol{\Theta}}_k) + (2\pi)^{-\frac{\alpha}{2}} \left(\frac{1+\alpha}{1+2\alpha}\right)^{1+\frac{k}{2}} (1+\alpha)k. \tag{5.27}$$

The goodness-of-fit terms in DIC and RIC_α are the same, but DIC has a different penalty term. However, we point out that their criterion was developed in the framework of estimating the mean of a multivariate normal distribution; in other words, the parameter vector is assumed to be the mean vector of a multivariate normal distribution. In a state-space model, the parameter vector is usually comprised of the parameters of the transition matrix as well as various variance/covariance parameters. Therefore, DIC cannot be used in the state-space framework.

5.2 Simulation Study

In this section, we present a comprehensive simulation study to compare the performance of the proposed criterion RIC_α and the traditional criterion AIC. Again, for the purpose of comparison, we list the forms of AIC and RIC_α.

$$\text{AIC} = -\left\{\sum_{t=1}^{N} \log f(\boldsymbol{\epsilon_t}|\widehat{\boldsymbol{\Theta}}_{k,MLE})\right\} + k, \tag{5.28}$$

$$\text{RIC}_\alpha = N\left\{\int f^{1+\alpha}(\boldsymbol{z}|\widehat{\boldsymbol{\Theta}}_k)d\boldsymbol{z} - \left(1+\frac{1}{\alpha}\right)\left\{\frac{1}{N}\sum_{t=1}^{N} f^\alpha(\boldsymbol{\epsilon_t}|\widehat{\boldsymbol{\Theta}}_k)\right\}\right\} + (1+\alpha)k. \tag{5.29}$$

Our proposed model selection criterion RIC_α is essentially an extension of AIC: as α approaches 0^+, $\widehat{\boldsymbol{\Theta}}_k$ in RIC_α converges to $\widehat{\boldsymbol{\Theta}}_{k,MLE}$ in AIC, and the penalty

term in RIC_α converges to the penalty term in AIC. In our simulation study, we do not include the Bayesian information criterion (BIC) (Schwarz, 1978), since BIC has substantially different asymptotic properties than AIC (Neath and Cavanaugh, 2012), and thus also RIC_α. The two key large-sample optimality properties of model selection criteria are consistency and asymptotic efficiency. A consistent criterion will asymptotically select the fitted candidate model having the correct structure with probability one; an efficient criterion will asymptotically select the fitted candidate model that minimizes the mean squared error of prediction. AIC is asymptotically efficient yet not consistent, and BIC is consistent yet not asymptotically efficient. Therefore, we only compare the two parallel model selection criteria in terms of their large-sample optimality properties: RIC_α and AIC.

To evaluate our proposed selection criterion RIC_α, we compile twelve simulation sets, characterized by different sample sizes, different types of contamination, and different AR models for the state process. The study is organized as a three-way factorial experiment. The configuration for each set is specified in Table 5.3. The number of replications for all generating models is 100. For the state-space generating models, the observation equation is given by

$$y_t = x_t + v_t, \tag{5.30}$$

and the observation noise v_t and the state noise ω_t are both distributed as $N(0, 1)$. For the tuning parameter α, we choose the following values for all simulation sets: $0.01, 0.05, 0.08, 0.1, 0.15, 0.2, 0.3, 0.5, 0.6$. Most of the α values are relatively close to 0^+, but we also include some larger α values for the sake of completeness.

The selection results are summarized from Table 5.4 to Table 5.15. The ordering of these tables corresponds to the labeling in the first column of Table 5.3.

In the sets with no contamination (sets I, IV, VII, and X), RIC_α performs slightly better than traditional AIC, even for large α values. In general, AIC has a

tendency to overfit (i.e., to choose more complex models than the generating model). The penalty term $(1+\alpha)k$ helps improve the performance of RIC_α with positive α values, because the criterion penalizes the overfitted models to a greater degree than AIC. However, for RIC_α with α values that are too large, the problem of underfitting (i.e., choosing simpler models than the generating model) arises. Table 5.10 provides a good illustration of this phenomenon: the performance of RIC_α becomes better as α increases from 0.0 up to 0.1, yet deteriorates as α increases beyond this point.

For the simulation sets based on the larger sample sizes, both AIC and RIC_α perform better than in the corresponding sets based on the smaller sample sizes. Notably, RIC_α always outperforms AIC. As the degree of perturbation becomes more severe, the performance of AIC becomes worse; however, RIC_α stays relatively robust. Not surprisingly, as α converges to 0^+, the RIC_α selection patterns begin to resemble the AIC patterns. For some simulation sets (e.g., sets IV, V, X, and XII), RIC_α with relatively large α values perform the best; but overall, if we choose $\alpha = 0.15$ or 0.2, the performance of RIC_α stays relatively stable and yields satisfactory results.

Table 5.3: Generating models for simulation sets.

Set	Model for State Process	Contamination	Sample Size
I	$x_t = 0.99x_{t-1} - 0.8x_{t-2} + \omega_t$	No perturbation	50
II	$x_t = 0.99x_{t-1} - 0.8x_{t-2} + \omega_t$	5%; magnitude of 8	50
III	$x_t = 0.99x_{t-1} - 0.8x_{t-2} + \omega_t$	5%; magnitude of 18	50
IV	$x_t = 0.99x_{t-1} - 0.8x_{t-2} + \omega_t$	No perturbation	100
V	$x_t = 0.99x_{t-1} - 0.8x_{t-2} + \omega_t$	5%; magnitude of 8	100
VI	$x_t = 0.99x_{t-1} - 0.8x_{t-2} + \omega_t$	5%; magnitude of 18	100
VII	$x_t = -0.9x_{t-3} + \omega_t$	No perturbation	50
VIII	$x_t = -0.9x_{t-3} + \omega_t$	5%; magnitude of 8	50
IX	$x_t = -0.9x_{t-3} + \omega_t$	5%; magnitude of 18	50
X	$x_t = -0.9x_{t-3} + \omega_t$	No perturbation	100
XI	$x_t = -0.9x_{t-3} + \omega_t$	5%; magnitude of 8	100
XII	$x_t = -0.9x_{t-3} + \omega_t$	5%; magnitude of 18	100

Table 5.4: Results for simulation set I. Generating model for the state process: AR(2). No observations are perturbed. Sample size is 50.

	AR Order			
Selection Criterion	1	2	3	4
AIC	3	73	15	9
$\text{RIC}_{\alpha=0.01}$	3	76	13	8
$\text{RIC}_{\alpha=0.05}$	3	78	12	7
$\text{RIC}_{\alpha=0.08}$	3	79	12	6
$\text{RIC}_{\alpha=0.1}$	3	80	12	5
$\text{RIC}_{\alpha=0.15}$	3	82	10	5
$\text{RIC}_{\alpha=0.2}$	3	88	5	4
$\text{RIC}_{\alpha=0.3}$	5	92	2	1
$\text{RIC}_{\alpha=0.5}$	8	92	0	0
$\text{RIC}_{\alpha=0.6}$	16	84	0	0

Table 5.5: Results for simulation set II. Generating model for the state process: AR(2). In each sample, 5% of the observations are additively perturbed by a magnitude shift of 8. Sample size is 50.

	AR Order			
Selection Criterion	1	2	3	4
AIC	8	67	13	12
$\text{RIC}_{\alpha=0.01}$	7	70	11	12
$\text{RIC}_{\alpha=0.05}$	8	74	11	7
$\text{RIC}_{\alpha=0.08}$	9	77	9	5
$\text{RIC}_{\alpha=0.1}$	10	78	9	3
$\text{RIC}_{\alpha=0.15}$	10	82	6	2
$\text{RIC}_{\alpha=0.2}$	9	86	3	2
$\text{RIC}_{\alpha=0.3}$	13	83	2	2
$\text{RIC}_{\alpha=0.5}$	21	78	1	0
$\text{RIC}_{\alpha=0.6}$	30	70	0	0

Table 5.6: Results for simulation set III. Generating model for the state process: AR(2). In each sample, 5% of the observations are additively perturbed by a magnitude shift of 18. Sample size is 50.

	AR Order			
Selection Criterion	1	2	3	4
AIC	34	37	14	15
$\text{RIC}_{\alpha=0.01}$	30	44	13	13
$\text{RIC}_{\alpha=0.05}$	34	49	10	7
$\text{RIC}_{\alpha=0.08}$	21	64	11	4
$\text{RIC}_{\alpha=0.1}$	18	66	12	4
$\text{RIC}_{\alpha=0.15}$	16	66	12	6
$\text{RIC}_{\alpha=0.2}$	19	67	10	4
$\text{RIC}_{\alpha=0.3}$	25	67	7	1
$\text{RIC}_{\alpha=0.5}$	41	57	1	1
$\text{RIC}_{\alpha=0.6}$	49	49	2	0

Table 5.7: Results for simulation set IV. Generating model for the state process: AR(2). No observations are perturbed. Sample size is 100.

	AR Order			
Selection Criterion	1	2	3	4
AIC	0	79	12	9
$\text{RIC}_{\alpha=0.01}$	0	81	11	8
$\text{RIC}_{\alpha=0.05}$	0	85	9	6
$\text{RIC}_{\alpha=0.08}$	0	89	6	5
$\text{RIC}_{\alpha=0.1}$	0	91	5	4
$\text{RIC}_{\alpha=0.15}$	0	93	5	2
$\text{RIC}_{\alpha=0.2}$	0	94	4	2
$\text{RIC}_{\alpha=0.3}$	0	96	3	1
$\text{RIC}_{\alpha=0.5}$	0	99	1	0
$\text{RIC}_{\alpha=0.6}$	1	98	1	0

Table 5.8: Results for simulation set V. Generating model for the state process: AR(2). In each sample, 5% of the observations are additively perturbed by a magnitude of 8. Sample size is 100.

	AR Order			
Selection Criterion	1	2	3	4
AIC	4	71	10	15
$RIC_{\alpha=0.01}$	4	73	8	15
$RIC_{\alpha=0.05}$	2	84	9	5
$RIC_{\alpha=0.08}$	1	86	9	4
$RIC_{\alpha=0.1}$	2	87	7	4
$RIC_{\alpha=0.15}$	2	92	4	2
$RIC_{\alpha=0.2}$	2	94	3	1
$RIC_{\alpha=0.3}$	2	95	2	1
$RIC_{\alpha=0.5}$	5	94	0	1
$RIC_{\alpha=0.6}$	12	87	0	1

Table 5.9: Results for simulation set VI. Generating model for the state process: AR(2). In each sample, 5% of the observations are additively perturbed by a magnitude shift of 18. Sample size is 100.

	AR Order			
Selection Criterion	1	2	3	4
AIC	26	40	20	14
$RIC_{\alpha=0.01}$	27	47	12	14
$RIC_{\alpha=0.05}$	34	54	8	4
$RIC_{\alpha=0.08}$	11	74	10	5
$RIC_{\alpha=0.1}$	9	73	10	8
$RIC_{\alpha=0.15}$	8	77	10	5
$RIC_{\alpha=0.2}$	10	79	7	4
$RIC_{\alpha=0.3}$	12	80	6	2
$RIC_{\alpha=0.5}$	21	74	3	2
$RIC_{\alpha=0.6}$	28	68	3	1

Table 5.10: Results for simulation set VII. Generating model for the state process: AR(3). No observations are perturbed. Sample size is 50.

	AR Order			
Selection Criterion	1	2	3	4
AIC	0	4	76	20
$\text{RIC}_{\alpha=0.01}$	0	4	76	20
$\text{RIC}_{\alpha=0.05}$	1	4	80	15
$\text{RIC}_{\alpha=0.08}$	3	4	82	11
$\text{RIC}_{\alpha=0.1}$	3	5	83	9
$\text{RIC}_{\alpha=0.15}$	4	7	81	8
$\text{RIC}_{\alpha=0.2}$	5	8	81	6
$\text{RIC}_{\alpha=0.3}$	10	9	75	6
$\text{RIC}_{\alpha=0.5}$	17	15	65	3
$\text{RIC}_{\alpha=0.6}$	21	22	55	2

Table 5.11: Results for simulation set VIII. Generating model for the state process: AR(3). In each sample, 5% of the observations are additively perturbed by a magnitude shift of 8. Sample size is 50.

	AR Order			
Selection Criterion	1	2	3	4
AIC	10	16	49	25
$\text{RIC}_{\alpha=0.01}$	10	18	51	21
$\text{RIC}_{\alpha=0.05}$	14	17	50	19
$\text{RIC}_{\alpha=0.08}$	14	15	51	20
$\text{RIC}_{\alpha=0.1}$	13	14	55	18
$\text{RIC}_{\alpha=0.15}$	15	14	58	13
$\text{RIC}_{\alpha=0.2}$	15	16	58	11
$\text{RIC}_{\alpha=0.3}$	17	15	59	9
$\text{RIC}_{\alpha=0.5}$	23	18	54	5
$\text{RIC}_{\alpha=0.6}$	29	23	46	2

Table 5.12: Results for simulation set IX. Generating model for the state process: AR(3). In each sample, 5% of the observations are additively perturbed by a magnitude shift of 18. Sample size is 50.

	AR Order			
Selection Criterion	1	2	3	4
AIC	39	25	26	10
$\text{RIC}_{\alpha=0.01}$	37	30	21	12
$\text{RIC}_{\alpha=0.05}$	35	27	26	12
$\text{RIC}_{\alpha=0.08}$	19	18	44	19
$\text{RIC}_{\alpha=0.1}$	15	17	48	20
$\text{RIC}_{\alpha=0.15}$	15	12	53	20
$\text{RIC}_{\alpha=0.2}$	16	14	58	12
$\text{RIC}_{\alpha=0.3}$	24	15	53	8
$\text{RIC}_{\alpha=0.5}$	29	19	47	5
$\text{RIC}_{\alpha=0.6}$	36	21	40	3

Table 5.13: Results for simulation set X. Generating model for the state process: AR(3). No observations are perturbed. Sample size is 100.

	AR Order			
Selection Criterion	1	2	3	4
AIC	0	1	76	23
$RIC_{\alpha=0.01}$	0	1	77	22
$RIC_{\alpha=0.05}$	0	1	81	18
$RIC_{\alpha=0.08}$	0	1	82	17
$RIC_{\alpha=0.1}$	0	1	83	16
$RIC_{\alpha=0.15}$	0	3	83	14
$RIC_{\alpha=0.2}$	0	3	85	12
$RIC_{\alpha=0.3}$	0	4	89	7
$RIC_{\alpha=0.5}$	2	4	93	1
$RIC_{\alpha=0.6}$	2	5	92	1

Table 5.14: Results for simulation set XI. Generating model for the state process: AR(3). In each sample, 5% of the observations are additively perturbed by a magnitude shift of 8. Sample size is 100.

	AR Order			
Selection Criterion	1	2	3	4
AIC	1	8	65	26
$RIC_{\alpha=0.01}$	0	8	68	24
$RIC_{\alpha=0.05}$	0	7	71	22
$RIC_{\alpha=0.08}$	0	8	66	26
$RIC_{\alpha=0.1}$	1	8	67	24
$RIC_{\alpha=0.15}$	1	8	68	23
$RIC_{\alpha=0.2}$	2	8	68	22
$RIC_{\alpha=0.3}$	2	7	74	17
$RIC_{\alpha=0.5}$	5	8	77	10
$RIC_{\alpha=0.6}$	7	13	72	8

Table 5.15: Results for simulation set XII. Generating model for the state process: AR(3). In each sample, 5% of the observations are additively perturbed by a magnitude shift of 18. Sample size is 100.

	AR Order			
Selection Criterion	1	2	3	4
AIC	28	20	30	22
$RIC_{\alpha=0.01}$	21	23	33	23
$RIC_{\alpha=0.05}$	27	25	33	15
$RIC_{\alpha=0.08}$	11	16	46	27
$RIC_{\alpha=0.1}$	5	12	51	32
$RIC_{\alpha=0.15}$	3	8	58	31
$RIC_{\alpha=0.2}$	4	12	58	26
$RIC_{\alpha=0.3}$	6	12	64	18
$RIC_{\alpha=0.5}$	13	12	71	4
$RIC_{\alpha=0.6}$	12	14	72	2

From the preceding simulation study, we conclude that our proposed model selection criterion RIC_α with $\alpha = 0.15$ or 0.2 performs favorably compared to AIC, especially in instances where the data is contaminated.

5.3 Application

In this section, we revisit the application based on the cardiovascular mortality time series from the Los Angeles area. In Chapter 4, we used an AR(2) process to model the centered 6-day average cardiovascular mortality signal. We will use our proposed criterion RIC_α to further validate the choice of the AR order.

To select an AR order for the state process, candidate models based on AR orders of 1 through 4 are fit to the data. Models with higher AR orders are deemed to be unnecessarily complex. We consider the selection criteria AIC and RIC_α, with various α values employed for the latter: $0.01, 0.05, 0.08, 0.1, 0.15, 0.2, 0.3, 0.5, 0.6$.

Table 5.16 features the model selection results. For each criterion, the differences in criterion values (relative to the minimum) are featured for the four candidate state-space. Thus, the selected model corresponds to a criterion difference of zero. AIC selects AR(2) as the best model for the state process. With relatively small values for α (less than 0.6), RIC_α also selects AR(2). However, RIC_α with $\alpha = 0.6$ selects AR(1), which is not too surprising – the bias correction term tends to penalize excessively for more complex models if α is too far away from 0. The agreement of the selection results from AIC and RIC_α with small α values is not surprising, since the degree of contamination in the data is not severe. In Chapter 6, we will diagnose the influential data points in the series.

Table 5.16: Differences in criterion values (relative to the criterion minimum) for candidate state-space models for the cardiovascular mortality application.

	AR Order (State Process)			
Selection Criterion	1	2	3	4
AIC	7.84	0	0.97	1.16
$\text{RIC}_{\alpha=0.01}$	7.54	0	0.98	1.16
$\text{RIC}_{\alpha=0.05}$	6.58	0	1.04	1.30
$\text{RIC}_{\alpha=0.08}$	5.90	0	1.07	1.38
$\text{RIC}_{\alpha=0.1}$	5.48	0	1.09	1.43
$\text{RIC}_{\alpha=0.15}$	4.53	0	1.15	1.57
$\text{RIC}_{\alpha=0.2}$	3.69	0	1.20	1.71
$\text{RIC}_{\alpha=0.3}$	2.32	0	1.30	2.00
$\text{RIC}_{\alpha=0.5}$	0.44	0	1.49	2.59
$\text{RIC}_{\alpha=0.6}$	0	0.19	1.79	3.07

CHAPTER 6
INFLUENCE DIAGNOSTICS USING THE BHHJ DISCREPANCY

An important aspect of statistical modeling involves the identification of cases that have an undue influence on certain inferential objectives. In the state-space modeling framework, since the recovery of the latent signal is often a primary objective, identifying cases that substantially impact state prediction is often of interest.

This chapter proposes two versions of an influence diagnostic in the state-space framework based on the BHHJ discrepancy. In particular, the modeling diagnostics identify influential data points by comparing the conditional densities of each state both with and without the corresponding data point. Such a measure is often referred to as a "case-deletion diagnostic" (Christensen, Pearson, and Johnson, 1992; Shi and Chen, 2008). Intuitively, a relatively large disparity between the two densities indicates that the deleted case exhibits a relatively high degree of predictive influence compared to other points in the series. In our procedural development, the disparity between the two densities is measured using the BHHJ discrepancy.

6.1 Procedural Development

In this section, we discuss in detail how we develop the two versions of the influence diagnostic for the state-space framework. For simplicity, we assume that the observations are all scalars.

In state-space modeling, prediction of the latent states can be accomplished using the one-step predictor, the Kalman filter, or the Kalman smoother. We develop two versions of our influence diagnostic, one based on the filter and the other based on the smoother. In the presentation of our development, we focus on the former. The development of the latter is analogous, and is briefly discussed.

First, let $y_1, y_2, ..., y_N$ denote the cases that comprise the composite sample

vector $\boldsymbol{Y}$. For $t = 1, 2, ..., N$, let $\boldsymbol{Y}[t]$ denote the data $\boldsymbol{Y}$ with the t^{th} case y_t deleted, and let $\boldsymbol{Y}_{1,2,...t}$ denote the collection of cases $y_1, y_2, ..., y_t$. Let α^* denote the choice of the estimation tuning parameter based on $\text{MSE}^{\mu,\sigma}_{\text{std}}$. Let $\widehat{\boldsymbol{\Theta}}_{\alpha^*}$ denote the BHHJ MDE with α^* based on the full data $\boldsymbol{Y}$, and let $\widehat{\boldsymbol{\Theta}}_{\alpha^*}[t]$ denote the case-deleted BHHJ MDE with α^* based on $\boldsymbol{Y}[t]$.

In Kalman filtering, for $t = 1, 2, ..., N$, if we make use of the current observation y_t, then the predictor for the signal $\boldsymbol{x}_t$ is based on the mean of the conditional density of $\boldsymbol{x}_t$ given $\boldsymbol{Y}_{1,...,t}$:

$$f(\boldsymbol{x}_t|\boldsymbol{Y}_{1,...,t}; \widehat{\boldsymbol{\Theta}}_{\alpha^*}). \tag{6.1}$$

This conditional density corresponds to

$$\mathcal{N}(\boldsymbol{x}_t^t(\widehat{\boldsymbol{\Theta}}_{\alpha^*}), P_t^t(\widehat{\boldsymbol{\Theta}}_{\alpha^*})), \tag{6.2}$$

where the parameter vector is specified as the BHHJ MDE $\widehat{\boldsymbol{\Theta}}_{\alpha^*}$. If we do not use the current observation y_t, then the Kalman filter reduces to the one-step predictor. In this case, the predictor for the signal $\boldsymbol{x}_t$ is based on the mean of the conditional density of $\boldsymbol{x}_t$ given $\boldsymbol{Y}_{1,...,t-1}$:

$$f(\boldsymbol{x}_t|\boldsymbol{Y}_{1,...,t-1}; \widehat{\boldsymbol{\Theta}}_{\alpha^*}[t]). \tag{6.3}$$

This conditional density corresponds to

$$\mathcal{N}(\boldsymbol{x}_t^{t-1}(\widehat{\boldsymbol{\Theta}}_{\alpha^*}[t]), P_t^{t-1}(\widehat{\boldsymbol{\Theta}}_{\alpha^*}[t])) \tag{6.4}$$

where the parameter vector is specified as $\widehat{\boldsymbol{\Theta}}_{\alpha^*}[t]$.

The case-specific influence diagnostic based on the Kalman filter is then reflected by the BHHJ discrepancy between

$$f(\boldsymbol{x}_t|\boldsymbol{Y}_{1,...,t}; \widehat{\boldsymbol{\Theta}}_{\alpha^*}) \tag{6.5}$$

and

$$f(\boldsymbol{x}_t|\boldsymbol{Y}_{1,\ldots,t-1};\widehat{\boldsymbol{\Theta}}_{\alpha^*}[t]). \tag{6.6}$$

We use the same tuning parameter α^* in the computation of the diagnostic as in the estimation of the parameters. We define the *Kalman filter based predictive influence function* (PIF_f) for assessing the influence of the observed case y_t on the filter for the corresponding state $\boldsymbol{x}_t$ as

$$\begin{aligned}
\text{PIF}_f(t) &= \Delta^{\alpha^*}_{BHHJ}(f(\boldsymbol{x}_t|\boldsymbol{Y}_{1,\ldots,t};\widehat{\boldsymbol{\Theta}}_{\alpha^*}), f(\boldsymbol{x}_t|\boldsymbol{Y}_{1,\ldots,t-1};\widehat{\boldsymbol{\Theta}}_{\alpha^*}[t])) \\
&= \int \Bigg\{ f^{1+\alpha^*}(\boldsymbol{x}_t|\boldsymbol{Y}_{1,\ldots,t-1};\widehat{\boldsymbol{\Theta}}_{\alpha^*}[t]) \\
&\quad - \left(1+\frac{1}{\alpha^*}\right) f(\boldsymbol{x}_t|\boldsymbol{Y}_{1,\ldots,t};\widehat{\boldsymbol{\Theta}}_{\alpha^*}) f^{\alpha^*}(\boldsymbol{x}_t|\boldsymbol{Y}_{1,\ldots,t-1};\widehat{\boldsymbol{\Theta}}_{\alpha^*}[t]) \\
&\quad + \frac{1}{\alpha^*} f^{1+\alpha^*}(\boldsymbol{x}_t|\boldsymbol{Y}_{1,\ldots,t};\widehat{\boldsymbol{\Theta}}_{\alpha^*}) \Bigg\} d\boldsymbol{x}_t.
\end{aligned} \tag{6.7}$$

The subscript "f" in $\text{PIF}_f(t)$ denotes the reliance on the filter for prediction. Since both $f(\boldsymbol{x}_t|\boldsymbol{Y}_{1,\ldots,t};\widehat{\boldsymbol{\Theta}}_{\alpha^*})$ and $f(\boldsymbol{x}_t|\boldsymbol{Y}_{1,\ldots,t-1};\widehat{\boldsymbol{\Theta}}_{\alpha^*}[t])$ are normal distributions, we can evaluate $\text{PIF}_f(t)$ by using the following theorem, for $t = 1, 2, ..., N$.

Theorem 8. *If $f(\boldsymbol{x})$ and $g(\boldsymbol{y})$ are two n dimensional multivariate normal distribution functions with $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \Sigma_x)$ and $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_y, \Sigma_y)$, then for any $\alpha \in \mathbb{R}^+$,*

$$\begin{aligned}
\Delta^{\alpha}_{BHHJ}(g, f) &= \left(\frac{1}{1+\alpha}\right)^{\frac{n}{2}} (2\pi)^{-\frac{n\alpha}{2}} |\Sigma_x|^{-\frac{\alpha}{2}} \\
&\quad - \left(1+\frac{1}{\alpha}\right) (2\pi)^{-\frac{n\alpha}{2}} |\Sigma_x|^{-\frac{\alpha}{2}} |\Sigma_y|^{-\frac{1}{2}} \left(\frac{1}{|\Sigma_y^{-1} + \alpha\Sigma_x^{-1}|}\right)^{\frac{1}{2}} \times \\
&\quad \exp\Bigg[-\frac{1}{2}\left(\boldsymbol{\mu}_y^T\Sigma_y^{-1}\boldsymbol{\mu}_y + \alpha\boldsymbol{\mu}_x^T\Sigma_x^{-1}\boldsymbol{\mu}_x\right) \\
&\qquad + \frac{1}{2}\left(\boldsymbol{\mu}_y^T\Sigma_y^{-1} + \alpha\boldsymbol{\mu}_x^T\Sigma_x^{-1}\right)\left(\Sigma_y^{-1} + \alpha\Sigma_x^{-1}\right)^{-1}\left(\boldsymbol{\mu}_y^T\Sigma_y^{-1} + \alpha\boldsymbol{\mu}_x^T\Sigma_x^{-1}\right)^T \Bigg] \\
&\quad + \frac{1}{\alpha}\left(\frac{1}{1+\alpha}\right)^{\frac{n}{2}} (2\pi)^{-\frac{n\alpha}{2}} |\Sigma_y|^{-\frac{\alpha}{2}}.
\end{aligned} \tag{6.8}$$

Proof. See Appendix. □

The magnitude of $\text{PIF}_f(t)$ will reflect the divergence of $f(\boldsymbol{x}_t|\boldsymbol{Y}_{1,\ldots,t};\widehat{\boldsymbol{\Theta}}_{\alpha^*})$ from

$f(\boldsymbol{x}_t|\boldsymbol{Y}_{1,\ldots,t-1};\widehat{\boldsymbol{\Theta}}_{\alpha^*}[t])$. The larger the value of $\text{PIF}_f(t)$, the more influential the case y_t in the prediction of the corresponding state $\boldsymbol{x}_t$. This version of the influence diagnostic based on the BHHJ discrepancy could be applied when one is interested in making inferences based upon the most recent data point.

Similarly, the other version of the influence diagnostic can be developed if one chooses to make inferences based on the whole of the time series. In this version, the Kalman smoother is used to predict the state. Subsequently, we change the conditional distributions of interest to the two conditional distributions

$$f(\boldsymbol{x}_t|\boldsymbol{Y};\widehat{\boldsymbol{\Theta}}_{\alpha^*}) \tag{6.9}$$

and

$$f(\boldsymbol{x}_t|\boldsymbol{Y}[t];\widehat{\boldsymbol{\Theta}}_{\alpha^*}[t]), \tag{6.10}$$

which respectively correspond to

$$\mathcal{N}(\boldsymbol{x}_t^N(\widehat{\boldsymbol{\Theta}}_{\alpha^*}), P_t^N(\widehat{\boldsymbol{\Theta}}_{\alpha^*})) \tag{6.11}$$

and

$$\mathcal{N}(\boldsymbol{x}_t^{N[t]}(\widehat{\boldsymbol{\Theta}}_{\alpha^*}[t]), P_t^{N[t]}(\widehat{\boldsymbol{\Theta}}_{\alpha^*}[t])). \tag{6.12}$$

The notation $N[t]$ in (6.12) refers to the exclusion of y_t in obtaining the Kalman smoother. We implement the missing value modification for the Kalman smoother as described in Section 6.4 of Shumway and Stoffer (2010). Similar to PIF_f, we define the *Kalman smoother based predictive influence function* (PIF_s) for assessing the influence of the observed case y_t on the Kalman smoother of the corresponding state $\boldsymbol{x}_t$ as

$$\begin{aligned}\text{PIF}_s(t) &= \Delta_{BHHJ}^{\alpha^*}(f(\boldsymbol{x}_t|\boldsymbol{Y};\widehat{\boldsymbol{\Theta}}_{\alpha^*}), f(\boldsymbol{x}_t|\boldsymbol{Y}[t];\widehat{\boldsymbol{\Theta}}_{\alpha^*}[t])) \\ &= \int \Big\{ f^{1+\alpha^*}(\boldsymbol{x}_t|\boldsymbol{Y}[t];\widehat{\boldsymbol{\Theta}}_{\alpha^*}[t])\end{aligned}$$

$$- \left(1 + \frac{1}{\alpha^*}\right) f(\boldsymbol{x}_t|\boldsymbol{Y}; \widehat{\boldsymbol{\Theta}}_{\alpha^*}) f^{\alpha^*}(\boldsymbol{x}_t|\boldsymbol{Y}[t]; \widehat{\boldsymbol{\Theta}}_{\alpha^*}[t])$$
$$+ \frac{1}{\alpha^*} f^{1+\alpha^*}(\boldsymbol{x}_t|\boldsymbol{Y}; \widehat{\boldsymbol{\Theta}}_{\alpha^*}) \Big\} d\boldsymbol{x}_t. \tag{6.13}$$

The subscript "s" in $\text{PIF}_s(t)$ denotes the reliance on the Kalman smoother for prediction. Again, both $f(\boldsymbol{x}_t|\boldsymbol{Y}; \widehat{\boldsymbol{\Theta}}_{\alpha^*})$ and $f(\boldsymbol{x}_t|\boldsymbol{Y}[t]; \widehat{\boldsymbol{\Theta}}_{\alpha^*}[t])$ are normal distributions; therefore, Theorem 8 can also be applied to evaluate $\text{PIF}_s(t)$, for $t = 1, 2, ..., N$.

Next, we present a simulation study to illustrate the effectiveness of the two versions of the proposed influence diagnostic in the state-space framework.

6.2 Simulation Study

In this simulation study, we hope to illustrate that our proposed modeling diagnostics are effective in identifying deliberately perturbed data points. Specifically, the generating model for each Monte Carlo trial is

$$y_t = x_t + v_t, \tag{6.14}$$

$$x_t = \phi x_{t-1} + \omega_t, \tag{6.15}$$

where $\phi = 0.8$, $v_t \sim N(0, R = \sigma_v^2 = 1)$, and $\omega_t \sim N(0, Q = \sigma_\omega^2 = 1)$. After generating each replicated sample, we perturb three pre-determined observations by additively shifting each data point by a magnitude of 5. The length of the time series is 100, and the indices of the perturbed observations in each sample are 10, 39, and 58. The number of replications is 100.

For each trial, we calculate $\text{PIF}_f(t)$ and $\text{PIF}_s(t)$, for $t = 1, 2, ..., 100$; we then compute the average of $\text{PIF}_f(t)$ and $\text{PIF}_s(t)$ over the 100 trials at each time point. By inspecting the average values PIF_f and PIF_s, we anticipate that the influence of any anomalous values that appear due to chance in specific samples should be dissipated. Thus, only the deliberately perturbed values should be highlighted.

The averaged $\text{PIF}_f(t)$ and $\text{PIF}_s(t)$ are displayed in Figure 6.1 and 6.2, respectively. We see that cases corresponding to time indices 10, 39, and 58 are identified as influential by both PIF_f and PIF_s. In fact, if we compare the diagnostic patterns in Figure 6.1 and 6.2, both diagnostics perform very similarly in this simulation study.

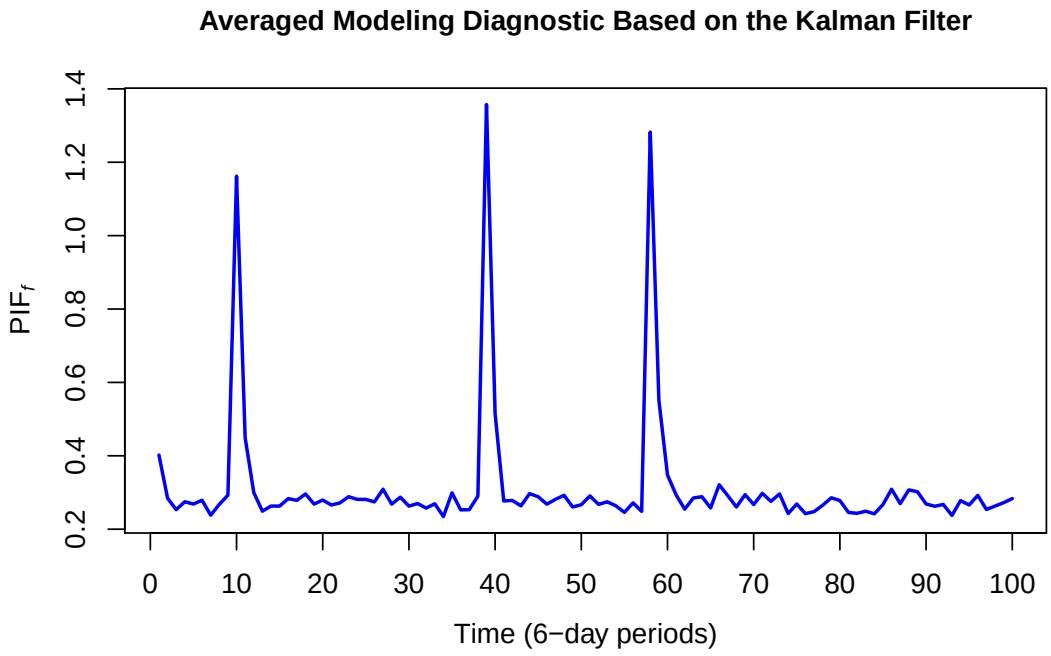

Figure 6.1: Simulation results based on $\text{PIF}_f(t)$. Note that cases corresponding to time indices 10, 39, and 58 are diagnosed.

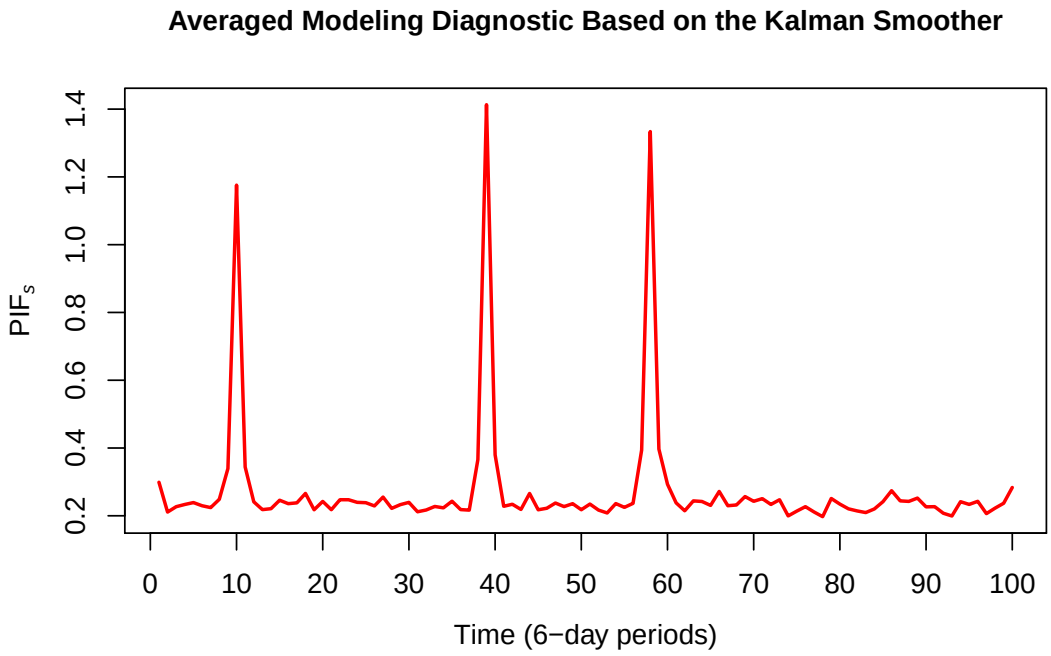

Figure 6.2: Simulation result based on $\text{PIF}_s(t)$. Note that cases corresponding to time indices 10, 39, and 58 are diagnosed.

In this study, we successfully identify the three deliberately perturbed observations by using our proposed influence diagnostics. However, in practice, we acknowledge that the mechanism of the perturbation might be more complex as well as more subtle. Thus, in the next section, we apply our methodology to a real data set.

6.3 Application

In this section, we revisit the cardiovascular mortality application. Based on the model selection results in Chapter 5, we model the signal using an AR(2) process. From Chapter 4, the tuning parameter is chosen as $\alpha = 0.180$. For $t = 1, 2, ..., 180$, we calculate $\text{PIF}_f(t)$ and $\text{PIF}_s(t)$. Again, the larger the value of $\text{PIF}_f(t)$ or $\text{PIF}_s(t)$, the more influential the case y_t in predicting the corresponding state $\boldsymbol{x}_t$.

Figure 6.3 and Figure 6.4 respectively feature the plot of $\text{PIF}_f(t)$ and $\text{PIF}_s(t)$ against the time index t. Note that cases 77, 91, and 151 appear to be influential in estimating the corresponding states. In particular, case 151 is the most influential point based on PIF_f, and case 77 is the most influential point based on PIF_s. The plot of the cardiovascular mortality series, with the three diagnosed observations marked as "$*$", is featured in Figure 6.5. A close inspection of Figure 6.5 reveals that case 77 corresponds to an unusually high spike that appears during the low part of the cardiovascular mortality cycle in the second year; case 91, on the other hand, corresponds to a relatively low value that appears during a subsequent rise in the mortality cycle. Case 151 is likely diagnosed due to its extremely high magnitude, defining the peak of the cycle in the third year.

We point out that Cavanaugh and Oleson (2001) also diagnosed cases 77, 91, and 151 using the same series. They used a pure AR(2) process to model the observations. Further, they assessed the influence of each observation by comparing the conditional densities of six future values (cases 181 to 186) both with and without a particular case. To measure the disparity between densities, they employed the K-L discrepancy. The fact that the diagnostic result based on our method agrees with theirs further validates the effectiveness of our proposed influence diagnostics.

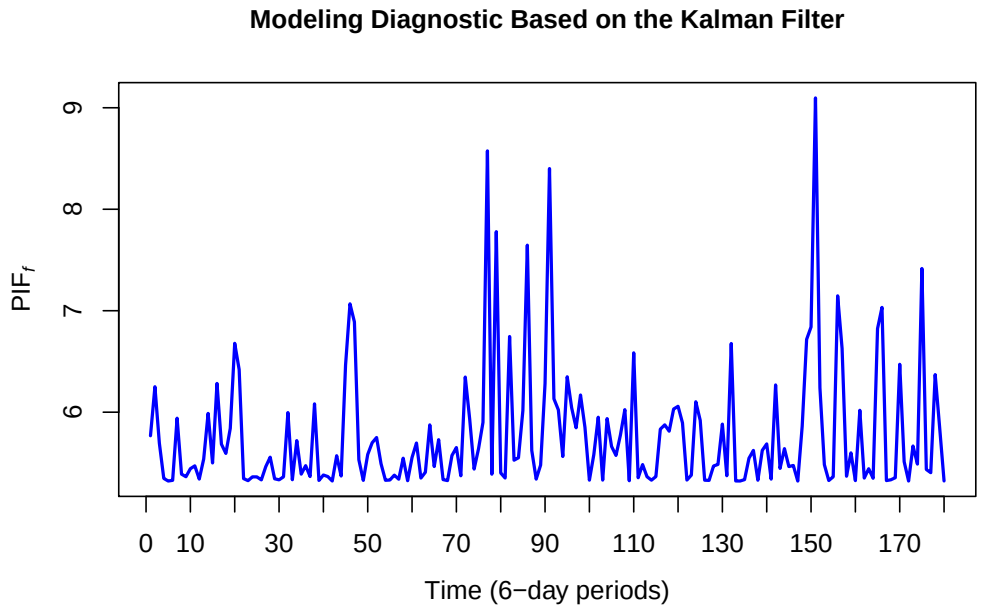

Figure 6.3: State-space influence diagnostic based on the Kalman filter.

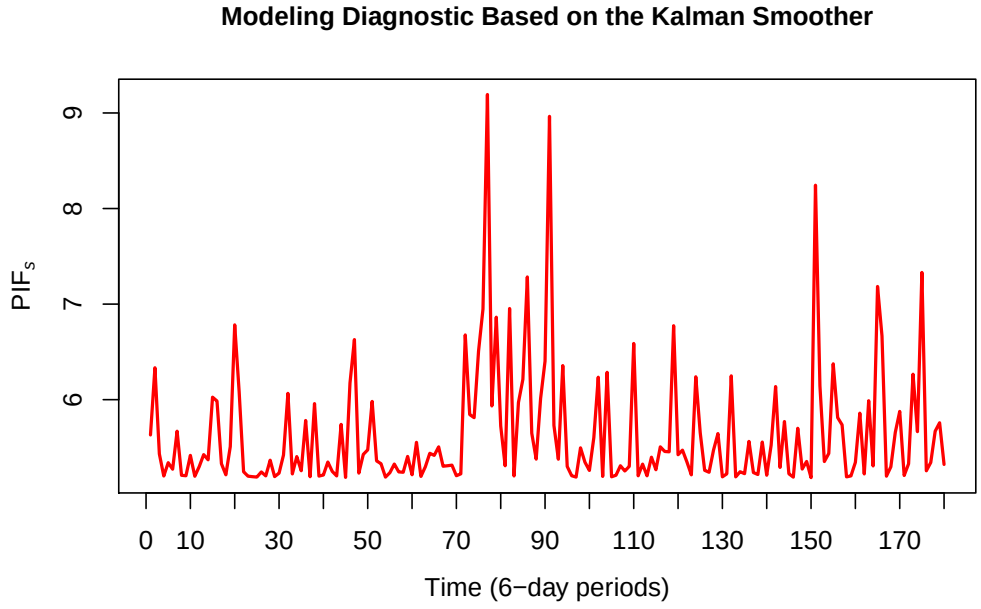

Figure 6.4: State-space influence diagnostic based on the Kalman smoother.

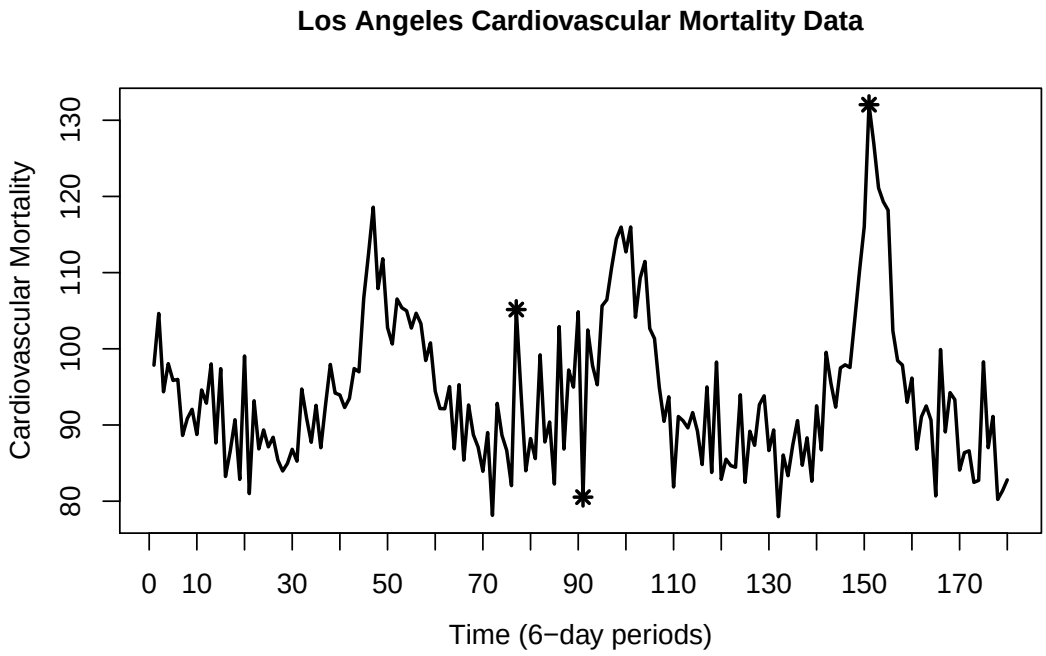

Figure 6.5: Three-year segment of the 1970's Los Angeles cardiovascular mortality incidence series. Cases 77, 91, and 151 are identified as influential points.

CHAPTER 7
CONCLUSIONS AND DISCUSSION

7.1 Conclusions and Discussion

Parameter estimation is often a formidable task in statistical modeling, particularly in frameworks where the model formulations might be complex. The traditional maximum likelihood approach suffers from a lack of robustness when the data are contaminated. For fitting state-space time series models, this thesis has proposed a new parameter estimation procedure that balances efficiency and robustness. The procedure is based on the minimization of the empirical BHHJ discrepancy. In the implementation of the method, as the non-negative tuning parameter α becomes larger, the estimator becomes more robust but less efficient.

Since the asymptotic variance of a parameter estimator is generally of interest for large-sample inferential procedures, we have also proposed a numerical method to approximate the asymptotic variance of the BHHJ estimator. Our procedure is specifically designed for state-space models. Another possible approach to estimating the variance is to employ bootstrapping (Shumway and Stoffer, 2010, Section 6.7). However, in the state-space setting, bootstrapping requires intensive computation. Furthermore, in implementing the bootstrap, one would need to make a choice between parametric bootstrapping, semi-parametric bootstrapping, and non-parametric bootstrapping.

In practice, the choice of the tuning parameter α could be a challenging task. This thesis has provided a data-driven approach for automatically selecting a suitable value of α.

Based on the proposed estimation procedure, we have developed a model selection criterion in the state-space modeling framework. In instances where the data is contaminated, our criterion is shown to perform favorably compared to the

traditional Akaike information criterion (AIC). We note that we have developed selection criteria where both bootstrapping and cross-validation are employed to approximate the bias correction term (Efron, 1983, 1986; Cavanaugh and Neath, 2012). Unfortunately, neither method works well in simulation studies, and thus the details of the two methods are not included in this thesis. We are not certain as to why the bootstrapping method performs poorly; Cavanaugh and Shumway (1997) successfully applied semi-parametric bootstrapping to develop an AIC variant for state-space model selection. Thus, the reason bootstrapping fails in the present setting may be due to the reliance of the methodology on the BHHJ discrepancy as opposed to the K-L discrepancy. Alternatively, bootstrapping may fail to provide an adequate bias correction in the presence of contamination.

One possible problem with the cross-validation method is that time series data are not independent, thus violating an essential assumption behind the classical method. Although we have considered different cross validatory approaches designed to weaken the temporal correlation between data points, based on our simulation studies, an optimal approach did not emerge. A more detailed and comprehensive investigation may be needed to successfully apply the method of cross-validation for the development of a discrepancy-based model selection criterion.

Lastly, we have proposed two versions of an influence diagnostic in the state-space framework. Our diagnostics help to identify cases that impact the recovery of the latent signal, thereby providing initial guidance and insight for further exploration.

In our simulation studies, we note that the manner in which we introduce contamination is to perturb the observation noise at isolated time points; thus, the generated outliers are additive outliers. Another approach for contaminating observations in a state-space process is to introduce innovative outliers, which can be

done by perturbing the state noise. Further work should consider the performance of the proposed inferential tools based on the BHHJ discrepancy in instances where innovative outliers are present.

APPENDIX

Proof for Theorem 1.

Proof. Explicitly,

$$\begin{aligned}&\int f^{1+\alpha}(\boldsymbol{x})d\boldsymbol{x}\\ =&\frac{1}{(2\pi)^{\frac{n(1+\alpha)}{2}}|\Sigma|^{\frac{1+\alpha}{2}}}\int_{\mathbb{R}^n}\exp\left[-\frac{1+\alpha}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]d\boldsymbol{x}.\end{aligned}\tag{A.1}$$

Using the substitution $\boldsymbol{t}=\sqrt{1+\alpha}(\boldsymbol{x}-\boldsymbol{\mu})$, expression (A.1) reduces to

$$\begin{aligned}&\frac{1}{(2\pi)^{\frac{n(1+\alpha)}{2}}|\Sigma|^{\frac{1+\alpha}{2}}}\ \frac{1}{(1+\alpha)^{\frac{n}{2}}}\int_{\mathbb{R}^n}\exp\left(-\frac{1}{2}\boldsymbol{t}^T\Sigma^{-1}\boldsymbol{t}\right)d\boldsymbol{t}\\ =&\frac{1}{(2\pi)^{\frac{n(1+\alpha)}{2}}|\Sigma|^{\frac{1+\alpha}{2}}}\ \frac{1}{(1+\alpha)^{\frac{n}{2}}}(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}\\ =&\left(\frac{1}{1+\alpha}\right)^{\frac{n}{2}}(2\pi)^{-\frac{n\alpha}{2}}|\Sigma|^{-\frac{\alpha}{2}},\end{aligned}$$

which completes the proof. □

Proof for Theorem 3.

Proof. Since we treat $\boldsymbol{\epsilon}_1,\boldsymbol{\epsilon}_2,...,\boldsymbol{\epsilon}_N$ as i.i.d., we may denote the density function for $\boldsymbol{\epsilon}_t$ as $f(\boldsymbol{x})$, for $t=1,2,...,N$. We have

$$\begin{aligned}\delta_t^{'}(\overline{\boldsymbol{\Theta}})&=\frac{1}{1+\alpha}\frac{\partial\left\{\frac{1}{N}\int f^{1+\alpha}(\boldsymbol{z}|\boldsymbol{\Theta})d\boldsymbol{z}-\left(1+\frac{1}{\alpha}\right)\frac{1}{N}f^{\alpha}(\boldsymbol{\epsilon}_t|\boldsymbol{\Theta})\right\}}{\partial\boldsymbol{\Theta}}\Bigg|_{\boldsymbol{\Theta}=\overline{\boldsymbol{\Theta}}}\\ &=\frac{1}{N}\left\{\int f^{\alpha}(\boldsymbol{z})\frac{\partial f(\boldsymbol{z})}{\partial\boldsymbol{\Theta}}d\boldsymbol{z}-f^{\alpha-1}(\boldsymbol{\epsilon}_t)\frac{\partial f(\boldsymbol{\epsilon}_t)}{\partial\boldsymbol{\Theta}}\right\}\Bigg|_{\boldsymbol{\Theta}=\overline{\boldsymbol{\Theta}}}.\end{aligned}\tag{A.2}$$

Therefore,

$$E_g\left(\delta_t^{'}(\overline{\boldsymbol{\Theta}})\right)=\frac{1}{N}\int f^{\alpha}(\boldsymbol{z})\frac{\partial f(\boldsymbol{z})}{\partial\boldsymbol{\Theta}}d\boldsymbol{z}\Bigg|_{\boldsymbol{\Theta}=\overline{\boldsymbol{\Theta}}}-\frac{1}{N}\int g(\boldsymbol{z})f^{\alpha-1}(\boldsymbol{z})\frac{\partial f(\boldsymbol{z})}{\partial\boldsymbol{\Theta}}d\boldsymbol{z}\Bigg|_{\boldsymbol{\Theta}=\overline{\boldsymbol{\Theta}}}.\tag{A.3}$$

Since $\overline{\boldsymbol{\Theta}}$ minimizes $\Delta^{\alpha}_{BHHJ}(g, f)$, we have

$$\left.\frac{\partial \Delta^{\alpha}_{BHHJ}(g,f)}{\partial \boldsymbol{\Theta}}\right|_{\boldsymbol{\Theta}=\overline{\boldsymbol{\Theta}}} = \mathbf{0}, \tag{A.4}$$

which reduces to

$$\left.\int f^{\alpha}(\boldsymbol{z})\frac{\partial f(\boldsymbol{z})}{\partial \boldsymbol{\Theta}}d\boldsymbol{z}\right|_{\boldsymbol{\Theta}=\overline{\boldsymbol{\Theta}}} - \left.\int g(\boldsymbol{z})f^{\alpha-1}(\boldsymbol{z})\frac{\partial f(\boldsymbol{z})}{\partial \boldsymbol{\Theta}}d\boldsymbol{z}\right|_{\boldsymbol{\Theta}=\overline{\boldsymbol{\Theta}}} = \mathbf{0}. \tag{A.5}$$

Equation (A.5) thus establishes that $E_g\left(\delta_t^{'}(\overline{\boldsymbol{\Theta}})\right) = \mathbf{0}$, for all $t = 1, 2, ..., N$. □

Proof for Theorem 5.

Proof. We will only outline the proof for the asymptotic variance of the BHHJ MDE $\widehat{\mu}$; the asymptotic variance of the BHHJ MDE $\widehat{\sigma}$ follows the same development and thus we leave the proof to the reader.

The explicit form of the density function for the normal random variable is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{A.6}$$

Warwick (2002) showed that the asymptotic variance of $\sqrt{N}$ times the BHHJ MDE $\widehat{\mu}$ is $\widehat{J}_{\mu}^{-1}\widehat{K}_{\mu}\widehat{J}_{\mu}^{-1}$, where

$$\begin{aligned}\widehat{J}_{\mu} = &\int u_{\widehat{\mu}}(x)u_{\widehat{\mu}}^{T}(x)f^{\alpha+1}(x)dx - \int \left(i_{\widehat{\mu}}(x) - \alpha u_{\widehat{\mu}}(x)u_{\widehat{\mu}}^{T}(x)\right)f^{\alpha+1}(x)dx \\ &+ \frac{1}{N}\sum_{t=1}^{N}\left[\left(i_{\widehat{\mu}}(e_t) - \alpha u_{\widehat{\mu}}(e_t)u_{\widehat{\mu}}^{T}(e_t)\right)f^{\alpha}(e_t)\right],\end{aligned} \tag{A.7}$$

$$\widehat{K}_{\mu} = \frac{1}{N}\sum_{t=1}^{N} f^{2\alpha}(e_t)u_{\widehat{\mu}}(e_t)u_{\widehat{\mu}}^{T}(e_t) - \frac{1}{N^2}\left[\sum_{t=1}^{N} f^{\alpha}(e_t)u_{\widehat{\mu}}(e_t)\right]^2. \tag{A.8}$$

Here, $u_{\widehat{\mu}}(x)$ is the score function in terms of $\widehat{\mu}$, and $i_{\widehat{\mu}}(x)$ is the observed Fisher information function. It can be easily shown that $u_{\widehat{\mu}}(x) = \widehat{\sigma}^{-2}(x - \widehat{\mu})$, and $i_{\widehat{\mu}}(x) = \widehat{\sigma}^{-2}$.

For simplicity, we write $\widehat{J}_\mu = A - B + C$, where

$$A := \int u_{\widehat{\mu}}(x) u_{\widehat{\mu}}^T(x) f^{\alpha+1}(x) dx, \tag{A.9}$$

$$B := \int \left(i_{\widehat{\mu}}(x) - \alpha u_{\widehat{\mu}}(x) u_{\widehat{\mu}}^T(x) \right) f^{\alpha+1}(x) dx, \tag{A.10}$$

$$C := \frac{1}{N} \sum_{t=1}^{N} \left[\left(i_{\widehat{\mu}}(e_t) - \alpha u_{\widehat{\mu}}(e_t) u_{\widehat{\mu}}^T(e_t) \right) f^{\alpha}(e_t) \right]. \tag{A.11}$$

Substituting the expressions for $f(x)$, $u_{\widehat{\mu}}(x)$, and $i_{\widehat{\mu}}(x)$ into A, B, and C, we have

$$\begin{aligned} A &= \int \widehat{\sigma}^{-4} (x - \widehat{\mu})^2 \left[\frac{1}{\sqrt{2\pi}\widehat{\sigma}} \exp\left(-\frac{(x-\widehat{\mu})^2}{2\widehat{\sigma}^2} \right) \right]^{\alpha+1} dx \\ &= \widehat{\sigma}^{-5-\alpha} (\alpha+1)^{-\frac{3}{2}} (2\pi)^{-\frac{\alpha+1}{2}} \int t^2 \exp\left(-\frac{t^2}{2\widehat{\sigma}^2} \right) dt \\ &\quad \text{[using variable substitution } t = \sqrt{1+\alpha}(x - \widehat{\mu})] \\ &= \widehat{\sigma}^{-4-\alpha} (\alpha+1)^{-\frac{3}{2}} (2\pi)^{-\frac{\alpha}{2}} (\widehat{\mu}^2 + \widehat{\sigma}^2), \end{aligned} \tag{A.12}$$

$$\begin{aligned} B &= \int \widehat{\sigma}^{-2} (2\pi)^{-\frac{\alpha+1}{2}} \widehat{\sigma}^{-(\alpha+1)} \exp\left(-\frac{(\alpha+1)(x-\widehat{\mu})^2}{2\widehat{\sigma}^2} \right) dx - \alpha A \\ &= \int \widehat{\sigma}^{-2} (2\pi)^{-\frac{\alpha+1}{2}} \widehat{\sigma}^{-(\alpha+1)} \exp\left(-\frac{t^2}{2\widehat{\sigma}^2} \right) (\alpha+1)^{-\frac{1}{2}} dt - \alpha A \\ &\quad \text{[using variable substitution } t = \sqrt{1+\alpha}(x - \widehat{\mu})] \\ &= \widehat{\sigma}^{-(2+\alpha)} (2\pi)^{-\frac{\alpha}{2}} (\alpha+1)^{-\frac{1}{2}} - \alpha \widehat{\sigma}^{-4-\alpha} (\alpha+1)^{-\frac{3}{2}} (2\pi)^{-\frac{\alpha}{2}} (\widehat{\mu}^2 + \widehat{\sigma}^2), \end{aligned} \tag{A.13}$$

$$C = (2\pi)^{-\frac{\alpha}{2}} \widehat{\sigma}^{-(2+\alpha)} \left\{ \frac{1}{N} \sum_{t=1}^{N} \left[(1 - \alpha \widehat{\sigma}^{-2} (e_t - \widehat{\mu})^2) \exp\left(-\frac{\alpha (e_t - \widehat{\mu})^2}{2\widehat{\sigma}^2} \right) \right] \right\}. \tag{A.14}$$

The expression for $\widehat{J}_\mu$ then follows.

Next, substituting the expressions for $f(x)$ and $u_{\widehat{\mu}}(x)$ into $\widehat{K}_\mu$, we have

$$\begin{aligned} \widehat{K}_\mu &= \frac{1}{N} \sum_{t=1}^{N} \left[(2\pi)^{-\frac{1}{2}} \widehat{\sigma}^{-1} \exp\left(-\frac{(e_t - \widehat{\mu})^2}{2\widehat{\sigma}^2} \right) \right]^{2\alpha} \widehat{\sigma}^{-4} (e_t - \widehat{\mu})^2 \\ &\quad - \frac{1}{N^2} \left[\sum_{t=1}^{N} (2\pi)^{-\frac{\alpha}{2}} \widehat{\sigma}^{-\alpha} \exp\left(-\frac{\alpha (e_t - \widehat{\mu})^2}{2\widehat{\sigma}^2} \right) \widehat{\sigma}^{-2} (e_t - \widehat{\mu}) \right]^2 \\ &= \frac{1}{N} \sum_{t=1}^{N} \left[(2\pi)^{-\alpha} \widehat{\sigma}^{-2\alpha-4} \exp\left(-\frac{\alpha (e_t - \widehat{\mu})^2}{\widehat{\sigma}^2} \right) (e_t - \widehat{\mu})^2 \right] \end{aligned}$$

$$-\frac{1}{N^2}\left\{\sum_{t=1}^{N}\left[(2\pi)^{-\frac{\alpha}{2}}\widehat{\sigma}^{-(\alpha+2)}\exp\left(-\frac{\alpha(e_t-\widehat{\mu})^2}{2\widehat{\sigma}^2}\right)(e_t-\widehat{\mu})\right]\right\}^2, \qquad \text{(A.15)}$$

which completes the proof for the explicit asymptotic variance form for the BHHJ MDE $\widehat{\mu}$. The proof for the explicit asymptotic variance form for the BHHJ MDE $\widehat{\sigma}$ follows similar reasoning with minor changes. □

Proof for Theorem 6.

Proof. It suffices to prove that

$$\textcircled{2}+\textcircled{3} = E_g\left\{\frac{1}{N}(1+\alpha)\text{tr}\left(\widehat{J}^{-1}(\widehat{\boldsymbol{\Theta}}_k)\widehat{K}(\widehat{\boldsymbol{\Theta}}_k)\right)\right\}+o(1).$$

The proof closely follows the development of the model selection criterion TIC. Specifically, we have

$$\begin{aligned}
&E_g\left\{\frac{1}{N}(1+\alpha)\text{tr}\left(\widehat{J}^{-1}(\widehat{\boldsymbol{\Theta}}_k)\widehat{K}(\widehat{\boldsymbol{\Theta}}_k)\right)\right\}\\
&=\frac{1}{N}(1+\alpha)E_g\left\{\text{tr}\left(\widehat{J}^{-1}(\widehat{\boldsymbol{\Theta}}_k)\widehat{K}(\widehat{\boldsymbol{\Theta}}_k)\right)\right\}\\
&=\frac{1}{N}(1+\alpha)E_g\left\{\text{tr}\left(J^{-1}(\widehat{\boldsymbol{\Theta}}_k)K(\widehat{\boldsymbol{\Theta}}_k)\right)\right\}+o(1)\\
&\quad\text{[because } \widehat{J} \text{ and } \widehat{K} \text{ are estimates for } J \text{ and } K\text{, respectively; the weak law}\\
&\quad\text{of large numbers (WLLN) is applied]}\\
&=\frac{1}{N}(1+\alpha)E_g\left\{\text{tr}\left(J^{-1}(\overline{\boldsymbol{\Theta}}_k)K(\overline{\boldsymbol{\Theta}}_k)\right)\right\}+o(1)\\
&\quad\text{[because } \widehat{\boldsymbol{\Theta}}_k \text{ converges to } \overline{\boldsymbol{\Theta}}_k \text{ in probability]}\\
&=\frac{1}{N}(1+\alpha)\text{tr}\left(J^{-1}(\overline{\boldsymbol{\Theta}}_k)K(\overline{\boldsymbol{\Theta}}_k)\right)+o(1)\\
&=\frac{1}{N}(1+\alpha)\text{tr}\left(K(\overline{\boldsymbol{\Theta}}_k)J^{-1}(\overline{\boldsymbol{\Theta}}_k)\right)+o(1)\\
&=\frac{1}{N}(1+\alpha)\text{tr}\left(J(\overline{\boldsymbol{\Theta}}_k)J^{-1}(\overline{\boldsymbol{\Theta}}_k)K(\overline{\boldsymbol{\Theta}}_k)J^{-1}(\overline{\boldsymbol{\Theta}}_k)\right)+o(1)\\
&=(1+\alpha)\text{tr}\left(J(\overline{\boldsymbol{\Theta}}_k)E_g\left\{\left(\overline{\boldsymbol{\Theta}}_k-\widehat{\boldsymbol{\Theta}}_k\right)\left(\overline{\boldsymbol{\Theta}}_k-\widehat{\boldsymbol{\Theta}}_k\right)^T\right\}\right)+o(1)\\
&\quad\text{[because the asymptotic covariance matrix of } \widehat{\boldsymbol{\Theta}}_k \text{ is}\\
&\quad\frac{1}{N}J^{-1}(\overline{\boldsymbol{\Theta}}_k)K(\overline{\boldsymbol{\Theta}}_k)J^{-1}(\overline{\boldsymbol{\Theta}}_k)]
\end{aligned}$$

$$= (1+\alpha)E_g\left\{\left(\overline{\boldsymbol{\Theta}}_k - \widehat{\boldsymbol{\Theta}}_k\right)^T J(\overline{\boldsymbol{\Theta}}_k)\left(\overline{\boldsymbol{\Theta}}_k - \widehat{\boldsymbol{\Theta}}_k\right)\right\} + o(1)$$
$$= \frac{1}{2}(1+\alpha)E_g\left\{\left(\overline{\boldsymbol{\Theta}}_k - \widehat{\boldsymbol{\Theta}}_k\right)^T J(\overline{\boldsymbol{\Theta}}_k)\left(\overline{\boldsymbol{\Theta}}_k - \widehat{\boldsymbol{\Theta}}_k\right)\right\}$$
$$+ \frac{1}{2}(1+\alpha)E_g\left\{\left(\overline{\boldsymbol{\Theta}}_k - \widehat{\boldsymbol{\Theta}}_k\right)^T J(\overline{\boldsymbol{\Theta}}_k)\left(\overline{\boldsymbol{\Theta}}_k - \widehat{\boldsymbol{\Theta}}_k\right)\right\} + o(1)$$
$$= \textcircled{2} + \textcircled{3} + o(1)$$

[because of Lemma 1 and Lemma 2]. □

Proof for Theorem 7.

Proof. Based on Theorem 6, we need only show that expression (5.19) converges to expression (5.11) under the assumptions given in the theorem. Since g is a member of every candidate class $\mathcal{F}(k)$ for $k = k_1, k_2, ..., k_L$, we can replace g with $f(\cdot|\overline{\boldsymbol{\Theta}}_k)$. From equations (4.1) and (4.2) in Chapter 4, providing the explicit expressions for $J(\overline{\boldsymbol{\Theta}}_k)$ and $K(\overline{\boldsymbol{\Theta}}_k)$, it may be established that $J(\overline{\boldsymbol{\Theta}}_k) \approx K(\overline{\boldsymbol{\Theta}}_k)$ for α close to 0. Since $\widehat{J}(\widehat{\boldsymbol{\Theta}}_k)$ and $\widehat{K}(\widehat{\boldsymbol{\Theta}}_k)$ are estimates for $J(\overline{\boldsymbol{\Theta}}_k)$ and $K(\overline{\boldsymbol{\Theta}}_k)$, respectively, we could treat $\widehat{J}^{-1}(\widehat{\boldsymbol{\Theta}}_k)\widehat{K}(\widehat{\boldsymbol{\Theta}}_k)$ as the identity matrix with the dimension being the number of parameters k in the model. This completes the proof. □

Proof for Theorem 8.

Proof. To prove Theorem 8, we first present the following Lemma.

Lemma. *If C is an $n \times n$ symmetric matrix, $\boldsymbol{x}$ and $\boldsymbol{b}$ are both n dimensional vectors, and d is a scalar, the following result holds:*

$$\frac{1}{2}\boldsymbol{x}^T C\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + d = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T C(\boldsymbol{x}-\boldsymbol{m}) + v, \tag{A.16}$$

where $\boldsymbol{m} = -C^{-1}\boldsymbol{b}$, and $v = d - \frac{1}{2}\boldsymbol{b}^T C^{-1}\boldsymbol{b}$.

Proof. The proof is fairly straightforward and therefore omitted. □

Using the result from Theorem 1, it suffices to derive the explicit form of

$$\int\left\{-\left(1+\frac{1}{\alpha}\right)g(\boldsymbol{z})f^{\alpha}(\boldsymbol{z})\right\}d\boldsymbol{z}. \tag{A.17}$$

Specifically,

$$
\begin{aligned}
&\int \left\{ -\left(1+\frac{1}{\alpha}\right) g(\boldsymbol{z}) f^{\alpha}(\boldsymbol{z}) \right\} d\boldsymbol{z} \\
&= -\left(1+\frac{1}{\alpha}\right) (2\pi)^{-\frac{n}{2}} |\Sigma_y|^{-\frac{1}{2}} (2\pi)^{-\frac{\alpha n}{2}} |\Sigma_x|^{-\frac{\alpha}{2}} \\
&\quad \int \exp\left[-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_y)^T \Sigma_y^{-1} (\boldsymbol{z}-\boldsymbol{\mu}_y) - \frac{1}{2}\alpha (\boldsymbol{z}-\boldsymbol{\mu}_x)^T \Sigma_x^{-1} (\boldsymbol{z}-\boldsymbol{\mu}_x) \right] d\boldsymbol{z} \\
&= -\left(1+\frac{1}{\alpha}\right) (2\pi)^{-\frac{(\alpha+1)n}{2}} |\Sigma_y|^{-\frac{1}{2}} |\Sigma_x|^{-\frac{\alpha}{2}} \\
&\quad \int \exp\left\{ -\frac{1}{2}\Big[\boldsymbol{z}^T \left(\Sigma_y^{-1} + \alpha\Sigma_x^{-1}\right) \boldsymbol{z} - 2\left(\boldsymbol{\mu}_y^T \Sigma_y^{-1} + \alpha \boldsymbol{\mu}_x^T \Sigma_x^{-1}\right) \boldsymbol{z} \right. \right. \\
&\qquad \left. \left. + \boldsymbol{\mu}_y^T \Sigma_y^{-1} \boldsymbol{\mu}_y + \alpha \boldsymbol{\mu}_x^T \Sigma_x^{-1} \boldsymbol{\mu}_x \Big] \right\} d\boldsymbol{z} \\
&= -\left(1+\frac{1}{\alpha}\right) (2\pi)^{-\frac{(\alpha+1)n}{2}} |\Sigma_y|^{-\frac{1}{2}} |\Sigma_x|^{-\frac{\alpha}{2}} \\
&\quad \int \exp\left\{ -\left[\frac{1}{2} \boldsymbol{z}^T \left(\Sigma_y^{-1} + \alpha\Sigma_x^{-1}\right) \boldsymbol{z} + \left(-\Sigma_y^{-1}\boldsymbol{\mu}_y - \alpha\Sigma_x^{-1}\boldsymbol{\mu}_x\right)^T \boldsymbol{z} \right. \right. \\
&\qquad \left. \left. + \frac{1}{2}\left(\boldsymbol{\mu}_y^T \Sigma_y^{-1} \boldsymbol{\mu}_y + \alpha \boldsymbol{\mu}_x^T \Sigma_x^{-1} \boldsymbol{\mu}_x\right) \right] \right\} d\boldsymbol{z}. \qquad \text{(A.18)}
\end{aligned}
$$

In order to use the preceding Lemma, we define

$$C = \Sigma_y^{-1} + \alpha\Sigma_x^{-1}, \qquad \text{(A.19)}$$

$$\boldsymbol{b} = -\Sigma_y^{-1}\boldsymbol{\mu}_y - \alpha\Sigma_x^{-1}\boldsymbol{\mu}_x, \qquad \text{(A.20)}$$

$$d = \frac{1}{2}\left(\boldsymbol{\mu}_y^T \Sigma_y^{-1} \boldsymbol{\mu}_y + \alpha \boldsymbol{\mu}_x^T \Sigma_x^{-1} \boldsymbol{\mu}_x\right). \qquad \text{(A.21)}$$

It is easy to check that the matrix C is symmetric. Thus, using the result from the Lemma, we have

$$
\begin{aligned}
(A.17) = &-\left(1+\frac{1}{\alpha}\right) (2\pi)^{-\frac{(\alpha+1)n}{2}} |\Sigma_y|^{-\frac{1}{2}} |\Sigma_x|^{-\frac{\alpha}{2}} \\
&\exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}_y^T \Sigma_y^{-1} \boldsymbol{\mu}_y + \alpha \boldsymbol{\mu}_x^T \Sigma_x^{-1} \boldsymbol{\mu}_x\right) + \right. \\
&\left. \frac{1}{2}\left(\boldsymbol{\mu}_y^T \Sigma_y^{-1} + \alpha \boldsymbol{\mu}_x^T \Sigma_x^{-1}\right) \left(\Sigma_y^{-1} + \alpha\Sigma_x^{-1}\right)^{-1} \left(\boldsymbol{\mu}_y^T \Sigma_y^{-1} + \alpha \boldsymbol{\mu}_x^T \Sigma_x^{-1}\right)^T \right]
\end{aligned}
$$

$$\int \exp\left[-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{m})^T\left(\Sigma_y^{-1}+\alpha\Sigma_x^{-1}\right)(\boldsymbol{z}-\boldsymbol{m})\right]d\boldsymbol{z}, \tag{A.22}$$

where

$$\boldsymbol{m}=\left(\Sigma_y^{-1}+\alpha\Sigma_x^{-1}\right)^{-1}\left(\Sigma_y^{-1}\boldsymbol{\mu}_y+\alpha\Sigma_x^{-1}\boldsymbol{\mu}_x\right). \tag{A.23}$$

After some simplifications, we have

$$\begin{aligned}(A.17) = &-\left(\frac{1}{1+\alpha}\right)(2\pi)^{-\frac{n\alpha}{2}}|\Sigma_x|^{-\frac{\alpha}{2}}|\Sigma_y|^{-\frac{1}{2}}\left(\frac{1}{|\Sigma_y^{-1}+\alpha\Sigma_x^{-1}|}\right)^{\frac{1}{2}}\\ &\exp\Bigg[-\frac{1}{2}\left(\boldsymbol{\mu}_y^T\Sigma_y^{-1}\boldsymbol{\mu}_y+\alpha\boldsymbol{\mu}_x^T\Sigma_x^{-1}\boldsymbol{\mu}_x\right)\\ &\quad+\frac{1}{2}\left(\boldsymbol{\mu}_y^T\Sigma_y^{-1}+\alpha\boldsymbol{\mu}_x^T\Sigma_x^{-1}\right)\left(\Sigma_y^{-1}+\alpha\Sigma_x^{-1}\right)^{-1}\left(\boldsymbol{\mu}_y^T\Sigma_y^{-1}+\alpha\boldsymbol{\mu}_x^T\Sigma_x^{-1}\right)^T\Bigg].\end{aligned}$$

Using the result from Theorem 1, we have completed the proof of Theorem 8. □

REFERENCES

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, in: B. N. Petrov and F. Csaki, (Eds.). *2nd International Symposium on Information Theory* (Akademia Kiado, Budapest), 267–281.

Akaike, H. (1974), A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Ali, S. M. and Silvey, S. D. (1966), A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B* **28**, 131–142.

Anderson, B. D. O. and Moore, J. B. (1979), *Optimal Filtering* (Prentice-Hall, New Jersey).

Basu, A., Harris, I., Hjort, N., and Jones, M. (1998), Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85**, 549–559.

Bengtsson, T. and Cavanaugh, J. E. (2008), State-space discrimination and clustering of atmospheric time series data based on Kullback information measures. *Environmetrics* **19**, 103–121.

Bernoulli, D. (1777), Dijudicatio maxime probabilis plurium observationum discrepantium atque verisimillima inductio inde formanda. *Acta Acad. Sci. Petropolit* **1**, 3–33.

Bickel, P. J. and Lehmann, E. L. (1975), Descriptive statistics for nonparametric models II. Location. *The Annals of Statistics* **3**, 1045–1069.

Buckland, S. T., Newman, K. B., Thomas, L., and Koesters, N. B. (2004), State-space models for the dynamics of wild animal populations. *Ecological Modelling* **171**, 157–175.

Cavanaugh, J. E. and Johnson, W. O. (1999), Assessing the predictive influence of cases in a state space process. *Biometrika* **86**, 183–190.

Cavanaugh, J. E. and Neath, A. A. (2012), Model selection criteria based on computationally intensive estimators of the expected optimism. *Mathematics in Engineering, Science and Aerospace* **3**, 343–356.

Cavanaugh, J. E. and Oleson, J. O. (2001), A diagnostic for assessing the influence of cases on the prediction of missing data. *Journal of the Royal Statistical Society. Series D (The Statistician)* **50**, 427–440.

Cavanaugh, J. E. and Shumway, R. H. (1997), A bootstrap variant of AIC for state-space model selection. *Statistica Sinica* **7**, 473–496.

Christensen, R., Pearson, L., and Johnson, W. (1992), Case-deletion diagnostics for mixed models. *Technometrics* **34**, 38–45.

Cressie, N. and Read, T. R. C. (1984), Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)* **46**, 440–464.

Csiszár, I. (1963), Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten. *Magyar. Tud. Akad. Mat. Kutato Int. Kozl* **8**, 85–108.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Efron, B. (1983), Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association* **78**, 316–331.

Efron, B. (1986), How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* **81**, 461–470.

Fujisawa, H. and Eguchi, S. (2006), Robust estimation in the normal mixture model. *Journal of Statistical Planning and Inference* **136**, 3989–4011.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions* (John Wiley and Sons, New York).

Hanna, E. J. and Quinn, B. G. (1979), The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)* **41**, 190–195.

Harvey, A. C. (1991), *Forecasting, Structural Time Series Models and the Kalman Filter* (Cambridge University Press, Cambridge).

Hodges, J. L. and Lehmann, E. L. (1963), Estimates of location based on rank tests. *The Annals of Mathematical Statistics* **34**, 598–611.

Hong, C. and Kim, Y. (2001), Automatic selection of the tuning parameter in the minimum density power divergence estimation. *Journal of the Korean Statistical Association* **30**, 453–465.

Huber, P. J. (1964), Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35**, 73–101.

Hurvich, C. M. and Tsai, C.-L. (1989), Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

Ibragimov, M., Ibragimov, R., and Walden, J. (2015), *Heavy-Tailed Distributions and Robustness in Economics and Finance* (Springer, New York).

Jazwinski, A. H. (1970), *Stochastic Processes and Filtering Theory* (Academic Press, New York).

Jeffreys, H. (1946), An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **186**, 453–461.

Kalman, R. E. (1960), A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* **82**, 35–45.

Kalman, R. E. and Bucy, R. S. (1961), New results in linear filtering and prediction theory. *Journal of Basic Engineering* **83**, 95–108.

Kitagawa, G. (1987), Non-Gaussian state-space modeling of nonstationary time series: rejoinder. *Journal of the American Statistical Association* **82**, 1060–1063.

Kullback, S. and Leibler, R. A. (1951), On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79–86.

Linhart, H. and Zucchini, W. (1986), *Model Selection* (Wiley, New York).

Mantalos, P., Mattheou, K., and Karagrigoriou, A. (2010), An improved divergence information criterion for the determination of the order of an AR process. *Communications in Statistics - Simulation and Computation* **39**, 865–879.

Maronna, R., Martin, D., and Yohai, V. (2006), *Robust Statistics: Theory and Methods* (Wiley, New York).

Mattheou, K., Lee, S., and Karagrigoriou, A. (2009), A model selection criterion based on the BHHJ measure of divergence. *Journal of Statistical Planning and Inference* **139**, 228–235.

Morimoto, T. (1963), Markov processes and the H-theorem. *The Physical Society of Japan* **18**, 328–331.

Neath, A. A. and Cavanaugh, J. E. (2012), The Bayesian information criterion: background, derivation, and applications. *WIREs Computational Statistics* **4**, 199–203.

Pardo, L. (2005), *Statistical Inference Based on Divergence Measures* (Chapman and Hall/CRC, New York).

Peel, D. and McLachlan, G. J. (2000), Robust mixture modelling using the t distribution. *Statistics and Computing* **10**, 339–348.

Rindfuss, R. and Ladinsky, J. (1976), Patterns of births: implications for the incidence of elective induction. *Medical Care* **14**, 685–693.

Schwarz, G. (1978), Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Shi, L. and Chen, G. (1978), Case deletion diagnostics in multilevel models. *Journal of Multivariate Analysis* **99**, 1860–1877.

Shumway, R. H. and Stoffer, D. S. (1982), An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis* **3**, 253–264.

Shumway, R. H. and Stoffer, D. S. (2010), *Time Series Analysis and Its Applications: With R Examples (3rd edition)* (Springer, New York).

Takeuchi, K. (1976), Distribution of information statistics and criteria for adequacy of models. *Mathematical Sciences* **153**, 12–18 (in Japanese).

Takeuchi, I., Bengio, Y., and Kanamori, T. (2002), Robust regression with asymmetric heavy-tail noise distributions. *Neural Computation* **14**, 2469–2496.

Tandeo, P., Ailliot, P., and Autret, E. (2011), Linear Gaussian state-space model with irregular sampling: application to sea surface temperature. *Stochastic Environmental Research and Risk Assessment* **25**, 793–804.

Toma, A. (2014), Model selection criteria using divergences. *Entropy* **16**, 2686–2698.

Warwick, J. (2002), Selecting tuning parameters in minimum distance estimators. Ph.D. Thesis, The Open University.

Warwick, J. (2005), A data-based method for selecting tuning parameters in minimum distance estimators. *Computational Statistics and Data Analysis* **48**, 571–585.

Yang, M., Cavanaugh, J. E., and Zamba, G. K. (2015), State-space models for count time series with excess zeros. *Statistical Modelling* **15**, 70–90.

Ypma, T. (1995), Historical development of the Newton-Raphson method. *SIAM Review* **37**, 531–551.

Zeng, Y. and Wu, S. (2013), *State-Space Models: Applications in Economics and Finance* (Springer, New York).