ABSTRACT

Topics in Multivariate Covariance Estimation
and Time Series Analysis

John D. Beeson, Ph.D.

Chairperson: Jane L. Harvill, Ph.D.

In this dissertation we will discuss two topics relevant to statistical analysis. The first is a new test of linearity for a stationary time series, that extends the bootstrap methods of Berg et al. (2010) to goodness-of-fit (GoF) statistics specified in Harvill (1999) and Jahan and Harvill (2008). Berg's bootstrap method utilizes the statistics specified in Hinich (1982) in the framework of an autoregressive bootstrap procedure, however we show that by utilizing GoF methods, we can increase the power of the test.

In Chapter three we discuss an alternative way of approaching the Friedman (1989) regularized discriminant method. Regularized discriminant analysis (RDA) is a well-known method of covariance regularization for the multivariate-normal based discriminant function. RDA generalizes the ideas of linear (LDA), quadratic (QDA), and mean-eigenvalue covariance regularization methods into one framework. The original idea and known extensions involve cross-validating in potentially high dimensions, and can be highly computational. We propose using the Kullback-Leibler divergence as an optimization method to estimate a linear combination of class covariance structures, which increases the accuracy of the RDA method, an limits the use of leave one out cross validation.

Topics in Multivariate Covariance Estimation
and Time Series Analysis

by

John D. Beeson, B.S., M.B.A

A Dissertation

Approved by the Department of Statistical Science

Jack D. Tubbs, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

Jane L. Harvill, Ph.D., Chairperson

David J. Kahle, Ph.D.

James D. Stamey, Ph.D.

Joseph D. White, Ph.D.

Dean M. Young, Ph.D.

Accepted by the Graduate School
December 2013

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

To attempt to acknowledge everyone who has supported me and my research throughout my time in graduate school is a difficult task indeed. Baylor University has been a special place for me during my six years here as both a MBA and Ph.D. student.

Two professors in the business school played a significant role in my decision to enter the doctoral program here at Baylor: Dr. James Roberts and Dr. Kris Moore. Dr. James Roberts is more than just a professor I served as a graduate assistant. He is both a mentor and a friend. Without Dr. Moore's encouragement, I do not think I would be here. It was he who enabled me to realize my goal of exploring advanced analytics, and I am very grateful.

I would also like to thank Dr. Bratcher for his belief in me and my potential. I will never forget to always strive for a high level of precision and excellence in my work and in life. Dr. Young's guidance was also essential, especially in the beginning of my program, when I was first attempting to explore my interests.

I had multiple instances of adversity come up during my time here, and without the support of Dr. Stamey and my advisor Dr. Harvill I surely would not have been able to complete my studies. They are both perfect examples of the quality of faculty here at Baylor and it makes me proud to have done my graduate work here.

Last, but surely not least, I consider myself very lucky to have such a supporting wife and family. They truly are a blessing, and a major factor in my success.

DEDICATION

To my parents, David, Connie, and Vicki Beeson

and

To my wife Stephanie Beeson

CHAPTER ONE

Introduction

1.1 Bootstrap-Based Goodness-of-Fit Test for Linearity of a Stationary Time Series

Time series are prevalent in many areas of investigation. In biology, time series data are collected in the observance of predator-prey interactions, or population studies. Financial data is inherently time-dependent – the realization of the closing prices of a stock being a common place example of a time series. Meteorological data, such as daily high temperatures, are time series. Speech recognition technology uses time series analysis methods. EEG records are a time series, as are seismographic records. One would be hard-pressed to find an area of investigation that does not record data across time.

Two primary objectives in time series analysis are modeling and forecasting. To successfully model a series – and thus forecast well – an understanding of the underlying process for model determination is essential. However, it is usually the case that partial or little such knowledge exists. Consequently, data-based methods for making such determinations are extremely useful. Most approaches for doing so rely on the basic presumption that there are two extremely broad classes of models: the class of linear models and the class of nonlinear models. The theory of linear models is well-established and rich in history. An excellent, classic monograph is that of Priestley (1981), or more recently Schumway and Stoffer (2011). Linear models are simple to understand. Thus, if a linear model is appropriate for modeling the data, then one should be used.

However, it is often the case that linear models fail to adequately describe the observed behavior. Tong (1990) is a well-known source for presenting many such

series and dynamic systems. Over the course of recent years, great strides have been made in parametric and nonparametric methods for analyzing nonlinear time series. Some classic nonlinear models include the bilinear model (Subba Rao and Gabr, 1980), the exponential autoregressive model of Haggan and Ozaki (1980), or the self-exciting threshold autoregressive model of Tong (1983). If a class of nonlinear models is unknown, then a nonparametric model might be used, as in Chen and Liu (1993), Cai et al. (2000), or Harvill and Ray (2005).

One set of distinguishing probabilistic properties of nonlinear models are the existence of non-constant higher-order moments, and the corresponding higher-order spectrum. More specifically, if a process is a Gaussian linear process, then all of the information is contained in the second-order moments and the spectral density function. Higher-order moments are zero, or are constant, and contain no additional information about the series. On the other hand, it is rarely the case that the higher-order moments of a nonlinear process are zero, or constant. Thus for a given time series, methods for determining whether the underlying process is linear or nonlinear can be based on higher-order moments, or correspondingly the higher-order spectra. Hinich (1982) devised a two stage test for the Gaussianity and linearity of a stationary time series based on properties of the bispectral density function. The first stage tests the Gaussianity of the series, and the second stage tests time series linearity, but for non-Gaussian errors. The linearity test statistic is the interquartile range (IQR) of the square modulus of the estimated normalized bispectrum. *P*-values are computed using the asymptotic normality of the IQR.

Although Hinich's test is widely used, it suffers from two issues related to the power of the test. First of all it has low power in detecting many forms of nonlinearity. And secondly, the power of Hinich's test is dependent upon the choice of smoothing parameters used in estimating the normalized bispectrum Chan and Tong (1986). Moreover, the using the normal distribution to compute p-values for the

normality stage is suspect, as shown in Harvill and Newton (1995). To address this, Harvill (1999) first suggested using a goodness-of-fit (GoF) statistic in the linearity stage of the test. In that paper, Harvill illustrated via simulation that the GoF approach had the potential to outperform the IQR test. Jahan and Harvill (2008) developed the approach in Harvill (1999), and illustrated for a most of nonlinear forms that the GoF approach outperformed Hinich's approach, especially for series of small to moderate lengths. However, like Hinich's approach, the power of the GoF approach is dependent upon the choice of smoothing parameter.

Most recently, Berg et al. (2010) developed a bootstrap approach applied to Hinich's test, and used flattop estimator for the bispectrum. The coupling of these ideas addressed the problems with the power of Hinich's test. In Chapter Two, we extend Berg's bootstrap algorithm to use the GoF-based statistics of Jahan and Harvill (2008).

1.2 Regularized Discriminant Analysis

Regularized discriminant analysis (RDA), first proposed by Friedman (1989), is a well-known method of covariance estimation for the multivariate-normal based discriminant function. RDA attempts to generalize the ideas of linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and mean-eigenvalue covariance methods into one framework. This is achieved using a two dimensional grid consisting of the weighted average of the LDA and QDA covariance estimators, with weight parameter λ, and then averaging the mean-eigenvalue estimate, with weight parameter γ. Most extensions of this idea involve increasing the parameter space of pooling/shrinking weight-variables to allow increased flexibility of the final model. However, this flexibility comes at the cost of a large space of plausible values after cross validations are completed. In problems where the sample size is approximately equal to the parameter space, these new methods often yield poor results

and drastically increase the computation time. We propose using the Kullback-Leibler divergence as an optimization method to estimate a linear combination of class covariance structures that are similar enough to where pooling them increases the precision of the estimator for the population covariance matrix.

There are various interpretations of the KL divergence. In the engineering literature, KL divergence is a measure of the information lost using $\mathcal{N}_2$ to estimate $\mathcal{N}_1$. The KL divergence has also been interpreted by Eguchi and Copas (2006) as an optimal property of likelihood ratios in connection to the Neyman-Pearson lemma. The KL divergence can also be generalized as a special case of the Bregman divergence

$$D_F(\lambda_1, \lambda_2) = F(\lambda_1) - F(\lambda_2) - \langle \nabla F(\lambda_2), \lambda_1 - \lambda_2 \rangle, \tag{1.1}$$

where $F : \Omega \to \mathbb{R}$ is a continuously-differentiable real-valued and strictly convex function defined on a convex closed set Ω, $\langle \cdot \rangle$ is the inner-product, and (1.1) is the first order Taylor series expansion of F evaluated around point p at point q. The KL divergence uses entropy for the function F. For further details see Bregman (1967). Vemuri et al. (2011) has used the KL divergence as a clustering criterion between two multivariate normal distributions.

In Chapter Three we will show that using the KL divergence as an optimization criterion increases the accuracy of the RDA method and alleviates the high dimensional cross-validation present in more recent adaptations.

CHAPTER TWO

Test for Linearity of a Stationary Time Series Using Goodness-Of-Fit Statistics and the Autoregressive Bootstrap

2.1 Introduction

Understanding the properties of a time series is an important step to any temporal analysis. Linear models are common place in practice because they are often capable of offering parsimonious solutions with reasonable accuracy. Consequently methods for analyzing a time series based on linear model theory are offered in a wide array of statistical packages, and the literature guiding the use of linear models is vast. Although linear models are very effective when they are appropriate, many temporal processes have more complicated generating processes that cannot be well explained using a linear model. Tong (1993) provides several limitations of linear phenomena: (1) a linear model cannot allow for stable periodic solutions independent of the starting value; (2) the Gaussian assumption for the errors are ill-suited for asymmetric random processes; (3) linear models cannot easily explain irregular bursts in amplitude; (4) higher-order moment information is left unexplained; and (5) linear models are insufficient for modeling time irreversible data, which is common in thermodynamics.

There are two broad sets of approaches for testing time series linearity. The first set consists of methods based in the time domain. As discussed in Jahan and Harvill (2008), most time domain approaches make use of a specific nonlinear alternative. These methods have high power when the alternative is true, but the performance of the test suffers if the generating process is not as specified in the alternative. Cai et al. (2000) developed a non-parametric bootstrap-based test of linearity for a univariate time series using functional coefficient autoregressive

(FCAR) models. This FCAR method was later extended to a vector time series by Harvill and Ray (2006). The FCAR method compared favorably to the parametric tests, but was somewhat sensitive to the correct selection of the functional variable.

The other set of approaches for testing time series linearity are based in the frequency domain. Generally speaking, tests in the frequency domain have two stages. The first stage is a test for Gaussianity. If the null is rejected, the test proceeds to the second stage, which is a test of time series linearity, but with non-symmetric errors. The use of the bispectrum for testing Gaussianity and linearity was originally proposed by Subba Rao and Gabr (1980). Their test statistic is a complex analogue of the Hotelling T^2 statistic, and is based on a finite sample estimator of standard error of the bispectrum. Later Hinich (1982) modified the approach by Subba Rao and Gabr (1980) by using an asymptotic standard error, and the standardized interquartile range (IQR) as a test statistic in the second stage of the test. Computation of p-values for the Hinich IQR statistic were refined using a saddlepoint approximation by (Harvill and Newton, 1995). Goodness-of-fit (GoF) tests were proposed by Harvill (1999) and Jahan and Harvill (2008). Bootstrap tests based on Hinich's IQR approach have been proposed by Hinich et al. (2005), and Berg et al. (2010). The fundamental contribution in this chapter is the extension of the work in Berg et al. (2010) to a goodness of fit (GoF) approach. Specifically, in place of the IQR used in Berg's method, we use the empirical cumulative distribution function (EDF) GoF statistics to test Gaussianity and linearity properties of a stationary time series.

2.2 The Bispectral Density Function

Let $\{X_t : t \in \mathbb{Z}\}$, where $\mathbb{Z}$ is the set of integers, be a zero mean third-order stationary process. Then the autocovariance and third-order moment functions are defined as

$$\gamma_u = E[X_t X_{t+u}] \quad \text{and} \quad \gamma_{u,v} = E[X_t X_{t+u} X_{t+v}],$$

respectively. Then if $\{X_t\}$ is second-order stationary and $\sum_{u=-\infty}^{\infty} |\gamma_u| < \infty$, then the spectral density function is defined as

$$I(\omega) = \sum_{u=-\infty}^{\infty} \gamma_u e^{-2\pi i u \omega}, \quad \text{for } \omega \in [0, 1]. \tag{2.1}$$

Because of symmetries, we can restrict our consideration of $I(\omega)$ to $\omega \in [0, 0.5]$. If $\{X_t\}$ is sixth-order stationary and the third-moment function is absolutely summable, then the bispectral density function is defined as

$$I(\omega_1, \omega_2) = \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} \gamma_{u,v} e^{-2\pi i (u\omega_1 + v\omega_2)} \quad \text{for } (\omega_1, \omega_2) \in [0, 1] \times [0, 1]. \tag{2.2}$$

The principle domain of the bispectrum is the region defined by $\mathcal{D} = \{(\omega_1, \omega_2) : 0 \leq \omega_2 \leq \omega_1 \leq 1/2, \omega_1 \leq (1 - \omega_2)/2\}$. The normalized bispectral density is defined as

$$Z(\omega_1, \omega_2) = \frac{|I(\omega_1, \omega_2)|^2}{I(\omega_1) I(\omega_2) I(\omega_1 + \omega_2)}, \quad \text{for } (\omega_1, \omega_2) \in [0, 1] \times [0, 1]. \tag{2.3}$$

The principle domain of $Z(\omega_1, \omega_2)$ is $\mathcal{D}$.

If $\{X_t\}$ is linear, then it admits the representation

$$X_t = \sum_{j=0}^{\infty} \beta_j \epsilon_{t-j}, \tag{2.4}$$

where $\{\epsilon_t\}$ is an independent and identically distributed (iid) sequence with finite variance σ_ϵ^2. Under linearity, the spectral and bispectral density functions defined in (2.1) and (2.2) reduce to

$$I(\omega) = \sigma_\epsilon^2 |H(\omega)|^2 \quad \text{and} \tag{2.5}$$

$$I(\omega_1, \omega_2) = \mu_3 H(\omega_1) H(\omega_2) H^*(\omega_1 + \omega_2), \tag{2.6}$$

where $H(\omega) = \sum_{j=0}^{\infty} \beta_j e^{-2\pi i j \omega}$ is the linear transfer function, μ_3 is the third moment of $\{\epsilon_t\}$, and the asterisk (*) denotes complex conjugate. Therefore if $\{X_t\}$ is linear the normalized bispectrum in (2.3) simplifies to

$$Z(\omega_1, \omega_2) = \frac{\mu_3}{\sigma_\epsilon^6} \quad \text{for all } (\omega_1, \omega_2) \in \mathcal{D}. \tag{2.7}$$

The tests developed by Subba Rao and Gabr (1980), Hinich (1982), Harvill (1999), and Berg et al. (2010) all make use of the constancy of $Z(\omega_1, \omega_2)$ under linearity as expressed in (2.7). As previously mentioned, these tests have two stages. The first stage makes use of the property that, if the series is a Gaussian linear series, the bispectrum will be identically zero. The second stage tests linear but non-Gaussian errors, where the bispectrum is constant but non-zero.

The differences of these methods lie in how estimators of $Z(\omega_1, \omega_2)$ are used to construct the test statistic in the linearity stage. Subba Rao and Gabr (1980) use the estimates of $Z(\omega_1, \omega_2)$ to generate a data matrix, and make use of a complex analog of the Hotelling T^2 statistic to compute p-values. The Hinich (1982) test uses the interquartile range (IQR) of the estimates of $Z(\omega_1, \omega_2)$, computing p-values using the asymptotic normality of the IQR with an asymptotic approximation of the standard errors of the estimators of $Z(\omega_1, \omega_2)$. Harvill (1999) first proposed a GoF approach and illustrates via simulation that GoF tests have the potential to be more powerful. Jahan and Harvill (2008) used a robust transformation of estimators of $Z(\omega_1, \omega_2)$ to normality and applied GoF statistics to the transformed estimates, which proved to have higher power when compared to Hinich's test. Berg et al. (2010) used a modified version of Hinich's test that estimated the sampling distribution of the interquartile range using an autoregressive bootstrap.

2.3 Goodness-of-Fit Tests

Let $\mathbf{Y} = \{Y_1, \ldots, Y_N\}$ denote a sequence of N independent and identically distributed (iid) random variables. When testing whether $\mathbf{Y}$ are distributed according

to a proposed distribution $F(\mathbf{y}|\theta)$, we use a class of quadratic discrepancy measures known as empirical distribution function (EDF) statistics. This class of measures has the form

$$Q\left(\mathbf{y}|F\right) = N\int_{-\infty}^{\infty}\left[F_N(\mathbf{y}) - F(\mathbf{y})\right]^2\psi(\mathbf{x})dF(\mathbf{y}), \tag{2.8}$$

where $\psi(\mathbf{y})$ is a weight function and $F_N(\mathbf{y})$ is the empirical distribution function defined by

$$F_N(y) = \frac{\text{Number of } Y_i \leq y}{N}, \quad -\infty < y < \infty.$$

Different expressions for the weight function $\psi(\mathbf{y})$ yield different distance measures. For instance, when $\psi(\mathbf{y}) = 1$, the discrepancy measure Q is called the Cramér-von Mises (CvM) statistic. If $\psi(\mathbf{y}) = [\{F(\mathbf{y})\}\{1 - F(\mathbf{y})\}]^{-1}$ then Q is known as the Anderson Darling (AD) statistic. While the CvM statistic is an average of the distance of the EDF from the expected cumulative distribution function (CDF) under H_0, the AD statistic penalizes outlying observations, and in general has the higher power of the two. For a detailed look at these statistics and others see D'Agostino and Stephens (1986).

Figure 2.1 below contains a plot of $\psi(\mathbf{y})$ of the AD statistic for three distributions: the standard normal, exponential with mean two, and the $\chi^2_2(1)$. From the plots in Figure 2.1, the effect of the selection of F on the weight function of the AD statistic can be easily seen. For all three distributions, $\psi(\cdot)$ places greater weight on values in the tail(s); that is, on values with low probability. Consequently, the presence of such values in the sample results in a greater increase of the AD discrepancy measure compared to the CvM statistic, which places equal weight on all observations. Skewed distributions such as the exponential and noncentral chi-square have asymmetric weight functions, and the rate of increase of the weight of the function is related to the skewness of F. For right-skewed (left-skewed) distributions, since the median is always less than (greater than) the mean, less weight is placed on values

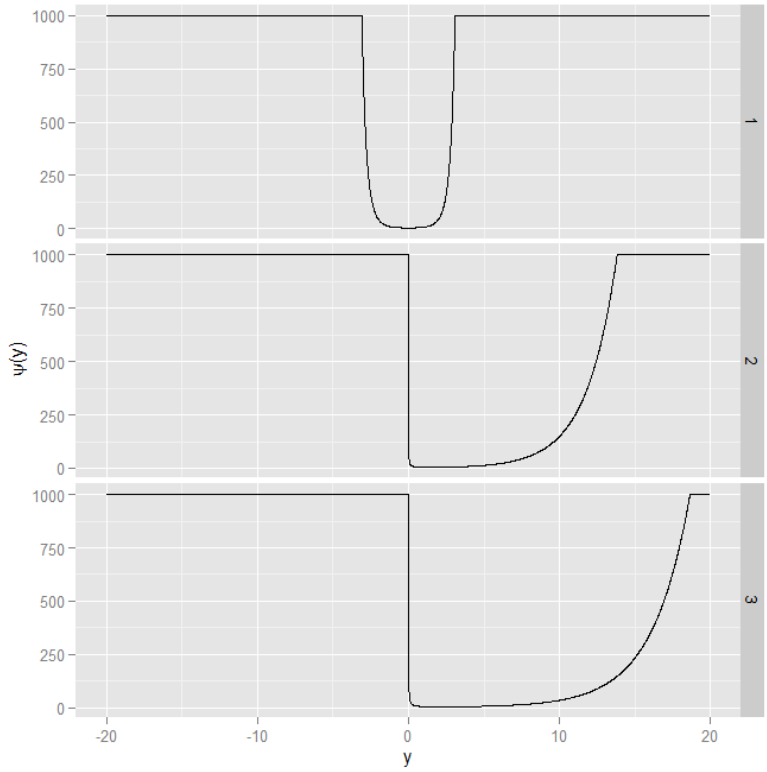

Figure 2.1: $\psi(\mathbf{y})$ for the standard normal (top), exponential with mean two (center), and $\chi^2_2(1)$ (bottom) distributions.

closer to the median (more weight is placed in the tail). This property of the AD weight function makes the AD test statistic somewhat median invariant.

Sample versions of the CvM and AD statistics are based on the Probability Integral Transform. Let $y_{(i)}$ denote the i-th order statistic of the sample, and $G_i = F(y_{(i)})$. Then

$$C = \frac{1}{12N} + \sum_{i=1}^{N} \left\{ G_i - \frac{(2i-1)}{2N} \right\}^2 \tag{2.9}$$

$$A = -N - \frac{1}{N} \sum_{i=1}^{N} \{2i-1\} \{\log G_i + \log(1 - G_{N+1-i})\} . \tag{2.10}$$

2.4 Estimating the Bispectrum

Let $X_1, \ldots, X_n$ be a realization from a zero mean, sixth-order stationary process $\{X_t\}$. Then the estimators of the autocovariance and third moment functions are

$$\hat{\gamma}_v = \frac{1}{n}\sum_{t=1}^{n-|v|} X_t X_{t+v} \quad \text{and} \quad \hat{\gamma}_{u,v} = \frac{1}{n}\sum_{t=1}^{n-s} X_t X_{t+u} X_{t+v}, \tag{2.11}$$

where $s = \max\{0, u, v\}$. Define the natural frequencies $\omega_j = (j-1)/n$, for $j = 1, \ldots, [n/2]+1$, where $[\cdot]$ is the greatest integer value. Then an estimator of the spectral density function is

$$\tilde{I}(\omega_j) = \sum_{u=-M_1}^{M_1} \hat{\gamma}_u e^{-2\pi i u \omega_j}, \tag{2.12}$$

where the truncation point M_1 is chosen such that $M_1 \to \infty$, $n \to \infty$, $M_1/n \to 0$.

To estimate the bispectrum, construct the lattice $\mathcal{L}$ of natural frequency pairs in $\mathcal{D}$ where

$$\mathcal{L} = \left\{ \left(\frac{(2j-1)M_2}{2n}, \frac{(2k-1)M_2}{2n} \right) : j = 1, \ldots, k;\ k < \frac{n}{2M_2} - \frac{k}{2} + \frac{3}{4} \right\},$$

and the truncation point M_2 is chosen so that $M_2 \to \infty$, $n \to \infty$ and $M_2/n \to \infty$. The estimator of the bispectrum in (2.2) is then given by

$$\tilde{I}(\omega_j, \omega_k) = \sum_{u=-M_2}^{M_2} \sum_{v=-M_2}^{M_2} \hat{\gamma}_{u,v} e^{-2\pi i (u\omega_j + v\omega_k)} \ \text{ for } \ (\omega_j, \omega_k) \in \mathcal{L}. \tag{2.13}$$

The estimators in equations (2.12) and (2.13) are not consistent estimators of the spectral density and bispectral density. The kernel approach that yields consistent estimators of the spectral density was developed in a series of seminal papers by Parzen (Parzen, 1957, 1961a,b). Then consistent kernel density estimators for higher-order spectra were developed in Brillinger (1965); Rosenblatt and Ness (1965); Van Ness (1966); Brillinger and Rosenblatt (1967); Brillinger (1969). Summarizing, let $\Lambda(\tau)$ be a one-dimensional symmetric lag window such that $\Lambda(0) = 1$,

and $\Lambda(\tau_1, \tau_2)$ a two-dimensional lag window satisfying

$$\Lambda(\tau_1, \tau_2) = \Lambda(\tau_2, \tau_1) = \Lambda(-\tau_1, \tau_2 - \tau_1) = \Lambda(\tau_1 - \tau_2, -\tau_2) = \Lambda(\tau_1)\Lambda(\tau_2)\Lambda(\tau_1 - \tau_2).$$

Then consistent kernel estimators of the spectral and bispectral densities are

$$\hat{I}(\omega_j) = \sum_{u=-M_1}^{M_1} \Lambda\left(\frac{u}{M_1}\right) \hat{\gamma}_u e^{-2\pi i u \omega_j} \quad \text{for } \omega_j = \frac{j-1}{n},\ j = 1, 2, \ldots \left[\frac{n}{2}\right] + 1, \tag{2.14}$$

$$\hat{I}(\omega_j, \omega_k) = \sum_{u=-M_2}^{M_2} \sum_{v=-M_2}^{M_2} \Lambda\left(\frac{u}{M_2}, \frac{v}{M_2}\right) \hat{\gamma}_{u,v} e^{-2\pi i (u\omega_j + v\omega_k)} \quad \text{for } (\omega_j, \omega_k) \in \mathcal{L}. \tag{2.15}$$

It is common to let $M_1 = M_2 = M$, with $M = n^c,\ 1/2 \leq c \leq 1$, where c controls the trade-off between bias (larger values of c) and variance (smaller values of c). The normalized bispectrum in (2.3) is estimated by

$$\hat{Z}(\omega_j, \omega_k) = \frac{\left|\hat{I}(\omega_j, \omega_k)\right|^2}{(M^2/n)\hat{I}(\omega_j)\hat{I}(\omega_k)\hat{I}^*(\omega_j + \omega_k)}, \quad \text{for } (\omega_j, \omega_k) \in \mathcal{L}. \tag{2.16}$$

Hinich (1982) estimates the spectral and bispectral density using a rectangular window. For a fixed (j, k), a square of dimension M is used to average the $\tilde{I}$, omitting from the average any points with a square that falls outside of $\mathcal{D}$.

2.5 Asymptotic Properties of Spectral Estimators

As was shown in Van Ness (1966), the bispectral estimator in (2.15) is approximately complex normal. If $\{X_t\}$ satisfies the regularity conditions of Theorem 1 of Van Ness (1966), then for fixed j and k,

$$n^{1/2} M \left\{\hat{I}(\omega_j, \omega_k) - \mathrm{E}\left[\hat{I}(\omega_k, \omega_k)\right]\right\} \tag{2.17}$$

converges in distribution to a complex normal variable $V + IW$, say, where V and W have zero mean, are jointly normal, are independent. The variances of V and W are as follows.

Define

$$h_1 = \left[\int_{-\infty}^{\infty} \Lambda(0,v)\,dv\right]^2 \quad \text{and} \quad h_2 = \frac{1}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Lambda^2(v_1,v_2)\,dv_1\,dv_2,$$

and

$$\delta(x) = \begin{cases} 1, & \text{if } x = 0, \\ 0, & \text{if } x \neq 0. \end{cases}$$

If (ω_j, ω_k) lies inside $\mathcal{D}$ and not on the boundaries, then

$$\sigma_V^2 = \sigma_W^2 = h_2 I(\omega_j) I(\omega_k) I(\omega_j + \omega_k).$$

If the boundaries are included, then the variances are

$$\begin{aligned} \sigma_V^2 &= h_1 I(\omega_j) I(\omega_k) I(\omega_j+\omega_k)\left\{8\delta(\omega_j) + \delta(\omega_k)\right\} + A + B, \\ \sigma_W^2 &= A - B, \end{aligned}$$

where

$$\begin{aligned} A &= h_2 I(\omega_j) I(\omega_k) I(\omega_j+\omega_k)\left[\left\{1+\delta(\omega_j-\omega_k)\right\}\left\{1+\delta(\omega_j+2\omega_k-1)\right.\right. \\ &\qquad \left.\left. +\delta(2\omega_j+\omega_k-1)\right\} + 4\delta(\omega_1)\right] \\ B &= h_2 I(\omega_j) I(\omega_k) I(\omega_j+\omega_k)\left[5\delta(\omega_j) + \delta(\omega_k)\left\{1+\delta(\omega_j-0.5)\right\}\right]. \end{aligned}$$

For appropriate choice of M, the estimators $\hat{I}(\omega_j, \omega_k)$ are asymptotically independent. (See Theorem 4 of Brillinger and Rosenblatt (1967).) Thus, if the estimators of $I(\omega_1, \omega_2)$ are restricted to those having frequencies that lie in the principle domain, and not on the boundaries, then $\hat{Z}(\omega_j, \omega_k)$ is asymptotically non-central χ^2 with two degrees of freedom, and noncentrality parameter

$$\lambda_{j,k} \propto 2n^{-(1-4c)} Z(\omega_j, \omega_k),$$

where the proportionality depends in a nontrivial way upon the choice of M. Hinich (1982) explains how to compute the constant of proportionality for the rectangular window, when all points in each average are given equal weight.

Under linearity, the noncentrality parameter $\lambda_{j,k}$ is constant. If there are n_l bispectral estimates, then under linearity, they can be considered approximately independent noncentral $\chi^2_\nu(\lambda)$, where $\nu = 2$ represents the degrees of freedom. An estimator of the (constant) noncentrality parameter due to Saxena and Alam (1982) is

$$\hat{\lambda} = \max\left\{0, \bar{Z} - \nu\right\}, \quad \text{where} \quad \bar{Z} = \frac{1}{n_l}\sum_{i=1}^{n_l} Z_i. \tag{2.18}$$

2.6 Frequency Domain Tests

Generally speaking, tests in the frequency domain have two stages. The first stage is a test for Gaussianity. If the null is rejected, the test proceeds to the second stage, which is a test of time series linearity, but with non-symmetric errors. The use of the bispectrum for testing Gaussianity and linearity was originally proposed by Subba Rao and Gabr (1980). Their test statistic is a complex analogue of the Hotelling T^2 statistic, and is based on a finite sample estimator of standard error of the bispectrum. Later Hinich (1982) modified the approach by Subba Rao and Gabr (1980) by using an asymptotic standard error, and the standardized interquartile range (IQR) as a test statistic in the second stage of the test. Computation of p-values for the Hinich IQR statistic were refined using a saddlepoint approximation by (Harvill and Newton, 1995). Goodness-of-fit (GoF) tests were proposed by Harvill (1999) and Jahan and Harvill (2008). Bootstrap tests based on Hinich's IQR approach have been proposed by Hinich et al. (2005), and Berg et al. (2010). The fundamental contribution in this chapter is the extension of the work in Berg et al. (2010) to a GoF approach. Specifically, in place of the IQR used in Berg's method, we use the EDF GoF statistics to test Gaussianity and linearity properties of a stationary time series. Both Hinich (1982) and Jahan and Harvill (2008) made use of this fact to construct their respective test using (2.16), which has an

asymptotic noncentral chi-square distribution, where as explained in Section 2.5, the noncentrality parameter is a function of the bispectrum.

2.6.1 Hinich (1982) Bispectral-Based Test

Hinich (1982) proposed using the asymptotic properties of the bispectrum to construct his test of Gaussianity and of linearity with non-Gaussian errors. The test is performed in two stages. The first stage tests the Gaussianity of the time series. Under Gaussianity, the bispectrum $Z(\omega_1, \omega_2)$ is zero for all (ω_1, ω_2) in the domain of $\mathcal{D}$.

Let n_l denote the number of frequency pairs in $\mathcal{L}$. The value of n_l is approximately $n^2/(12M^2)$. Reindex the frequency pairs, letting $\boldsymbol{\omega}_i = (\omega_j, \omega_k)$, $i = 1, 2, \ldots, n_l$. Then the test statistic for Gaussianity is

$$T_G = \sum_{i=1}^{n_l} \hat{Z}(\boldsymbol{\omega}_i) \dot{\sim} \chi^2_{2n_l}. \tag{2.19}$$

The p-value is computed using the upper-tail probability of a central χ^2 with $2n_l$ degrees of freedom.

The second stage of Hinich's procedure tests the linearity of the series, but with non-Gaussian errors. Under the linear hypothesis, the sampling distribution of the bispectrum is asymptotically a noncentral chi-square $\chi^2_{2n_l}(\lambda)$, with the noncentrality parameter λ, estimated by (2.18). The interquartile range (IQR) of the bispectral estimates is used as the test statistic. The approximate p-value is based on the $N(\xi_3 - \xi_1, \sigma^2_{\xi_3-\xi_1})$, where ξ_1 and ξ_3 are the first and third quartiles of the noncentral $\chi^2_{2n_l}$ distribution, $f_{\nu,\lambda}(\cdot)$ denotes the probability density function of a noncentral χ^2 distribution with ν degrees of freedom and noncentrality parameter λ, and

$$\sigma^2_{\xi_3-\xi_1} = \frac{1}{16n_l}\left[\frac{3}{f^2_{2n_l,\lambda}(\xi_1)} - \frac{2}{f^2_{2n_l,\lambda}(\xi_1) f^2_{2n_l,\lambda}(\xi_3)} + \frac{3}{f^2_{2n_l,\lambda}(\xi_3)}\right].$$

Hinich's test has been considered the standard bispectral-based test for the Gaussianity and linearity of a stationary time series. However, it has been shown that

the second stage has a number of issues. Three of the main problems with Hinich's test are that (1) it suffers from an inflated Type I error rate as shown in Jahan and Harvill (2008) and Berg et al. (2010) (2) the power of the test is low (again, see Jahan and Harvill (2008)), and (3) the power of the test is highly dependent upon the choice of smoothing parameters (Chang and Tong, 1986).

2.6.2 Goodness-of-Fit Test for Gaussianity and Linearity

Jahan and Harvill (2008) used GoF theory to construct a two stage test of the Gaussianity and linearity of a stationary time series. As with Hinich (1982), the first stage tests Gaussianity, and the second stage tests linearity with non-Gaussian errors. In stage one, the Jahan and Harvill test uses a GoF approach to compare the estimated normalized bispectrum to an exponential(2). The second stage is designed to test the fit of the $\hat{Z}$ to the $\chi^2_2(\lambda)$ distribution. Since the GoF test statistics and corresponding critical values are functions of the unknown λ, Jahan and Harvill use a robust transformation due to Abdel-Aty (1954) of the estimated normalized bispectrum to standard normal. The transformed bispectral estimates are compared to a standard normal via the GoF statistics in (2.9) or (2.10). The algorithm is

Stage 1: Testing Gaussianity.

(1) Compute the normalized bispectral estimates as given in (2.16).

(2) Apply GoF test of exponential(2) to normalized bispectral estimates from step 1.

(3) If the null is not rejected, then stop. Otherwise proceed to Stage 2.

Stage 2: Testing linearity with non-Gaussian errors.

(1) Apply transformation of Abdel-Aty (1954) to the normalized bispectral estimates from step 1 of Stage 1.

(2) Compute GoF statistic for standard normality.

(3) Compare GoF statistic to appropriate upper-tail critical value.

Jahan and Harvill (2008) conducted a numerical investigation of the size and power of Stage 1 of the test, and the power of the second stage of the test. However, there was no investigation of the size of the second stage of the test. Consequently, we ran a small simulation study to fill in that missing information. Four series lengths of $n = 100, 250, 500,$ and 1000, were considered for each of three models. The first model (AR) was an AR(2) given by

$$X_t - 0.4X_{t-1} + 0.3X_{t-2} = \epsilon_t, \quad t = 1, 2, \ldots, n,$$

where ϵ_t is standard normal white noise. The second model was asymmetric zero-mean white noise generated from a central χ^2_2, and centered around 2 (iid Chisq); that is, for $t = 1, 2, \ldots, n$,

$$X_t = \varepsilon_t - 2, \quad \text{where } \varepsilon_t \sim \chi^2_2(0).$$

To investigate the power, the third model, was a bilinear model given by

$$X_t - 0.4X_{t-1} + 0.3X_{t-2} = 0.8\epsilon_{t-1} + 0.5X_{t-1}\epsilon_{t-1} + \epsilon_t, \quad t = 1, 2, \ldots, n,$$

with ϵ_t being standard normal white noise. For each model and each value of n, 5,000 replications were generated.

Table 2.1 contains empirical rejection rates for a level 0.05 linearity tests. Rows without an asterisk contain empirical rejection rates of the Jahan-Harvill test, and rows with an asterisk contain empirical rejection rates for a modified version of the test, described in Section 2.6.3. Note that for the non-modified version the empirical sizes of the test are slightly greater than the nominal 0.05, and increase as the series lengthens. Consequently, the power study for the original versions of the tests are brought into question, given that the Type I error rate is so large. The behavior is more pronounced for the iid Chisq case. On the other hand, the modified test has much improved empirical Type I error rates. Not surprisingly, the power of the

modified test is less than the original tests. But since the original tests' powers are now suspect, then the modified test must be the preferred test.

Table 2.1: Empirical Type I error rates and powers based on 5,000 replications of each model for conventional tests and the modified test using a Monte Carlo estimate of the noncentrality parameter in a goodness of fit test for time series Gaussianity and linearity.

Model	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
AR: CvM	0.07	0.14	0.32	0.76
AR: CvM*	0.11	0.07	0.06	0.07
AR: AD	0.08	0.17	0.42	0.90
AR: AD*	0.12	0.07	0.07	0.07
iid Chisq: CvM	0.12	0.21	0.48	0.90
iid Chisq: CvM*	0.19	0.24	0.21	0.21
iid Chisq: AD	0.13	0.26	0.61	0.98
iid Chisq: AD*	0.18	0.23	0.21	0.21
Bilinear: CvM	0.11	0.24	0.54	0.92
Bilinear: CvM*	0.15	0.20	0.26	0.36
Bilinear: AD	0.12	0.29	0.66	0.98
Bilinear: AD*	0.14	0.20	0.26	0.36

2.6.3 Modification of the Jahan-Harvill Test

One solution for correcting the inflated Type I error would be estimate λ via simulation. Then, instead of transforming the estimated normalized bispectral points to normality, estimate the GoF critical values of a $\chi^2_2(\hat{\lambda})$ distribution. The algorithm for the modified stage two test is as follows.

Stage 2: Testing linearity with non-Gaussian errors.

(1) Using (2.18), estimate λ based on the n_l values of $\hat{Z}$. Call the estimate $\hat{\lambda}$.

(2) Compute GoF statistic for comparing the original $\hat{Z}$ to the $\chi^2_2(\hat{\lambda})$.

(3) For $\ell = 1, 2, \ldots, m$, generate n_l independent observations from $\chi^2_2(\hat{\lambda})$. Denote one replication of n_l observations by $\mathbf{Z}^*_\ell = \{Z^*_{\ell,1}, \ldots, Z^*_{\ell,n_l}\}$, $\ell = 1, \ldots, m$.

(4) For each $\mathbf{Z}_\ell^*,\ \ell = 1, \ldots, m$,

   (a) estimate λ from $\mathbf{Z}_\ell^*$ using (2.18). Call the estimate $\hat{\lambda}_\ell$, and

   (b) calculate GoF statistic for comparing the observations in $\mathbf{Z}_\ell^*$ to a $\chi_2^2(\hat{\lambda}_\ell)$.

(5) Compare the GoF statistic from Step 2 to the $(1-\alpha)^{th}$ quantile of the empirical GoF statistics computed in Step 4(b). If the GoF statistic exceeds the quantile, reject H_0.

Using this method we can estimate the sampling distribution of the CvM and AD statistics under the hypothesis of a $\chi_2^2(\lambda)$ distribution. These critical values will account for the error in estimating λ and produce lower Type I error rates.

2.6.4 Criticisms of Existing Bispectral-Based Tests

Hinich's test can be viewed as a goodness-of-fit test, as the IQR is a measure of fit between the distance of quantiles compared to the appropriate chi-square distribution function. On the other hand, Jahan and Harvill (2008) showed that by using the EDF via a goodness-of-fit approach, the power of the linear stage of the test can be increased significantly. However, as illustrated in Table 2.1, that power may be suspect, since it appears the Type I error rate of the second stage of the test is largely inflated.

The drawbacks of both methods stem from the way they employ asymptotic theory to calculate the p-values for the test statistics. For small, finite, series lengths, the noncentral chi-square is a poor fit to the bispectral estimates, and the resulting in an inflated Type I error rate, but a low power. Moreover, theory in Brillinger and Rosenblatt (1967) requires the series be strictly stationary for the estimated bispectrum to be approximately independent and asymptotically complex normal. Moreover, the variance of the bispectral estimates differ on the boundary. It is highly

likely that the violation of these properties could lead to any of the problems noted above.

Berg et al. (2010) showed that by using an autoregressive bootstrap to estimate the sampling distribution, Hinich's test is improved significantly. We extend this idea using GoF statistics in place of T_G an T_L.

2.7 Bootstrap Methods

Although adding in a simulation step to the second stage, which computed the GoF critical values with respect to the uncertainty of estimating λ, to the asymptotic GoF test yielded empirical Type I errors closer to the nominal α, the computation complexity increased dramatically. Therefore, if a simulation is necessary to compute critical points, we could use a bootstrap simulation to approximate the sampling distribution of the GoF statistics under the null hypothesis of Gaussianity and linearity respectively.

Bootstrapping a sequence of stationary time series data, involves slightly different methods than those of Efron and Tibshirani (1993), which deal with data that are iid. There are several ways to bootstrap a time series, the two most popular being the block bootstrap proposed by Künsch (1989), and the sieve bootstrap proposed by Bühlmann (1997). For a detailed review of these procedures see Bühlmann (2002).

2.7.1 AR(p) Bootstrap

The autoregressive order p (AR(p)) bootstrap stems from the idea that a zero-mean covariance stationary process $\{X_t\}$ can be written as an infinite order moving average process given by

$$X_t = \beta + \sum_{v=0}^{\infty} \beta_v \epsilon_{t-v}, \tag{2.20}$$

where ϵ_t is a white noise process with finite variance, and $\sum_{v=0}^{\infty} \beta_v^2 < \infty$ (Anderson, 1971). Moreover, if $\{X_t\}$ is invertible, then it can be written as an one-sided infinite-order autoregressive process

$$\sum_{j=0}^{\infty} \phi_j X_{t-j} = \epsilon_t, \tag{2.21}$$

with $\sum_{j=0}^{\infty} \phi_j^2 < \infty$. Moreover, any non-linear covariance stationary process can be arbitrarily well-approximated a large, possibly infinite, order autoregressive model. This is the motivating idea behind using an autoregressive model to estimate the sampling distributions under the Gaussian and linear hypothesis.

2.7.2 Bootstrap Adaptation of Hinich's Test

An adaptation of Hinich's linearity test using an $AR(p)$ bootstrap was proposed by Berg et al. (2010). If $X_1, \ldots, X_n$ is a time series of length n, then the Berg bootstrap algorithm is as follows:

(1) Fit an AR(p) model to $X_1, \ldots, X_n$ and obtain estimated coefficients

$$\left(\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p\right).$$

(2) For $\ell = 1, \ldots, n_b$, generate a series of pseudo-observations, $X_1^*, X_2^*, \ldots, X_n^*$, using

$$X_t^* = \sum_{j=1}^{p} \hat{\phi}_j X_{t-j}^* + u_t^*, \quad t = 1, \ldots, n \tag{2.22}$$

where $X_t^* = 0$ for $t \leq 0$ and $u_t^* \sim F_n$. The form of F_n will depend upon the null hypothesis, and may be data dependent. For example, if H_0 is testing the Gaussianity of the series

$$F_n \sim N(0, \hat{\sigma}_p^2),$$

where

$$\hat{\sigma}_p^2 = \frac{1}{n-p} \sum_{t=p+1}^{n} \left(\hat{u}_t - \bar{u}_n\right)^2.$$

As another example, when the null hypothesis is testing time series linearity with non-Gaussian errors, then F_n is the EDF of the centered residuals, $\hat{u}_t - \bar{u}_n$, where

$$\hat{u}_t = X_t - \sum_{j=1}^{p} \hat{\phi}_j X_{t-j} \quad t = p, p+1, \ldots, n,$$

and

$$\bar{u}_n = \frac{1}{n-p} \sum_{t=p+1}^{n} \hat{u}_t.$$

(3) For each pseudo-series, compute the estimates of the normalized bispectrum for each pseudo-series. For $\ell = 1, \ldots, n_b$, let $\hat{\mathbf{Z}}_\ell^\dagger$ denote the n_l-length vector of these estimates. For each series, compute the statistics

$$T_G^\dagger = \sum_{i=1}^{n_l} \hat{Z}_i^\dagger$$

$$T_L^\dagger = IQR(\hat{\mathbf{Z}}^\dagger)$$

and store them in vectors

$$\mathbf{T}_{bG} = \left\{ T_{G,1}^\dagger, \ldots, T_{G,n_b}^\dagger \right\}$$

$$\mathbf{T}_{bL} = \left\{ T_{L,1}^\dagger, \ldots, T_{L,n_b}^\dagger \right\},$$

where n_b is the number of bootstrap replications.

The empirical distribution functions of the bootstrap replications, denoted

$$F_{nG} = F_n(\mathbf{T}_{bG}) \tag{2.23}$$

$$F_{nL} = F_n(\mathbf{T}_{bL}), \tag{2.24}$$

are used to approximate the Gaussian and linear null sampling distributions respectively. The bootstrap p-values, p_G and p_L say, are found by calculating T_L and T_G from the bispectral estimates of the original data, and computing

$$p_G = 1 - F_{nG}(T_G)$$

$$p_L = 1 - F_{nL}(T_L).$$

2.8 Goodness-of-Fit Bootstrap Test for Gaussianity and Linearity of a Time Series

The test for Gaussianity and linearity of a stationary time series proposed in this section combines theory from Jahan and Harvill (2008) and Berg et al. (2010). The first two steps are identical to Berg's method, so they will not be repeated here. However, like the GoF asymptotic test detailed in Section 2.6.2, estimation of λ greatly affects the performance of the test. The bootstrap algorithm for using the GoF statistics specified in (2.9) and (2.10) to test time series Gaussianity and linearity is as follows.

(1) As outlined in Section 2.7.2, generate a sequence of pseudo-observations via the residuals of a fitted AR(p) model.

(2) For each pseudo-series calculate the normalized bispectrum. Under linearity, each of the normalized bispectral estimates will be approximately independent, and are approximately $\chi^2_{2n_l}(\lambda)$.

(3) To test the Gaussian hypothesis, use $\hat{\mathbf{Z}}^\dagger$ to calculate the Cramér von Mises statistics $C_G^\dagger$ or Anderson-Darling statistics $A_G^\dagger$ from (2.9) and (2.10), using $F \sim$ exponential(2), and store in vectors

$$\mathbf{A}_G = \left\{ A^\dagger_{G,1}, \ldots, A^\dagger_{G,n_b} \right\}$$
$$\mathbf{C}_G = \left\{ C^\dagger_{G,1}, \ldots, C^\dagger_{G,n_b} \right\}.$$

(4) Calculate the p-value in manner analogous to the p-value for Hinich's bootstrap test. If the null hypothesis is rejected, proceed to the next step. If not, then end.

(5) To test the hypothesis of non-Gaussian independent errors, use $\hat{\mathbf{Z}}^\dagger$ to calculate $C_L^\dagger$ and $A_L^\dagger$ from (2.9) and (2.10), using $F \sim \chi^2_2(0)$, and store into

vectors

$$\mathbf{A}_L = \left\{ A^{\dagger}_{L,1}, \ldots, A^{\dagger}_{L,n_b} \right\}$$
$$\mathbf{C}_L = \left\{ C^{\dagger}_{L,1}, \ldots, C^{\dagger}_{L,n_b} \right\}.$$

(6) Calculate the p-value in manner analogous to the p-value for Hinich's bootstrap test.

Berg et al. (2010) showed that the AR(p) algorithm yields a consistent estimator for any almost everywhere continuous function. The EDF statistics in (2.9) and (2.10) meet this criterion. Therefore our GoF algorithm is consistent.

2.8.1 Estimation of λ

As mentioned in the introduction of this section, estimation for λ has a definitive effect on the size and power of the test. In our simulation studies, we considered three methods for estimating the noncentrality parameter in the linearity step of the test.

(1) Estimate a global $\hat{\lambda}$ using (2.18) from the bispectral estimates of the original series. Then for each replication, use $\hat{\lambda}$ for λ.

(2) For each bootstrap replication, use the estimates of the normalized bispectrum (for that replication) to get a bootstrap estimate $\hat{\lambda}_\ell$.

(3) Use a naive constant estimator $\tilde{\lambda}$ for all estimates of λ.

In a simulation study (see Appendix A), method 3 performed best. A possible explanation is that for method (1), $\hat{\lambda}$ would be the best estimator for λ with respect to mean squared error discussed in Saxena and Alam (1982), for the bispectral estimates from the original series. But for each pseudo-series, $\hat{\lambda}$ is not best, since the pseudo-series is not the same data. This leads to method (2), which presupposes that estimating the noncentrality parameter separately for each pseudo-series would yield

a better estimate of λ with respect to a particular bootstrap replication. Therefore, $\hat{\lambda}$ will yield a GoF statistic in the lower tail of the sampling distribution, and the hypothesis of linearity will rarely be rejected. However, for method (2), estimation of λ for each pseudo-series increases the variability, and consequently decreases the overall power of the test. The empirical evidence provided in the simulation study suggests that, in most cases the power improved asymptotically, but was still less than both the Berg bootstrap and the method (3) GoF bootstrap.

The choice of the constant $\tilde{\lambda}$ is a crucial part of our GoF bootstrap method. In our simulations $\tilde{\lambda} = 0$ yielded the best performance. This can be explained by noting that for the null models considered the simulation study on Stage 2 of the linearity tests, most values of the estimated normalized bispectrum were close to zero. This idea can be generalized by utilizing the same AR bootstrap engine to estimated the sampling distribution of λ. If we let $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_{n_b}\}$, then choose $\tilde{\lambda} = \min\{\boldsymbol{\lambda}\}$. Other possibilities are to choose $\tilde{\lambda}$ as some lower percentile of the $\lambda_1, \ldots, \lambda_{n_b}$.

2.9 Simulation

In order to evaluate the performance proposed GoF bootstrap, we examines the size and power of both the Gaussianity test and the linearity test. To examine if the proposed GoF Gaussian bootstrap behaves as a level α test, we considered three null models.

(1) White Noise [WN(1)]:

$$X_t = \epsilon_t$$

(2) Autoregressive [AR(2)]:

$$X_t - 0.4X_{t-1} + 0.3X_{t-2} = \epsilon_t, \text{ and}$$

(3) AR Moving Average [ARMA(2,2)]:

$$X_t - 0.8897X_{t-1} + 0.4858X_{t-2} = \epsilon_t - 0.2279\epsilon_{t-1} + 0.2488\epsilon_{t-2},$$

where $\{\epsilon_t\}$ is Gaussian white noise with unit variance. To examine the size of the linearity test, we use

$$\text{Centered Chisquare}: \quad X_t = \chi^2_2(0) - 2,$$

which should be rejected by the Gaussian test for a large percentage of the replications. On the other hand, the level α test for time series linearity with non-Gaussian errors should reject for approximately $100\alpha\%$ of the replications. To examine the power of the linearity test we use several non-linear models from Jahan and Harvill (2008), and Berg et al. (2010).

(1) Bilinear:

$$X_t = 0.4X_{t-1} - 0.3X_{t-2} + 0.8\epsilon_{t-1} + 0.5X_{t-1}\epsilon_{t-1} + \epsilon_t$$

(2) AR Conditional Heteroscedasticity [ARCH]:

$$X_t = \epsilon_t\sqrt{1.5 + 0.9X^2_{t-1}}$$

(3) Nonlinear Moving Average [NLMA]:

$$X_t = 0.5\epsilon_{t-1} - 0.6\epsilon^2_{t-1} + \epsilon_t$$

(4) Self Exciting Threshold AR (SETAR):

$$X_t = \begin{cases} 1 - 0.5X_{t-1} + \epsilon_t & \text{if } X_{t-1} < 0, \\ -1 - 0.5X_{t-1} + \epsilon_t & \text{if } X_{t-1} \geq 0. \end{cases}$$

(5) Smoothed TAR [STAR]:

$$X_t = 1 - 0.5X_{t-1} - 0.5F\left(X_{t-1}\right) + \epsilon_t,$$

where $F\left(x\right) = 1 + \exp(-x/2)$

(6) Exponential AR [EXPAR]:

$$X_t = \left[0.3 + 100\exp\left(-X_{t-1}^2\right)\right] X_{t-1} + \epsilon_t$$

(7) Nonlinear AR [NLAR]:

$$X_t = -0.25X_{t-1} + 0.2X_{t-2} + 0.15X_{t-2}^2 - 0.1X_{t-2}^2 + \epsilon_t$$

(8) Generalized ARCH [GARCH]:

$$X_t = \sigma_t \epsilon_t,$$

where

$$\sigma_t^2 = 0.015 + 0.112X_{t-1}^2 + 0.492\sigma_{t-1}^2 - 0.034\sigma_{t-2}^2 + 0.420\sigma_{t-3}^2$$

The GARCH model specifically tests the behavior of the linearity and Gaussian test when the stationary assumption is violated.

For each model, 5000 replications were run, which bounded the simulation error by 0.0071. To examine the asymptotic behavior we used series lengths of $n = 100, 250, 500$ and 1000. The level of the test was $\alpha = 0.05$. The results of the simulation are listed in Tables 2.3 and 2.4. The acronyms for the results tables are listed in Table 2.2. For the three linear series, all Type I errors for the first stage of the test were within the simulation error when compared to Berg's Hinich-based bootstrap. The difference between the CvM and AD were negligible. Both of the bootstrap methods performed better than the asymptotic methods with respect to level α test.

Table 2.2. Acronyms for Tables 2.3 and 2.4

GCBG	GoF CvM Bootstrap Gaussian	GCBL	GoF Bootstrap CvM Linear
GABG	GoF AD Bootstrap Gaussian	GABL	GoF Bootstrap AD Linear
HBG	Hinich Bootstrap Gaussian	HBL	Hinich Bootstrap Linear
JCG	Jahan CvM Gaussian	JCG	Jahan CvM Linear
JAG	Jahan AD Gaussian	JAL	Jahan AD Linear
HG	Hinich Gaussian	HL	Hinich Linear

Table 2.3. Test of Gaussian Errors

Model	n	GCBG	GABG	HBG	JCG	JAG	HG
WN(1)	100	0.0522	0.0514	0.0580	0.0888	0.0878	0.0802
WN(1)	250	0.0554	0.0528	0.0512	0.0976	0.0980	0.0794
WN(1)	500	0.0498	0.0514	0.0530	0.0934	0.0946	0.0828
WN(1)	1000	0.0522	0.0524	0.0560	0.0820	0.0830	0.0800
AR(2)	250	0.0562	0.0570	0.0606	0.1470	0.1468	0.1450
AR(2)	500	0.0572	0.0548	0.0578	0.0964	0.0942	0.0906
AR(2)	1000	0.0538	0.0520	0.0602	0.0878	0.0878	0.0842
ARMA(2,2)	100	0.0524	0.0498	0.0592	0.0962	0.0962	0.0912
ARMA(2,2)	250	0.0482	0.0462	0.0518	0.0866	0.0854	0.0704
ARMA(2,2)	500	0.0500	0.0512	0.0518	0.0938	0.0924	0.0852
ARMA(2,2)	1000	0.0500	0.0488	0.0566	0.0780	0.0770	0.0792

For the second stage test of non-Gaussian but linear errors, the GoF bootstrap outperformed the Hinich bootstrap for the non-linear models considered. The Type I error of the second stage of the test yielded slightly higher rates for the GoF bootstrap compared to the Hinich bootstrap, however the difference was small, and improved asymptotically. The GoF bootstrap had the best performance with respect to the ARCH, bilinear, and SETAR models. And had a negligible gain on the other models considered. All methods performed poorly on the EXPAR and STAR models.

2.10 Conclusion

In this work we have extended the idea of using the AR(p) bootstrap to estimate the null sampling distribution used for the bispectral-based two stage linearity

test. We have added the use of EDF GoF statistics in place of the IQR in the second stage of the test. Although the performance of the first stage was not improved using GoF statistics, this is in line with the general idea that the sum of the bispectral estimates is sufficient for estimating the null distribution under the assumption of Gaussian errors.

There is clearly an advantage to using the GoF bootstrap to test the linearity properties of a time series in the second stage of the test. The GoF bootstrap performed better than the Hinich in almost all cases tested, and, in the cases of the ARCH, bilinear, and SETAR, did so by a large margin.

For example, the power of the GoF bootstrap for the bilinear model was greater by 0.083, at $n = 100$, but still had advantages at $n = 1000$. The TAR had similar performance gains. The cases that the GoF bootstrap was outperformed by the Hinich, as in the NLAR model with $n = 100$, the GoF test performed better asymptotically.

The Anderson-Darling statistic, generally performed better than the Cramer-Von-Misés statistic. This is likely due to the median invariance property of the AnD statistic, and thus has less sensitivity to misspecification of the parent distribution F. A natural extension of the GoF method described here would be to examine the effect of different weight functions in (2.10).

In addition, the idea of using GoF statistics in the bootstrap test should be extended to infinite order flat top window functions to estimate the bispectrum as done in Berg and Politis (2009). This was initially examined by the authors in simulation, however the GoF bootstrap test showed a large sensitivity to bandwidth selection. This is due to the fact that the GoF statistics are more sensitive to variation of the data than the IQR is in the Hinich based test.

Table 2.4. Test of Linear Errors

model	n	GCBL	GABL	HBL	JCL	JAL	HL
χ^2	100	0.1092	0.1098	0.0988	0.0002	0.0004	0.1970
χ^2	250	0.0908	0.0910	0.0884	0.0004	0.0010	0.2256
χ^2	500	0.0750	0.0762	0.0768	0.0090	0.0154	0.2286
χ^2	1000	0.0774	0.0782	0.0820	0.0776	0.1708	0.2224
Bilinear	100	0.4082	0.4156	0.3326	0.0016	0.0022	0.2114
Bilinear	250	0.6046	0.6164	0.5498	0.0038	0.0056	0.2246
Bilinear	500	0.8146	0.8256	0.7658	0.0354	0.0554	0.2504
Bilinear	1000	0.9510	0.9538	0.9172	0.1898	0.3144	0.2246
ARCH	100	0.4754	0.4842	0.4170	0.0132	0.0168	0.2056
ARCH	250	0.6354	0.6508	0.6028	0.0542	0.0670	0.2940
ARCH	500	0.7294	0.7490	0.7212	0.1748	0.2296	0.3760
ARCH	1000	0.8102	0.8378	0.8226	0.5376	0.6536	0.4206
NLMA	100	0.1614	0.1646	0.1434	0.0000	0.0000	0.1864
NLMA	250	0.2058	0.2078	0.1762	0.0002	0.0002	0.1848
NLMA	500	0.2340	0.2424	0.2124	0.0016	0.0040	0.1956
NLMA	1000	0.3106	0.3174	0.2688	0.0268	0.0758	0.1826
SETAR	100	0.2642	0.2680	0.1984	0.0000	0.0000	0.2014
SETAR	250	0.4392	0.4496	0.3480	0.0000	0.0002	0.2110
SETAR	500	0.6602	0.6722	0.5598	0.0016	0.0038	0.2222
SETAR	1000	0.9046	0.9154	0.8170	0.0194	0.0616	0.1994
STAR	100	0.0428	0.0420	0.0484	0.0634	0.684	0.1284
STAR	250	0.0562	0.0550	0.0492	0.1268	0.1538	0.1440
STAR	500	0.0576	0.0564	0.0558	0.3244	0.4286	0.1468
STAR	1000	0.0564	0.0550	0.0550	0.7404	0.8934	0.1372
EXPAR	100	0.0522	0.0444	0.0640	0.0000	0.0000	0.1950
EXPAR	250	0.0672	0.0646	0.0602	0.0000	0.0000	0.1784
EXPAR	500	0.0556	0.0574	0.0488	0.0012	0.0028	0.1570
EXPAR	1000	0.0248	0.0248	0.0250	0.0222	0.0754	0.1400
NLAR	100	0.0712	0.0714	0.0808	0.0000	0.0000	0.1446
NLAR	250	0.1206	0.1238	0.1048	0.0000	0.0000	0.1656
NLAR	500	0.1310	0.1378	0.1262	0.0004	0.0012	0.1660
NLAR	1000	0.1716	0.1798	0.1704	0.0208	0.0594	0.1674
GARCH	100	0.1488	0.1486	0.1366	0.0000	0.0000	0.1626
GARCH	250	0.5022	0.5146	0.4798	0.0004	0.0006	0.2104
GARCH	500	0.8346	0.8470	0.8202	0.0562	0.0972	0.2562
GARCH	1000	0.9732	0.9760	0.9694	0.4914	0.6192	0.3262

CHAPTER THREE

Regularized Discriminant Analysis: Regularized Covariance Estimation Using a Modified Kullback-Leibler Divergence Criterion

3.1 Discriminant Analysis

In statistical discriminant analysis we wish to accurately assign a new, unlabeled observation $\mathbf{x}_i \in \mathbb{R}^P$, where $\mathbb{R}^P$ is a p-dimensional real valued vector, into one of K known classes by using a training set $\mathcal{L}$ of labeled observations to construct a decision rule $D_0(\mathbf{x})$. Let $\mathcal{L} = \{(\mathbf{x}_i, y_i) : i = 1, \ldots, N\} \in \mathbb{R}^{[N\times(p+1)]}$, where $(\mathbf{x}_i, y_i)$ is a realization from a probability density function. The density function $f_k(\mathbf{x}|C_k)$ is the class conditional probability density function of the kth class, $y_i \in \{C_1, \ldots, C_K\}$ denotes the actual, unique membership of $\mathbf{x}_i$, and n_k is the number of observations in $\mathcal{L}$ from class C_k, so that $N = \sum_{k=1}^{K} n_k$.

The Bayes decision rule $D_0(\mathbf{x})$ that assigns an unlabeled observation $\mathbf{x}$ to one of K distinct classes is chosen with respect to the optimization of a loss function criterion. Notice that if $g_k(C_k)$ is the prior probability of class membership of the kth class then

$$P(C_k|\mathbf{x}) = \frac{f_k(\mathbf{x}|C_k)g_k(C_k)}{\sum_{j=1}^{K} f_k(\mathbf{x}|C_j)g_k(C_j)}. \tag{3.1}$$

is the posterior of C_k given $\mathbf{x}$. Therefore, the risk for classifying the unlabeled observation $\mathbf{x}$ into class C_k is

$$R(\hat{C}|\mathbf{x}) = \frac{\sum_{k=1}^{K} L(C_k, \hat{C}) f_k(\mathbf{x}|C_k)g_k(C_k)}{\sum_{j=1}^{K} f(\mathbf{x}|C_k)g_k(C_k)}, \tag{3.2}$$

where $L(C_k, \hat{C})$ is the loss incurred for misclassifying an observation. Under 0/1 loss,

$$L(C, \hat{C}) = 1 - \delta(C, \hat{C}),$$

where $\delta(C, \hat{C}) = 1$ if the observation was correctly classified, and zero otherwise. We define the Bayes rule classifier as

$$D_0(\mathbf{x}) \equiv \max_{k \in K} \left\{ f_k \left(\mathbf{x} | C_k \right) g_k(C_k) \right\}. \tag{3.3}$$

If the K distributions can be assumed to be approximately multivariate normal, then the class conditional density is modeled as

$$f_k(\mathbf{x}|C_k) = \mathrm{MVN}_p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-p/2} \left|\boldsymbol{\Sigma}_k\right|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right], \tag{3.4}$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^p$ is the p-dimensional mean vector and $\boldsymbol{\Sigma}_k \in \mathbb{R}^+_{p \times p}$ is a positive definite symmetric $p \times p$ matrix. We estimate $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, $k = 1, 2, \ldots, K$, using their maximum likelihood estimators

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} \mathbf{x}_j$$

and

$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} \left(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_k \right)^T \left(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_k \right).$$

One can readily determine two MVN-based discriminant functions defined by the homogeneity or heterogeneity of the class covariance matrices. For the linear discriminant analysis (LDA) classifier we assume $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ for $k = 1 \ldots, K$, so that

$$D_0(\mathbf{x}) = \max_{k \in K} \left\{ \mathrm{MVN}_p \left(\mathbf{x} | \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_{pooled} \right) g_k(C_k) \right\}, \tag{3.5}$$

where

$$\hat{\boldsymbol{\Sigma}}_{pooled} = \frac{1}{N} \sum_{k=1}^{K} n_k \hat{\boldsymbol{\Sigma}}_k. \tag{3.6}$$

If some subset of the K covariance matrices are not equal, then we have the quadratic discriminant analysis (QDA) classifier given by

$$D_0(\mathbf{x}) = \max_{k \in K} \left\{ \mathrm{MVN}_p \left(\mathbf{x} | \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k \right) g_k(C_k) \right\}. \tag{3.7}$$

Estimation of the covariance matrices can be problematic if the parameter space p is large relative to n_k. Let $\mathbf{v}_j$, $j = 1, \ldots, p$ represent the p eigenvectors of $\hat{\mathbf{\Sigma}}_k^{-1}$ and e_j the jth largest eigenvalue of $\hat{\mathbf{\Sigma}}_k^{-1}$ corresponding to $\mathbf{v}_j$. Then the spectral decomposition of $\hat{\mathbf{\Sigma}}_k^{-1}$ is

$$\hat{\mathbf{\Sigma}}_k^{-1} = \sum_{j=1}^{p} \frac{\mathbf{v}_j \mathbf{v}_j'}{e_j}.$$

Examination of the decomposition reveals that small values of e_j inflate the influence of $\mathbf{v}_j$ on the estimator $\hat{\mathbf{\Sigma}}_k^{-1}$. The smallest eigenvalues of $\hat{\mathbf{\Sigma}}_k^{-1}$ are underestimated, having a bias that is inversely proportional to n_k/p. The biased estimates of e_j cause highly variable and unstable estimators of $\hat{\mathbf{\Sigma}}_k^{-1}$ and, thus can cause poor classifier performance. For high-dimensional data, the number of training observations needed to sufficiently estimate the covariance matrices increases exponentially, so some form of covariance regularization is often needed. One common approach to covariance regularization akin to those used in ridge, Bayesian, and semi-parametric regression, is shrinkage regularization of the form

$$\hat{\mathbf{\Sigma}}_k(\gamma) = \hat{\mathbf{\Sigma}}_k + \gamma \mathbf{I}_p, \tag{3.8}$$

where $\mathbf{I}_p \in \mathbb{R}^{p \times p}$ is the $p \times p$ identity matrix, and γ is a positive scalar. Expression (3.8), known as a shrinkage estimator, shrinks $\hat{\mathbf{\Sigma}}_k$ towards the identity matrix, and thus the eigenvalues of $\hat{\mathbf{\Sigma}}_k^{-1}$ away from zero.

3.2 Regularized Discriminant Analysis

The regularized discriminant analysis (RDA) supervised classification method in Friedman (1989) proposes a generalization of the covariance estimation and regularization methods discussed in Section 3.1. Instead of imposing equality assumptions on the $\mathbf{\Sigma}_k$, Friedman proposed first computing a weighted average of the sample class covariance matrices and the pooled sample covariance matrix defined in (3.6). A weight parameter $\lambda \in [0, 1]$ dictates the amount of information shared between $\hat{\mathbf{\Sigma}}_k$

and $\hat{\mathbf{\Sigma}}_{pooled}$. Friedman's regularized covariance matrix estimator $\hat{\mathbf{\Sigma}}_k(\lambda)$ is computed using

$$\begin{aligned} n_k(\lambda) &= (1-\lambda)n_k + \lambda N, \\ \mathbf{S}_k &= n_k \hat{\mathbf{\Sigma}}_k, \\ \mathbf{S} &= \sum_{k=1}^{K} \mathbf{S}_k, \\ \mathbf{S}_k(\lambda) &= (1-\lambda)\mathbf{S}_k + \lambda \mathbf{S}, \\ \hat{\mathbf{\Sigma}}_k(\lambda) &= \frac{\mathbf{S}_k(\lambda)}{n_k(\lambda)}. \end{aligned} \tag{3.9}$$

If we examine MVN classifers of the form

$$D_0(\mathbf{x}) = \max_{k \in K} \left\{ \text{MVN}_p \left(\mathbf{x} | \hat{\boldsymbol{\mu}}_k, \hat{\mathbf{\Sigma}}_k(\lambda) \right) g_k(C_k) \right\}, \tag{3.10}$$

then when $\lambda = 0$, (3.10) reduces to the QDA classifier given in (3.7), and when $\lambda = 1$ (3.10) reduces to the LDA classifier given in (3.5). When $\hat{\mathbf{\Sigma}}_k(\lambda)$ is singular, further regularization is needed. Friedman (1989) proposed using a shrinkage regularization method similar to (3.8) in the form of

$$\hat{\mathbf{\Sigma}}_k(\lambda, \gamma) = (1-\gamma)\hat{\mathbf{\Sigma}}_k(\lambda) + \gamma \frac{\text{tr}\left[\hat{\mathbf{\Sigma}}(\lambda)\right]}{p} \mathbf{I}_p. \tag{3.11}$$

For a given value of λ, the parameter $\gamma \in [0, 1]$ controls shrinkage toward a multiple of the identity matrix. The multiplier $\text{tr}\left[\hat{\mathbf{\Sigma}}(\lambda)\right]/p$ is simply the average eigenvalue of $\hat{\mathbf{\Sigma}}_k(\lambda)$. This shrinkage has the effect of decreasing the larger eigenvalues and increasing the smaller eigenvalues. Thus the RDA classifier can be defined as

$$D_0(\mathbf{x}) = \max_{k \in K} \left\{ \text{MVN}_p \left(\mathbf{x} | \hat{\boldsymbol{\mu}}_k, \hat{\mathbf{\Sigma}}_k(\lambda, \gamma) \right) g_k(C_k) \right\}, \tag{3.12}$$

where $g_k(C_k)$ is the prior probability of class membership of the kth class.

3.3 Estimation of the RDA Classifier

As there are no known closed form estimators for the regularization parameters λ and γ, Friedman (1989) proposed a leave one out (LOO) cross validation procedure

to select the optimal values of λ, γ, denoted $\hat{\lambda}, \hat{\gamma}$, from a model selection lattice of dimensions $L \times G$. Specifically the model selection lattice is constructed using the Cartesian product of $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_L)'$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_G)'$, where $\lambda_l, \gamma_g \in [0, 1]$ for $l = 1, \ldots, L$, and $g = 1, \ldots, G$.

The following definition for the LOO error estimator follows from Hastie et al. (2009). Let $\hat{f}^{(i)}(\mathbf{x}_i)$ denote the trained classifier computed with the ith observation removed. If we consider the 0/1 loss function, then the LOO misclassification error for each pair (λ_l, γ_g) is

$$\widehat{err}(\lambda_l, \gamma_g) = \frac{1}{N} \sum_{i=1}^{N} \delta \left(y_i, f^{(i)}(\mathbf{x}_i) \right). \tag{3.13}$$

The optimal regularization parameters are chosen to be the (λ_l, γ_g) lattice pair with the smallest LOO error rate; that is

$$\left(\hat{\lambda}, \hat{\gamma} \right) = \arg\min_{l,g} \left\{ \widehat{err} \left(\lambda_l, \lambda_g \right) \right\}. \tag{3.14}$$

3.4 Alternative Covariance Matrix Regularization Methods

There have been several alternative covariance regularization methods for supervised classification in the literature that extend Friedman's RDA classifier. Most of the alternative regularization techniques involve increasing the covariance estimation flexibility by allowing for pooling and shrinking parameters (λ_k, γ_k), $k = 1, 2, \ldots, K$, for each class. Covariance estimation flexibility is most desirable when there are varying combinations of similarity between the covariance structures among the K classes.

One covariance regularization method, called mixed leave one out covariance (LOOC-1), was proposed by Kuo and Landgrebe (2002). For their covariance regularization technique the covariance estimator is chosen to be a linear combination

of six different components of $\hat{\mathbf{\Sigma}}_k$ and $\hat{\mathbf{\Sigma}}_{pooled}$ so that

$$\hat{\mathbf{\Sigma}}_k^{LOOC-1} = \alpha_k^{(1)} \frac{\mathrm{tr}[\hat{\mathbf{\Sigma}}_k]}{p} \mathbf{I}_p + \alpha_k^{(2)} \mathrm{diag}[\hat{\mathbf{\Sigma}}_k] + \alpha_k^{(3)} \hat{\mathbf{\Sigma}}_k + \alpha_k^{(4)} \frac{\mathrm{tr}[\hat{\mathbf{\Sigma}}_{pooled}]}{p} \mathbf{I}_p + \alpha_k^{(5)} \mathrm{diag}[\hat{\mathbf{\Sigma}}_{pooled}] + \alpha_k^{(6)} \hat{\mathbf{\Sigma}}_{pooled}, \quad k = 1, 2, \ldots, K. \tag{3.15}$$

The weights $\alpha_k^{(1)}, \ldots, \alpha_k^{(6)}$ are selected for each class through a LOO procedure detailed in Section 3.3, using a $6K$-dimensional hyper-grid. The estimation of $\mathbf{\Sigma}_k$ from this covariance regularization method is slightly biased, but has much smaller variability than the standard RDA procedure. However the computation cost is considerable if K is large.

To address the high computational load, Kuo and Landgrebe (2002) proposed a simpler covariance regularization method (LOOC-2) that use combinations of the covariance components in (3.15) above, and then selects among them using a LOO procedure. A weakness of this modification is that it assumes $\mathbf{\Sigma}$ can be best estimated using only two of the six covariance components.

Robinson (2009) proposed a covariance regularization method, known as cross validated covariance mixing (CVCM) that allows for individual weights to be used in a linear combination of estimated covariance structures. For CVCM the researcher selects the number and type of covariance structures to combine for estimating $\mathbf{\Sigma}_k$, $k = 1, 2, \ldots, K$. For example, if we select M covariance estimators, denoted by $\mathbf{W}_1, \ldots, \mathbf{W}_M$, for estimating $\mathbf{\Sigma}$ then

$$\hat{\mathbf{\Sigma}}_k^{CVCM}(\gamma_k, \boldsymbol{\lambda}_k) = \gamma_k \left[\lambda_1 \mathbf{W}_1 + \lambda_2 \mathbf{W}_2 + \cdots + \left(1 - \sum_{j=1}^{p-1} \lambda_j \right) \mathbf{W}_M \right]. \tag{3.16}$$

The CVCM covariance regularization method has the benefit of drastically decreasing the estimator variability but has two major drawbacks. The first problem is that, like the LOOC-1 estimator in (3.15), each proposed covariance matrix estimator must be obtained by using a LOO procedure. In particular if K is large, then this process can be highly computationally expensive. Second, the number of

terms in the proposed covariance estimator can be limited by the researcher, but the algorithmic automation of the RDA method will be lost.

3.5 Kullback Leibler Divergence: Kullback and Leibler (1951)

Let $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ represent two multivariate normal distributions. The Kullback Leibler (KL) divergence from $\mathcal{N}_1$ to $\mathcal{N}_2$ is

$$D_{KL}(\mathcal{N}_1||\mathcal{N}_2) \equiv \frac{1}{2}\left\{\text{tr}\left[\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right] + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - p - \log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right\} \tag{3.17}$$

(Vemuri et al., 2011). Let $\mathcal{N}_k(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ represent the multivariate normal distribution for group C_k, $k = 1, 2, \ldots, K$, and $N_{(k)}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_{(k)}[\boldsymbol{\lambda}_k]\right)$ represent a proposed multivariate normal distribution with

$$\boldsymbol{\Sigma}_{(k)}[\boldsymbol{\lambda}_k] = \lambda_1\boldsymbol{\Sigma}_1 + \cdots + \lambda_{k-1}\boldsymbol{\Sigma}_{k-1} + \lambda_{k+1}\boldsymbol{\Sigma}_{k+1} + \cdots + \lambda_K\boldsymbol{\Sigma}_K; \tag{3.18}$$

that is, $\boldsymbol{\Sigma}_{(k)}[\boldsymbol{\lambda}_k]$ is a linear combination of all but the kth covariance matrix. Then the KL divergence from $\mathcal{N}_{(k)}$ to $\mathcal{N}_k$ is

$$D^*_{KL}\left(\mathcal{N}_{(k)}||\mathcal{N}_k\right) \propto \frac{1}{2}\left\{\text{tr}\left[\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{(k)}\right] - p - \log\frac{\left|\boldsymbol{\Sigma}_{(k)}\right|}{|\boldsymbol{\Sigma}_k|}\right\}. \tag{3.19}$$

Because $\boldsymbol{\Sigma}_k$ is constant with respect to each new proposed covariance estimator $\boldsymbol{\Sigma}_{(k)}$, the modified KL divergence can be written

$$D^*_{KL}\left(\mathcal{N}_{(k)}||\mathcal{N}_k\right) \equiv \left\{\text{tr}\left[\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{(k)}\right] - \log\frac{\left|\boldsymbol{\Sigma}_{(k)}\right|}{|\boldsymbol{\Sigma}_k|}\right\}. \tag{3.20}$$

3.6 Covariance Matrix Regularization using the Kullback Leibler Divergence

The one essential requirement of the modified KL divergence is that both $\boldsymbol{\Sigma}_{(k)}$ and $\boldsymbol{\Sigma}_k$ must be nonsingular. Since singularity of any of the covariance matrices is not unusual, an algorithm that addresses nonsingularity when it exists is desirable. Theorem 1 below details both a new KL divergence computation that can adapt to the possible singularity of one of the covariance structures, and states the properties

of the algorithm. Moreover, if the covariance matrices are nonsingular, the new divergence measure is equivalent to the KL criterion in (3.20).

Theorem 1. *Let* $\mathbf{\Sigma}_k$ *and* $\mathbf{\Sigma}_{(k)}$ *represent two symmetric, nonnegative-definite matrices, and let* $\mathbf{\Phi}_k^T \mathbf{\Xi}_k \mathbf{\Phi}_k$ *and* $\mathbf{\Phi}_{(k)}^T \mathbf{\Xi}_{(k)} \mathbf{\Phi}_{(k)}$ *be the eigenvalue decomposition of* $\mathbf{\Sigma}_k$ *and* $\mathbf{\Sigma}_{(k)}$*, respectively. Then* D_{KL}^* *is a function of the generalized eigenvalues of* $\mathbf{\Sigma}_k$ *and* $\mathbf{\Sigma}_{(k)}$*.*

Proof. Select ν such that

$$\nu = \min \left\{ n_{\mathbf{\Xi}_k}, n_{\mathbf{\Xi}_{(k)}} \right\}, \tag{3.21}$$

where $n_{\mathbf{\Xi}}$ is the number of non-zero eigenvalues. We can define $\mathbf{\Phi}_k^* \in \mathbb{R}^{p \times \nu}$ to be a basis for the space spanned by the columns of $\mathbf{\Sigma}_k$, and corresponding to the non-zero eigenvalues of $\mathbf{\Sigma}_k$. Then,

$$\mathbf{\Phi}_k^* \mathbf{\Phi}_k^T \mathbf{\Xi}_k \mathbf{\Phi}_k \mathbf{\Phi}_k^{*T} = \mathbf{\Xi}_k^*$$

and

$$\mathbf{\Phi}_k^* \mathbf{\Phi}_{(k)}^T \mathbf{\Xi}_{(k)}^* \mathbf{\Phi}_{(k)} \mathbf{\Phi}_k^{*T} = \mathbf{\Omega},$$

where $\mathbf{\Omega}$ is a symmetric positive-definite matrix, and $\mathbf{\Xi}_k^*$ is a diagonal matrix of the non-zero eigenvalues of $\mathbf{\Sigma}_k$. Notice that

$$\left[\mathbf{\Xi}_k^*\right]^{-1/2} \mathbf{\Xi}_k^* \left[\mathbf{\Xi}_k^*\right]^{-1/2} = \mathbf{I}_\nu, \tag{3.22}$$

and

$$\left[\mathbf{\Xi}_k^*\right]^{-1/2} \mathbf{\Omega} \left[\mathbf{\Xi}_k^*\right]^{-1/2} = \mathbf{\Omega}^*, \tag{3.23}$$

where $\mathbf{\Omega}^* \in \mathbb{R}_{\nu \times \nu}^+$. Because both matrices (3.22) and (3.23) are nonsingular and symmetric, there exists a matrix $\mathbf{D} = \mathbf{\Phi}_{\Omega^*}$ that simultaneously diagonalizes (3.22) and (3.23) . The modified KL divergence in (3.20) is therefore,

$$\begin{aligned} D_{KL}^* \left(\mathbf{\Xi}_{(k)} || \mathbf{\Xi}_k \right) &= \left\{ \operatorname{tr} \left[\mathbf{I}_\nu^{-1} \mathbf{\Xi}_\Omega^* \right] - \log \frac{|\mathbf{\Xi}_\Omega^*|}{|\mathbf{I}_\nu|} \right\}, \\ &= \left\{ \operatorname{tr} \left[\mathbf{\Xi}_\Omega^* \right] - \log |\mathbf{\Xi}_\Omega^*| \right\}, \end{aligned} \tag{3.24}$$

where $\mathbf{\Xi}_\Omega^*$ are the generalized eigenvalues of $\left[\mathbf{\Xi}_k^*\right]^{-1} \mathbf{\Omega}$. □

To understand the computational gain of our proposed KL optimization regularization method, consider the case of $K = 4$ and suppose we wish to evaluate each regularization parameter from 0 to 1 using five equally spaced values. Then the LOOC-1 regularization method would involve cross validation in 24 dimensions resulting in $5.96e{+}16$ LOO computations. Although the number of computation routines was not available for the simulation results, preliminary investigation yielded a worst case of 30 KL evaluations for each class. Therefore a reasonable expectation for the number of computations would be 120 KL evaluations and 25 LOO computations.

3.7 Optimization Algorithm

We now outline our proposed covariance regularization algorithm based upon optimizing (3.24). By using optimization, we can reduce the reliance of cross validation in a high-dimensional space, while retaining the flexibility of the Robinson (2009) and Kuo and Landgrebe (2002) covariance regularization methods.

(1) Estimate $\boldsymbol{\Sigma}_k$ for $k = 1, 2, \ldots, K$ using

$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} \left(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_k\right)^T \left(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_k\right). \tag{3.25}$$

(2) For each $k = 1, 2, \ldots K$

(a) Let $\boldsymbol{\Lambda} \in \mathbb{R}^{K-1}$ represent the parameter space for a vector of weights representing a linear combination of covariance structures that comprise $\hat{\boldsymbol{\Sigma}}_{(k)}$. We wish to select an optimal weighting vector $\boldsymbol{\lambda}_k$ to satisfy

$$\min_{\boldsymbol{\lambda}_k \in \boldsymbol{\Lambda}} \left\{ D^*_{KL}\left(\mathcal{N}_{(k)} || \mathcal{N}_k\right) \right\}, \tag{3.26}$$

subject to $\sum_{j=1}^{K-1} \lambda_j = 1$, where $\boldsymbol{\lambda}_j = (\lambda_1, \lambda_2, \ldots, \lambda_{K-1})'$.

(b) Let $\boldsymbol{\Sigma}_{(k)}\left[\boldsymbol{\lambda}_k\right] \equiv \boldsymbol{\Sigma}_{p_k}$, where $\boldsymbol{\Sigma}_{p_k}$ represents the proposed pooled covariance matrix corresponding to the population C_k.

(3) Run an appropriate LOO cross validation to estimate the weights $(\hat{\lambda}_k, \hat{\gamma}_k)$ for each class in $\mathcal{L}$.

The gradient vector for D^*_{KL} is intractable, so numerical derivatives must be used with a constrained optimization function in the R computational package. We expect our KL-based RDA classifier to result in a lower misclassification rate compared to the standard RDA classifier.

3.8 Simulation

To examine the properties of our proposed regularized covariance estimation algorithm, we designed a simulation study to compare the error rate of the RDA classifier using the KL optimization routine to estimate $\mathbf{\Sigma}_k$ and the error rate of the Friedman's RDA classifier. We generated data from $K = 4$ classes according to the models

$$\mathbf{x}_{jk} \sim \text{MVN}(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k),$$

where $n_k = 20,\ k = 1, \ldots, 4$, and $j = 1, 2, \ldots, n_k$. The four mean vectors, each of length 20, are $\boldsymbol{\mu}_1 = \mathbf{0}_{20}$, $\boldsymbol{\mu}_2 = 0.5\mathbf{1}_{20}$, $\boldsymbol{\mu}_3 = \mathbf{1}_{20}$, and $\boldsymbol{\mu}_4 = 1.5\mathbf{1}_{20}$, where $\mathbf{0}_\kappa$ and $\mathbf{1}_\kappa$ are length κ vectors of zeros and ones respectively. We remark that the mean vectors contain a minimal amount of classification information.

We examined two general covariance structures. The first is an intra-class structure, where the covariance between the covariates is constant. For notational simplicity, we let $\mathbf{\Sigma}_\kappa[\sigma^2, \rho]$ denote an $\kappa \times \kappa$ intra-class covariance matrix with variance σ^2 and constant covariance ρ.

We consider the following three cases.

(1) Case I-1: All four population covariance matrices are 20×20 identity matrices; that is $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3 = \mathbf{\Sigma}_4 = \mathbf{I}_{20}$

(2) Case I-2: $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_{20}[5, 0.25]$, and $\mathbf{\Sigma}_3 = \mathbf{\Sigma}_4 = \mathbf{I}_{20}$.

(3) Case I-3:

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_{20}[5, 0.25],$$

$$\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_{20}[4.9, 0.24],$$

$$\boldsymbol{\Sigma}_3 = \mathbf{I}_{20},$$

and

$$\boldsymbol{\Sigma}_4 = \boldsymbol{\Sigma}_{20}[1.0, 0.10].$$

In case I-1 we use the most basic covariance structure the identity matrix for all four classes. In case I-2 the classes C_1 and C_2 have the identity covariance, and C_3 and C_4 have an equal non-identity intra-class covariance matrix. Case I-3 is the most complex of the three cases. In case I-3, the covariance matrices for classes one and two are not equal. However, the differences in the variances between cases one and two is very small (0.01); likewise for the difference in the covariances (0.01). The third class covariance matrix is the identity. The fourth class covariance has a variance of one (like class three), but a covariance of 0.1. Summarily, the covariance matrices of classes one and two differ only slightly; the covariance matrices for classes three and four differ very slightly. But the covariances between pairs of classes one and two, and classes three and four, are quite different.

The second set of cases considered in our simulation uses a Toeplitz matrix for defining the covariance matrix. Because a symmetric Toeplitz matrix is indexed by its first row, we use the notation

$$\text{Toeplitz}\,[a, b, c] = \begin{bmatrix} a & b & c \\ b & a & b \\ c & b & a \end{bmatrix}$$

to indicate a Toeplitz matrix with first row having elements a, b, and c.

We investigate three Toeplitz covariance structures that model diminishing covariance between covariates. Specifically we have the following three cases.

(1) Case T-1: For $k = 1, 2, 3, 4$,

$$\boldsymbol{\Sigma}_k = \text{Toeplitz}[\ 1 \quad 0.55 \quad 0.10 \quad \mathbf{0}_{17}\ ].$$

(2) Case T-2:

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \text{Toeplitz}[\ 5 \quad 2.55 \quad 0.10 \quad \mathbf{0}_{17}\ ]$$

and

$$\boldsymbol{\Sigma}_3 = \boldsymbol{\Sigma}_4 = \text{Toeplitz}[\ 1 \quad 0.55 \quad 0.10 \quad \mathbf{0}_{17}\ ].$$

(3) Case T-3:

$$\boldsymbol{\Sigma}_1 = \text{Toeplitz}[\ 5.0 \quad 2.55 \quad 0.10 \quad \mathbf{0}_{17}\ ],$$

$$\boldsymbol{\Sigma}_2 = \text{Toeplitz}[\ 4.9 \quad 2.50 \quad 0.10 \quad \mathbf{0}_{17}\ ],$$

$$\boldsymbol{\Sigma}_3 = \text{Toeplitz}[\ 4.8 \quad 2.45 \quad 0.10 \quad \mathbf{0}_{17}\ ],$$

and

$$\boldsymbol{\Sigma}_4 = \text{Toeplitz}[\ 1.0 \quad 0.55 \quad 0.10 \quad \mathbf{0}_{17}\ ].$$

Case T-1 is similar to case I-1 in that the Toeplitz covariance matrices across all classes are equal. In case T-2, class C_1 and C_2 have equal Toeplitz covariance matrices, and C_3 and C_4 have equal covariance matrices which are not equal to the covariance matrix of classes C_1 and C_2.

The estimated expected misclassification rate and its corresponding estimated standard error for each parameter configuration considered in our simulation are given in Table 3.1. The standard error of the error rate is estimated by

$$\widehat{s.e}(\epsilon) = \sqrt{\sum_{i=1}^{500} \frac{(\epsilon_i - \bar{\epsilon})^2}{499}},$$

where ϵ_i is the misclassification rate for the i^{th} replication, and $\bar{\epsilon} = 1/500 \sum_{i=1}^{500} \epsilon_i$. For each parameter configuration, there are two different figures shown in Appendix B. The first figure is a 2×2 grid of box plots, displaying the output of the KL optimization routine for each covariance matrix. On these figures (as in Table 3.1) RDA represents Friedman's method, and RDA-KL represents our new optimization method. To illustrate, refer to Figure B.1 on page 60. The box plot in the upper left corner displays the output of $\boldsymbol{\lambda}_1$ from the KL optimization routine for case I-1. The plot in the upper right hand corner corresponds to $\boldsymbol{\lambda}_2$ for each replication, and so on. On the following page, Figure B.2 is a box plot of the estimated misclassification rates for the RDA and RDA-KL regularization methods.

Table 3.1. Misclassification Rates and Standard Errors

Case	RDA	RDA-KL
I-1	0.3822 (0.0660)	0.3680 (0.0617)
I-2	0.3689 (0.0695)	0.3613 (0.0572)
I-3	0.3867 (0.0706)	0.3766 (0.0590)
T-1	0.4719 (0.0573)	0.4757 (0.0587)
T-2	0.4334 (0.0802)	0.4234 (0.0612)
T-3	0.4844 (0.0707)	0.5398 (0.0534)

3.9 Simulation Results

The simulation results indicate both the potential for our regularized covariance estimation method as well as some of its current limitations. The most promising aspect of the RDA-KL regularization method is the decreased standard error of the misclassification rate. This aspect allows for a more predictable and stable performance of the classifier compared to the standard RDA. The new RDA-KL regularization algorithm performed well for cases where each class had an equal or very similar pair of covariance matrices, specifically I-2, I-3, and T-2. However, as

the differences in misclassification rates are not large enough to be attributed to any factor except noise, a simulation with more replication is warranted.

We will first discuss the cases for which the KL optimization algorithm performed best. From Figures B.3, B.5, and B.9 containing the plots of $\boldsymbol{\lambda}_k$ in Appendix B, we can see that cases I-2, I-3, and T-2 have similar patterns for $\boldsymbol{\lambda}$. The estimated covariance matrices $\hat{\boldsymbol{\Sigma}}_1$ and $\hat{\boldsymbol{\Sigma}}_2$ had similar assigned weight and $\hat{\boldsymbol{\Sigma}}_3$ and $\hat{\boldsymbol{\Sigma}}_4$ were also weighted similarity. For each of these cases, classes C_1 and C_2 had either identical or very similar population covariance structures, and C_3 and C_4 shared a structure. The KL algorithm behaved as expected by pairing each sample class covariance matrix with its closest match in terms of their shared eigenvalue structure, and did so consistently across replications.

We will now examine the cases where the KL optimization algorithm failed to perform as expected. For cases I-1 and T-1, the selected weights were highly variable, giving rise to the conjecture that the KL algorithm has a tendency to heavily weight the one covariance matrix that is most similar in eigenvalue structure, while putting little weight on the others. However for I-1, the RDA-KL method had a smaller misclassification rate compared to RDA, as the parameter space was smaller relative to the sample size. In I-1 there were two parameters for the covariance matrix, and for case T-3 there were three, so the ratio n_k/p was 10 for I-1 and 6.67 for T-1.

The RDA-KL regularization routine performed the worst for case T-3. This fact is explainable through closer examination of the selected weight coefficients for each iteration. Similar to cases I-1 and T-1, the RDA-KL method had a tendency to place large weight on the one covariance matrix covariance matrix that was most similar in eigenvalue structure, and placed little to no weight on the remaining estimated covariance structures. It should be noted that this is advantageous in configurations such as I-2, I-3, and T-2 where pairs of classes share a similar covariance structure. This fact gives rise to an algorithm extension that would use a

forward model selection procedure, similar to variable selection algorithms in linear regression, where components are added until the difference between the remaining structures cannot be well established.

APPENDICES

APPENDIX A

Appendix: Simulation Results and Code for Chapter Two

A.1 Estimation of λ

Table A.1. Test for Gaussianity

model	n	GCBG.3	GABG.3	GCBG.1	GABG.1	GCBG.2	GABG.2
ARMA(2,2)	100	0.0524	0.0498	0.0524	0.0498	0.0524	0.0498
ARMA(2,2)	250	0.0482	0.0462	0.0482	0.0462	0.0482	0.0462
ARMA(2,2)	500	0.0500	0.0512	0.0500	0.0512	0.0500	0.0512
ARMA(2,2)	1000	0.0500	0.0488	0.0500	0.0488	0.0500	0.0488
WN(1)	100	0.0522	0.0514	0.0522	0.0514	0.0522	0.0514
WN(1)	250	0.0554	0.0528	0.0554	0.0528	0.0554	0.0528
WN(1)	500	0.0498	0.0514	0.0498	0.0514	0.0498	0.0514
WN(1)	1000	0.0522	0.0524	0.0522	0.0524	0.0522	0.0524
AR(2)	100	0.0446	0.0418	0.0446	0.0418	0.0446	0.0418
AR(2)	250	0.0562	0.0570	0.0562	0.0570	0.0562	0.0570
AR(2)	500	0.0572	0.0548	0.0572	0.0548	0.0572	0.0548
AR(2)	1000	0.0538	0.0520	0.0538	0.0520	0.0538	0.0520

Table A.2. Test for Linearity

Model	n	GCBL.3	GABL.3	GCBL.1	GABL.1	GCBL.2	GABL.2
ARCH	100	0.4754	0.4842	0.0020	0.0020	0.1102	0.1142
ARCH	250	0.6354	0.6508	0.0000	0.0000	0.2054	0.2184
ARCH	500	0.7294	0.7490	0.0000	0.0000	0.3836	0.4038
ARCH	1000	0.8102	0.8378	0.0000	0.0000	0.5698	0.5854
Bilinear	100	0.4082	0.4156	0.0006	0.0002	0.0804	0.0792
Bilinear	250	0.6046	0.6164	0.0002	0.0002	0.0914	0.0934
Bilinear	500	0.8146	0.8256	0.0000	0.0000	0.1628	0.1672
Bilinear	1000	0.9510	0.9538	0.0000	0.0000	0.2590	0.2598

Continued on Next Page...

Model	n	GCBL.3	GABL.3	GCBL.1	GABL.1	GCBL.2	GABL.2
χ^2	100	0.1092	0.1098	0.0004	0.0004	0.0666	0.0628
χ^2	250	0.0908	0.0910	0.0000	0.0000	0.0572	0.0576
χ^2	500	0.0750	0.0762	0.0002	0.0002	0.0604	0.0624
χ^2	1000	0.0774	0.0782	0.0000	0.0000	0.0596	0.0598
NLAR	100	0.0712	0.0714	0.0100	0.0112	0.0338	0.0312
NLAR	250	0.1206	0.1238	0.0042	0.0044	0.0438	0.0436
NLAR	500	0.1310	0.1378	0.0024	0.0036	0.0284	0.0268
NLAR	1000	0.1716	0.1798	0.0020	0.0026	0.0168	0.0160
NLMA	100	0.1614	0.1646	0.0022	0.0024	0.0382	0.0336
NLMA	250	0.2058	0.2078	0.0010	0.0010	0.0420	0.0412
NLMA	500	0.2340	0.2424	0.0002	0.0002	0.0482	0.0480
NLMA	1000	0.3106	0.3174	0.0002	0.0002	0.0432	0.0440
SETAR	100	0.2642	0.2680	0.0002	0.0004	0.0526	0.0486
SETAR	250	0.4392	0.4496	0.0002	0.0002	0.0610	0.0558
SETAR	500	0.6602	0.6722	0.0002	0.0000	0.0542	0.0520
SETAR	1000	0.9046	0.9154	0.0000	0.0000	0.0480	0.0458
GARCH	100	0.1488	0.1486	0.0070	0.0076	0.0310	0.0288
GARCH	250	0.5022	0.5146	0.0010	0.0020	0.0482	0.0538
GARCH	500	0.8346	0.8470	0.0002	0.0004	0.2396	0.2424
GARCH	1000	0.9732	0.9760	0.0000	0.0000	0.5620	0.5656
EXPAR	100	0.0522	0.0444	0.0020	0.0020	0.0734	0.0734
EXPAR	250	0.0672	0.0646	0.0002	0.0002	0.0638	0.0646
EXPAR	500	0.0556	0.0574	0.0000	0.0002	0.0694	0.0708
EXPAR	1000	0.0248	0.0248	0.0000	0.0000	0.0684	0.0692
STAR	100	0.0428	0.0420	0.0092	0.0104	0.0414	0.0360
STAR	250	0.0562	0.0550	0.0048	0.0052	0.0414	0.0440
STAR	500	0.0576	0.0564	0.0082	0.0074	0.0464	0.0452
STAR	1000	0.0564	0.0550	0.0126	0.0126	0.0398	0.0394

A.2 Code

```
dyn.load("allfortran.dll")

source("allrcode.R")

source("functions4.R")


hinich <- function(y)

{

n <- length(y)
```

```
m <- trunc(n^0.625)

zz      <- bispec(y)
z       <- zz$z[which(zz$z > 0)]
sz <- sum(z)
pvalsg <- 1 - pchisq(sz, df = 2*zz$nsq)

b.iqr <- IQR(z)
ncp.est <- max(0,mean(z)-2)
eta1    <- qchisq(0.25, df = 2, ncp = ncp.est)
eta3    <- qchisq(0.75, df = 2, ncp = ncp.est)
    d1      <- dchisq(0.25, df = 2, ncp = ncp.est)
    d3      <- dchisq(0.75, df = 2, ncp = ncp.est)
    viqr    <- (3/d1**2 + 3/d3**2 - 2/(d1*d3))/(16*zz$nsq)
l.ts <- (b.iqr - (eta3 - eta1))/sqrt(viqr)
    pvalsl <- 1 - pnorm(l.ts)

rej <- c(as.numeric(pvalsg < 0.05),as.numeric(pvalsl < 0.05))
return(rej)
}

boot.linear.gof  <- function(n,res,ar.coef)
{
  s <-sample(res,size=n,replace=TRUE)
  bootL <-filter(s,ar.coef,method="recursive")
  return(bootL)
}
```

```
boot.normal.gof <- function(n,sd.res,ar.coef)
{
  s <-rnorm(n,0,sd.res)
  bootG <-filter(s,ar.coef,method="recursive")
  return(bootG)
}

boot.symmetric <- function(res,ar.coef)
{
s <-abs(sample(res,replace=T))*rbinom(length(res),1,.5)
bootS <-filter(s,ar.coef,method="recursive")
return(bootS)
}

##
#################################################################
## BEGIN: Sim Code

norm.trans <- function(z,df,lam)
{
s <- df + 2*lam
r <- lam + df
h <- 1-(2*r*(df+3*lam)/(3*s^2))
Y <- (z/r)^h

return((Y-mean(Y))/sd(Y))
}
```

```
##
#####################################################################
## Calc of CvM and AD statistics

GOF.stat <- function(z,k,fun=NULL,param=NULL,stat=c(1,2))
{
k <- length(z)
CvM <- 0.0
AnD <- 0.0
i <- 1:k
tol <- .0001

if(fun == "exp")
{
Q <- sort(pexp(z,param))
}
if(fun == "chisq") {
Q <- sort(pchisq(z,df=2,ncp=param))
  #Q <- sort(pnorm(z))
#Q <- sort(pgamma(z,2,6))
#print(mean(Q))
}
if(fun == "norm")
{
z.t <- (z-mean(z))/sd(z)
Q <- sort(pnorm(z.t))
}
```

```
Q[Q >= (1-tol)] <- 1-tol
Q[Q <= tol] <- tol

CvM <- 1/(12*k) + sum((Q - (2*i-1)/(2*k))^2)
AnD <- -k - mean((2*i-1)*log(Q*(1-rev(Q))))
#result <- c(CvM,AnD)
#return(result[stat])
return(T.adjust(c(CvM,AnD),n=k,fun=fun))
}

T.adjust <- function(T,n,fun=NULL)
{

if(fun == "norm"){
T[1] <- T[1]*(1 + 0.5/n)
T[2] <- T[2]*(1 + 0.75/n + 2.25/(n^2))
}
if(fun == "exp"){
T[1] <- (T[1]- 0.4/n + 0.6/(n^2))*(1+1/n)
}

return(T)
}

##
################################################################
## GoF test as specified in Jahan & Harvill 2008
```

```
gof.jahan <- function (x,cv=NULL)
{
n <- length(x)
m <- trunc(n^0.625)
z <- as.vector(bispec(x)$z)
z <- z[z>0]
k <- length(z)

T.g <- GOF.stat(z=z,fun="exp",param=1/2)
T.g <- T.adjust(T.g,k,fun="exp")
g.rej <- c(as.numeric(T.g[1] > 0.461),
                as.numeric(T.g[2] > 2.492))

ncp <- max(0,mean(z)-2)

if(is.null(cv))
{
Y <- norm.trans(z,df=2,lam=ncp)
T.l <- GOF.stat(z=Y,k=k,fun="norm")
l.rej <- c(as.numeric(T.l[1] > 0.126),
                    as.numeric(T.l[2] > 0.752))
}else
{
T.l <- GOF.stat(z=z,k=k,fun="chisq",param=ncp)
l.rej <- c(as.numeric(T.l[1] > cv[1]),
                    as.numeric(T.l[2] > cv[2]))
}
```

```
#from simulation of iid variables: 0.3942588 2.5487131
#l.rej <- c(as.numeric(T.l[1] >0.3942588),
                 as.numeric(T.l[2] > 2.5487131))
return(c(g.rej,l.rej))
#return(c(T.g,T.l))
}

##
################################################################
## Get initial value of ncp for GoF test

lam.burnin <- function(Data,param,NITER = 1000,
                       ar_order=10,fun=NULL)
{
n <- length(Data)
fit <- ar(Data,FALSE,ar_order)
ar.coef <- fit$ar
res <- fit$resid[-(1:ar_order)]
centered.res<- res-mean(res)

A <- vector(length=NITER)
for(i in 1:NITER){
bootL <- boot.linear.gof(n,centered.res,ar.coef)
z.L <- bispec(bootL)$z
  z.L <- as.vector(z.L[z.L > 0])
A[i] <- GOF.stat(z=z.L,k=length(z.L),
                           fun=fun,param=param,stat=2)
```

```
}
plot(ecdf(A))
return(quantile(sort(A),.95))
}

##
################################################################
## Various ways of estimating the ncp

get.lam <- function(x,type=NULL,M_b=NULL)
{
n <- length(x)
if(type == "berg"){
lambda <- n*Cfun(0,0,x)^2/(M_b^2*1.375*R(0,x)^3)
}else if(type == "standard"){
lambda <- max(0,mean(x)-2)
}else if(type == "opt")
{
lambda <- optimize(
         f=function(ncp){sum(dchisq(x,df=2,ncp=ncp,log=TRUE))},
         interval=c(min(x),max(x)),maximum=TRUE)$maximum
}

  return(lambda)
}
##
################################################################
```

```
## Bootstrap Test

boot.gof.test <- function(bootnum,Data,ar_order=NULL,
                          alpha=NULL,method=3)
{
n <- length(Data)

#calculate centered residuals
fit <- ar(Data,FALSE,ar_order)
ar.coef <- fit$ar
ar_order <- length(ar.coef)
res <- fit$resid[-(1:ar_order)]
centered.res<- res-mean(res)
sd.res <- sd(res)

# Calculate Bispectrum for RAW data
z <- bispec(Data)$z
  z <- as.vector(z[z>0])

#Hinich gauss and lin test stat
T_Hg <- sum(z)
T_Hl <- IQR(z)

# Test Statistic for Gaussian hypothesis
T_g <- GOF.stat(z,k,fun="exp",param=1/2)

#Test statistic for linear hypothesis
```

```
if(method==1){ #use global ncp
lam <- get.lam(z,type="standard")
T_l <- GOF.stat(z,fun="chisq",param=lam)
}

if(method==2){ #unique ncp for each iteration
lam <- get.lam(z,type="standard")
T_l <- GOF.stat(z,fun="chisq",param=lam)
}

if(method==3){ #use ncp=0
T_l <- GOF.stat(z,fun="chisq",param=0)
}
if(method==4){ #use F=pnorm
z.m <- mean(z)
sd.m <- sd(z)
T_l <- GOF.stat(z,fun="norm",param=c(z.m,sd.m))
}

#test statistics for CvM and AnD
GOF_g <- cbind(double(bootnum),double(bootnum))
GOF_l <- cbind(double(bootnum),double(bootnum))
Hin_g <- vector(length=bootnum)
Hin_l <- vector(length=bootnum)
#lambda.L<- vector(length=bootnum)

for(i in 1:bootnum)
```

```
{
#########################################################
#gaussian test bootstrap

bootG <- boot.normal.gof(n,sd.res,ar.coef)
z.G <- bispec(bootG)$z
  z.G <- as.vector(z.G[z.G > 0])

#Hinich gauss est.
Hin_g[i] <- sum(z.G)

#GOF gauss est.
GOF_g[i,] <- GOF.stat(z.G,fun="exp",param=1/2)

#########################################################
# linear test bootstrap

bootL <- boot.linear.gof(n,centered.res,ar.coef)
z.L <- bispec(bootL)$z
  z.L <- as.vector(z.L[z.L > 0])

#Hinich lin est.
Hin_l[i] <- IQR(z.L)

#GOF lin est.
if(method==1){ #use global ncp
GOF_l[i,] <- GOF.stat(z.L,fun="chisq",param=lam)
```

```
}

if(method==2){
lam <- get.lam(z.L,type="standard")
GOF_l[i,] <- GOF.stat(z.L,fun="chisq",param=lam)
}
if(method==3){
GOF_l[i,] <- GOF.stat(z.L,fun="chisq",param=0)
}
if(method==4){
GOF_l[i,] <- GOF.stat(z,fun="norm",
                                    param=c(z.m,sd.m))
}
}

# Calculate p-values
C_pval_g <- 1-ecdf(GOF_g[,1])(T_g[1])
C_pval_l <- 1-ecdf(GOF_l[,1])(T_l[1])
A_pval_g <- 1-ecdf(GOF_g[,2])(T_g[2])
A_pval_l <- 1-ecdf(GOF_l[,2])(T_l[2])
H_pval_g <- 1-ecdf(Hin_g)(T_Hg)
H_pval_l <- 1-ecdf(Hin_l)(T_Hl)
return(c(C_pval_g,A_pval_g,H_pval_g,C_pval_l,A_pval_l,H_pval_l))
}
```

APPENDIX B

Appendix: Output of the KL Algorithm & Misclassification Rate

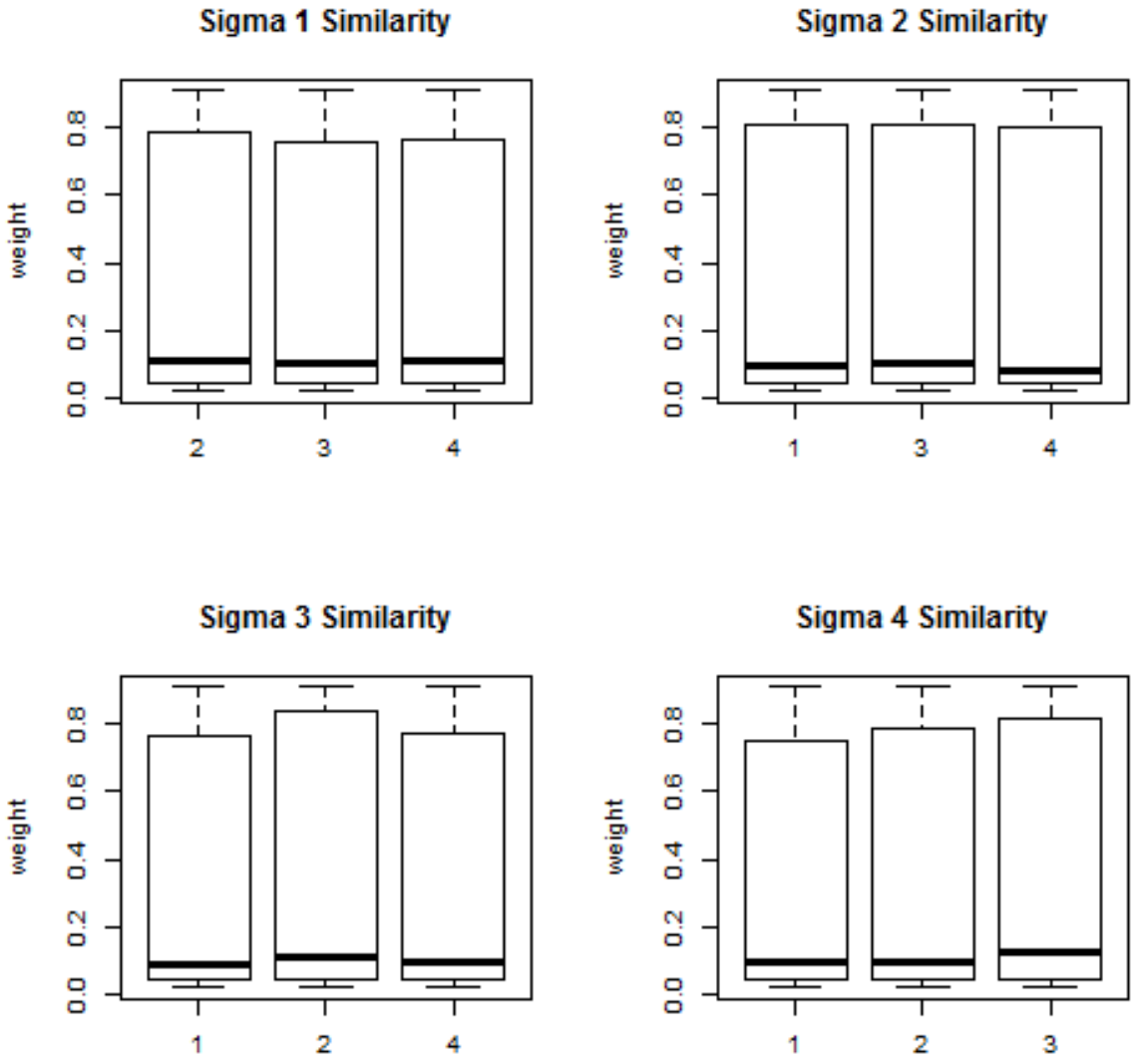

Figure B.1. Case I-1: KL Similarity Performance

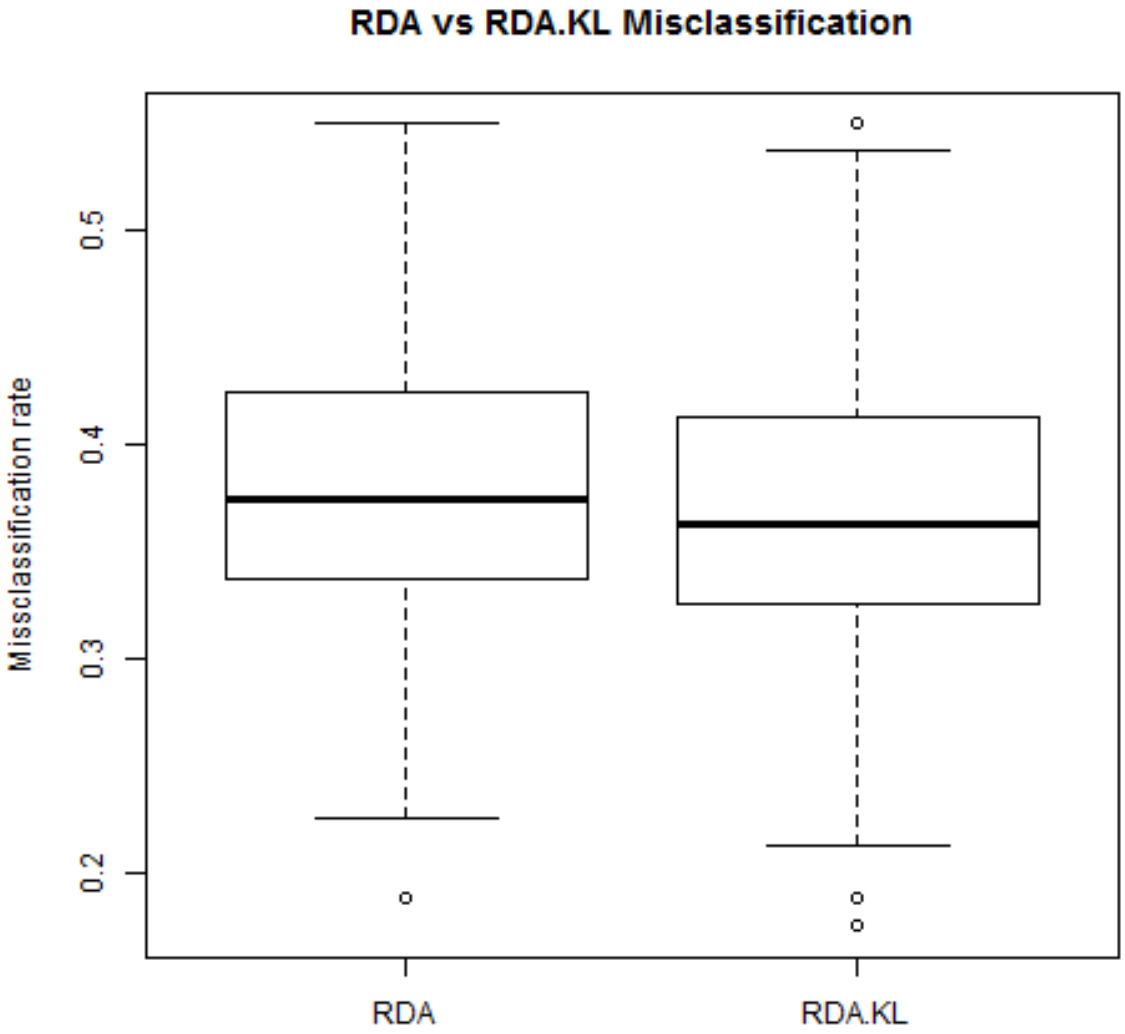

Figure B.2. Case I-1: Misclassification Rate

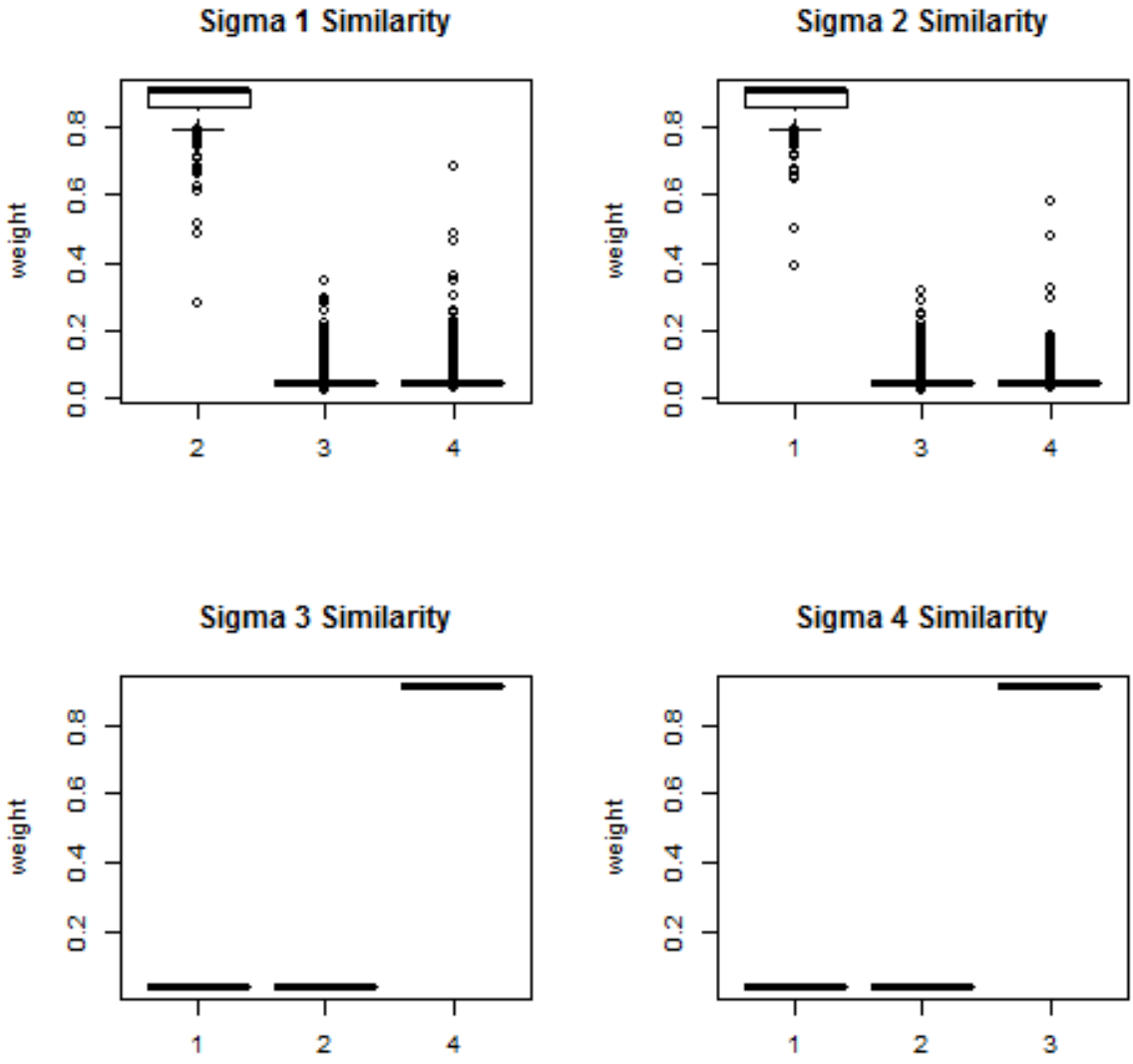

Figure B.3. Case I-2: KL Similarity Performance

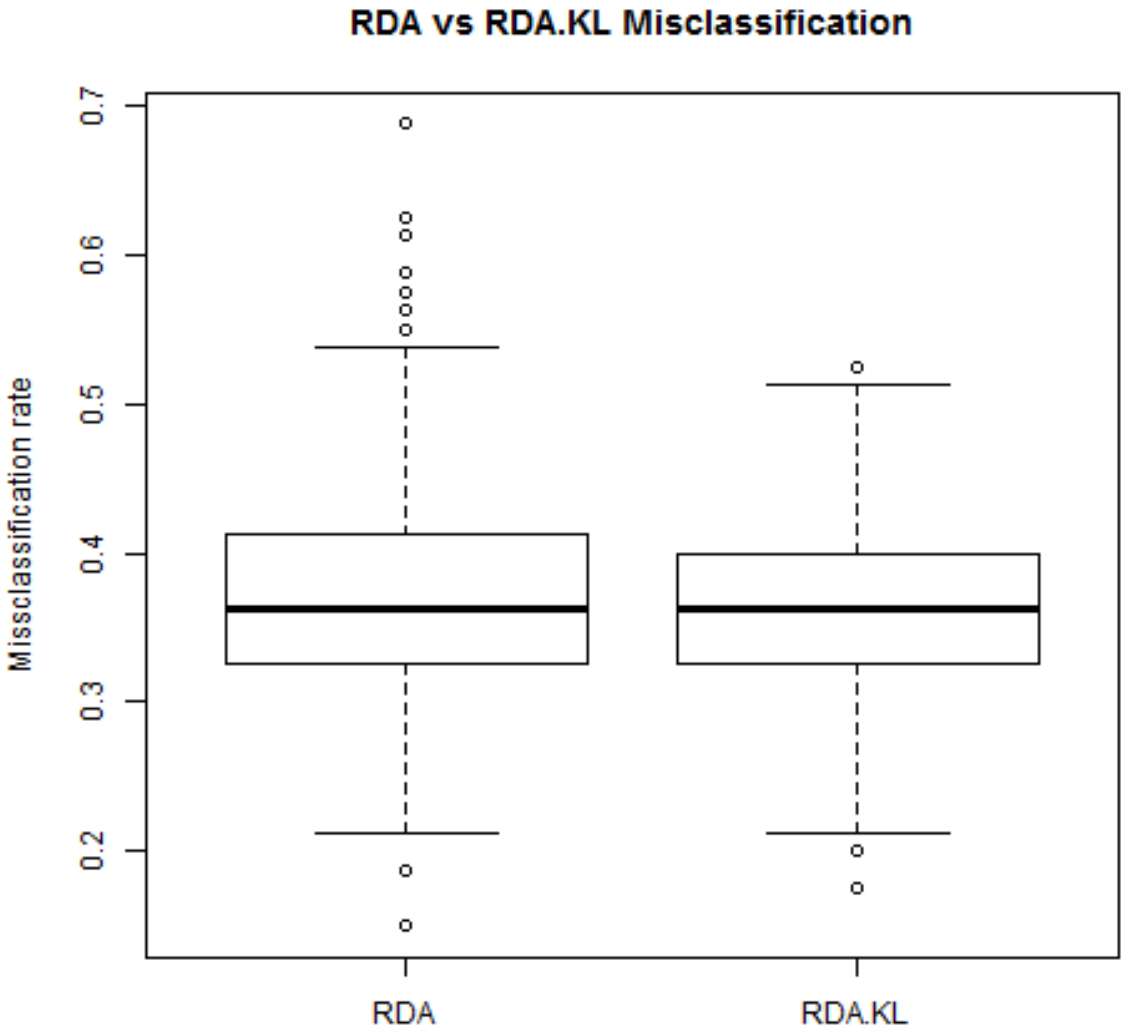

Figure B.4. Case I-2: Misclassification Rate

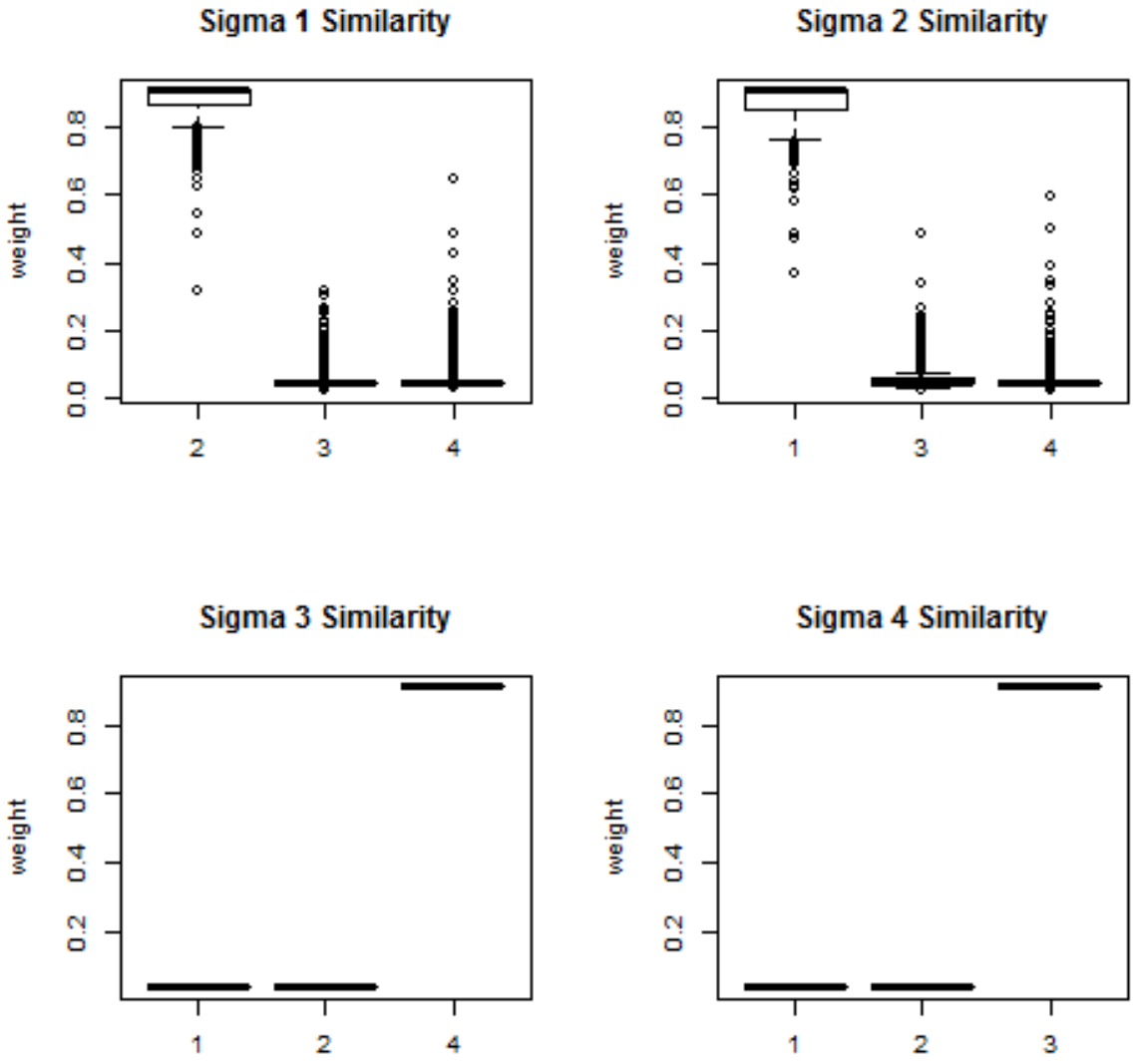

Figure B.5. Case I-3: KL Similarity Performance

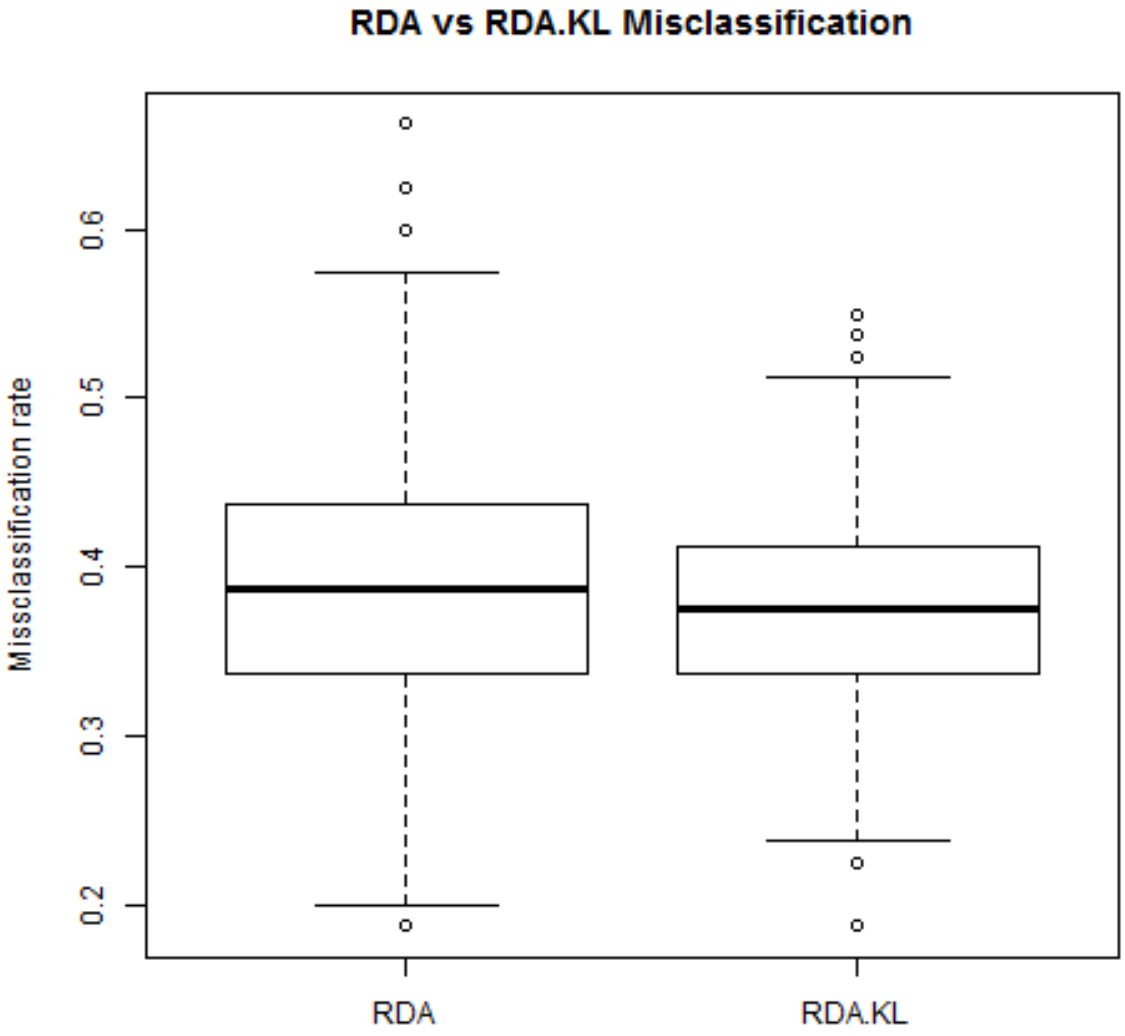

Figure B.6. Case I-3: Misclassification Rate

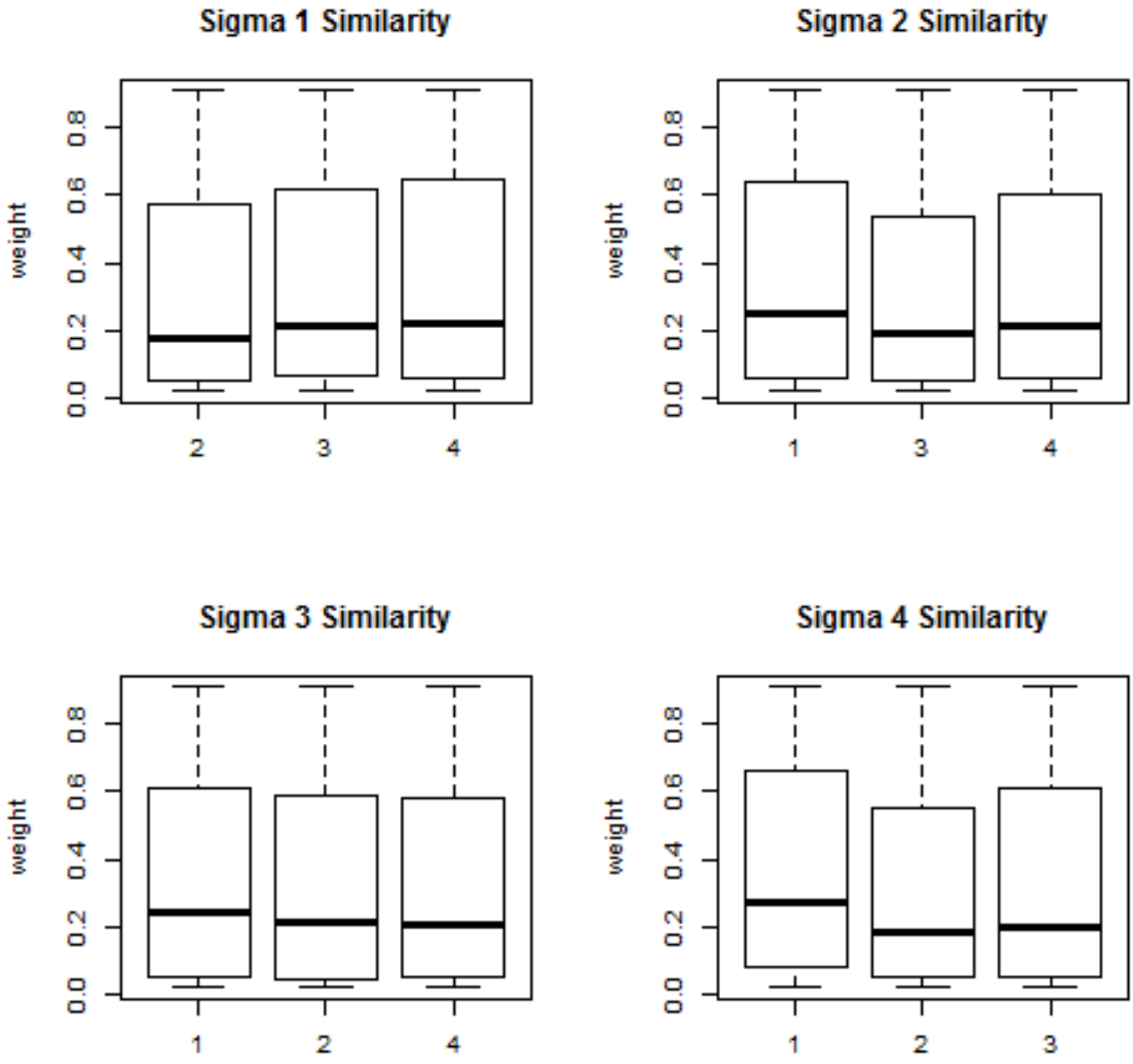

Figure B.7. Case T-1: KL Similarity Performance

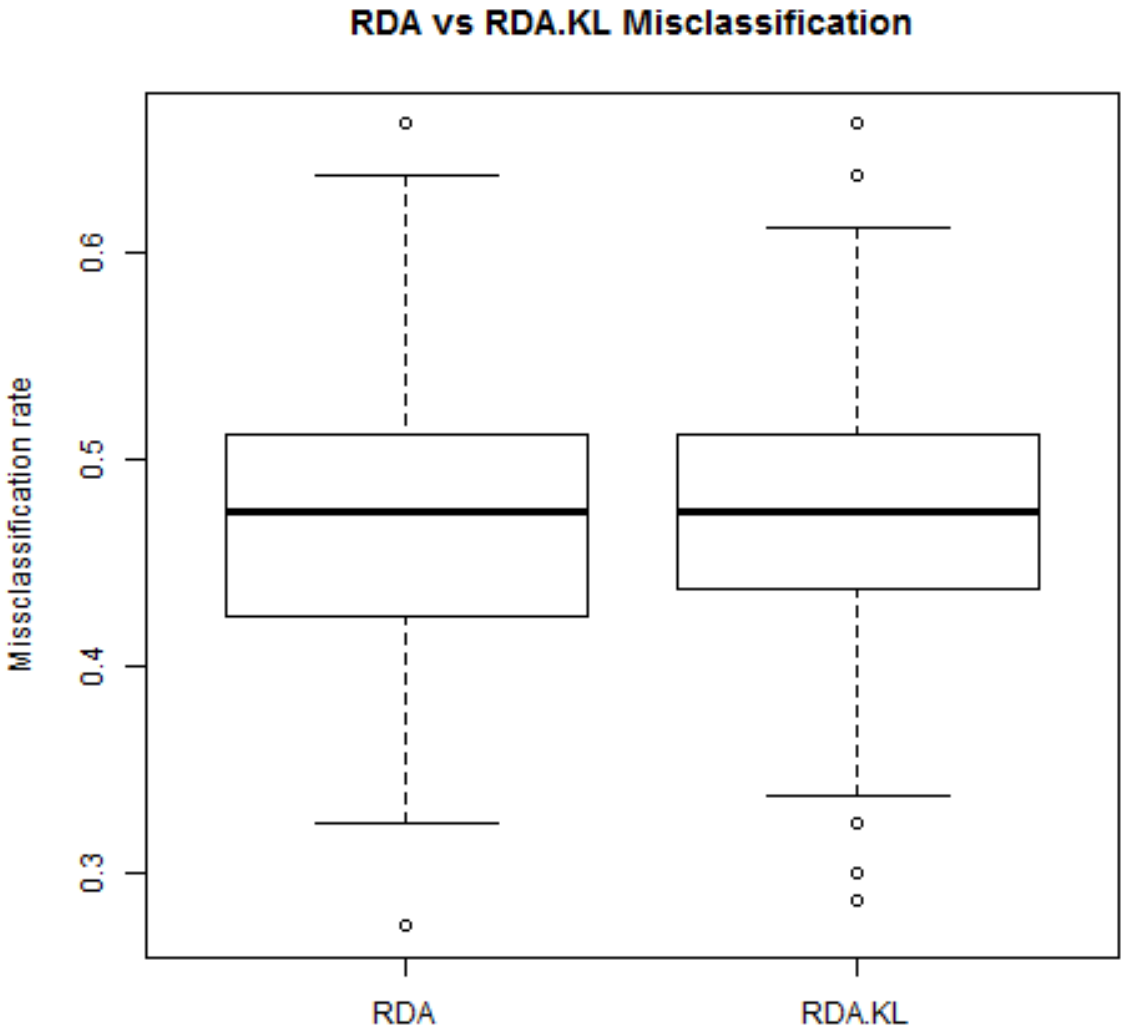

Figure B.8. Case T-1: Misclassification Rate

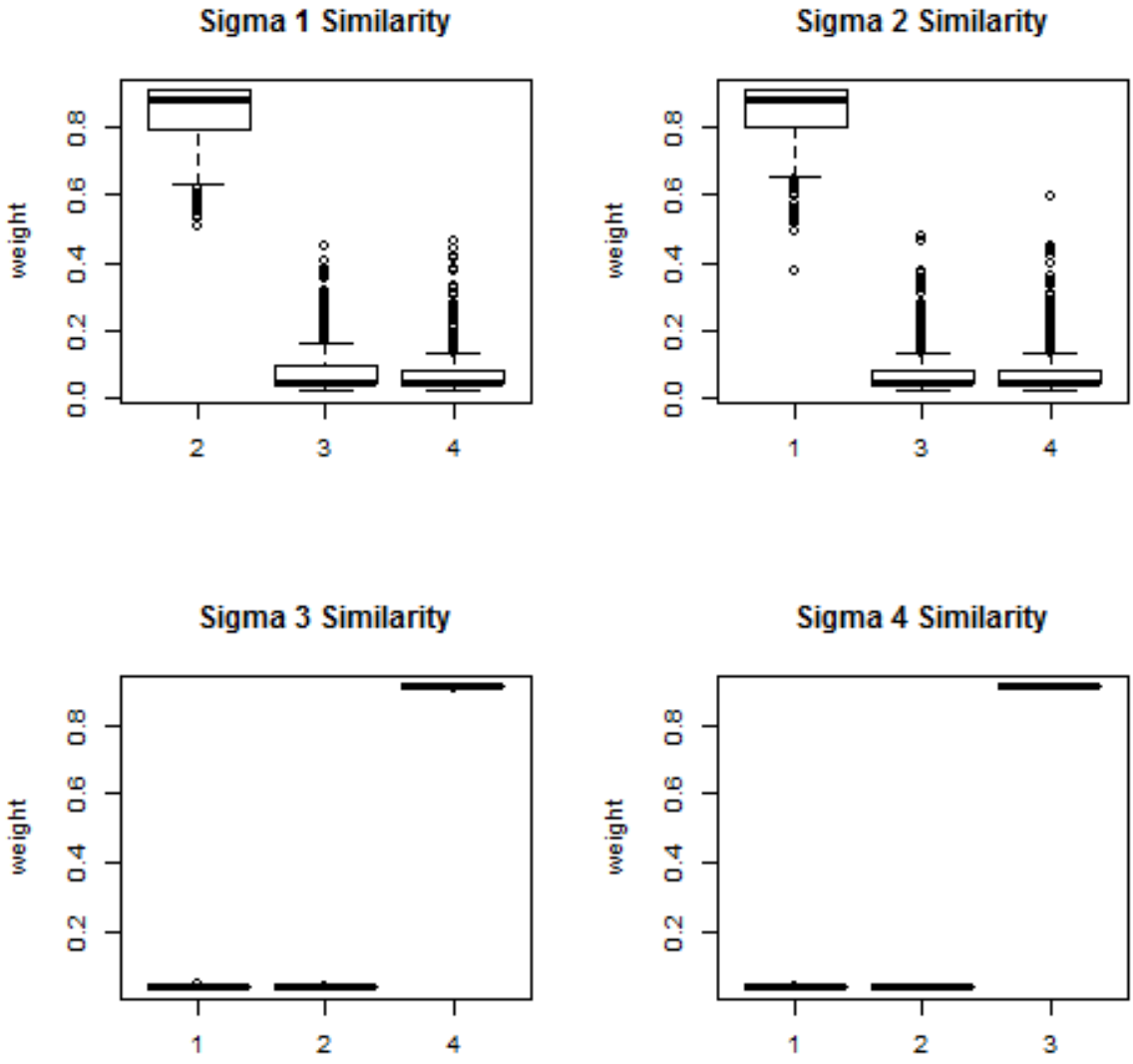

Figure B.9. Case T-2: KL Similarity Performance

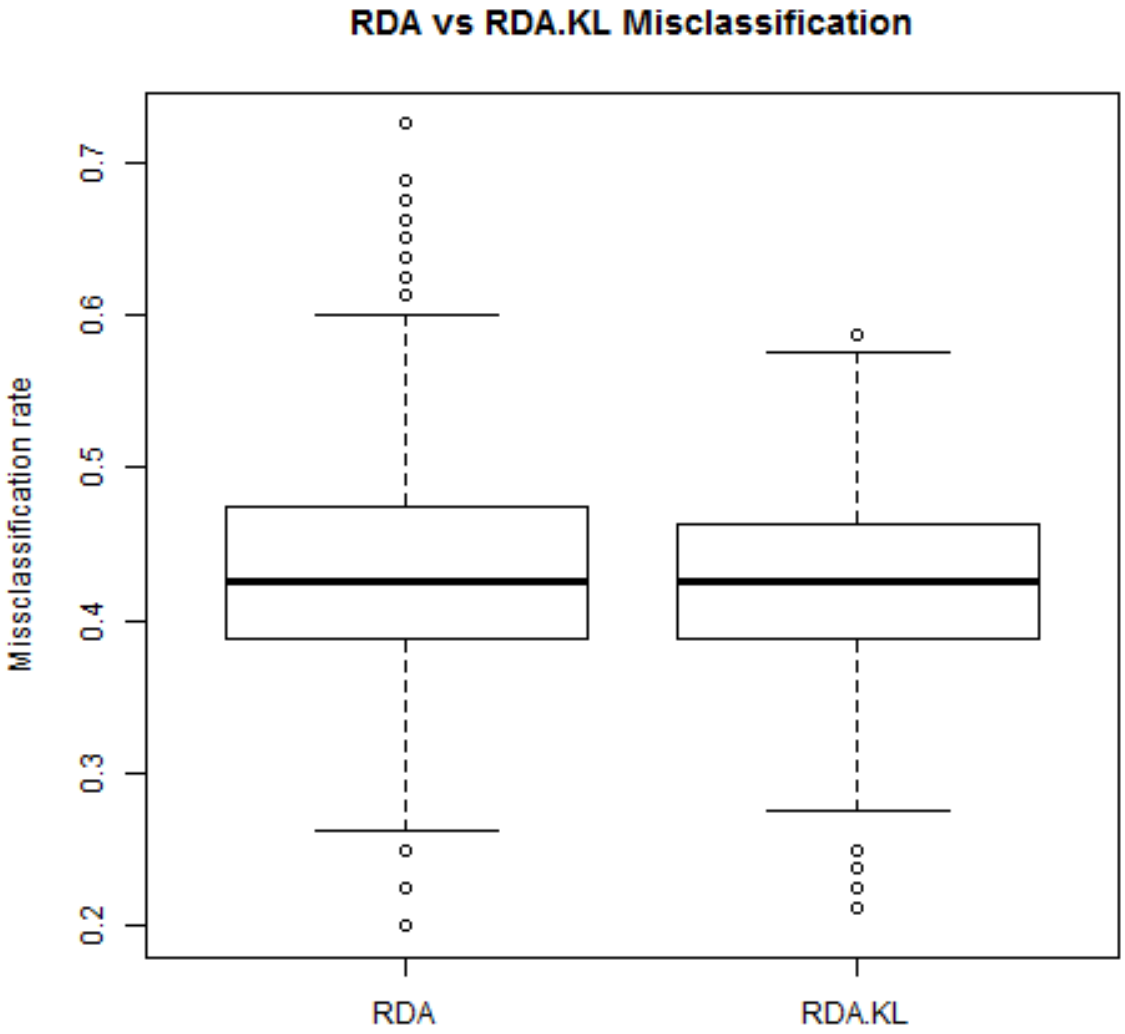

Figure B.10. Case T-2: Misclassification Rate

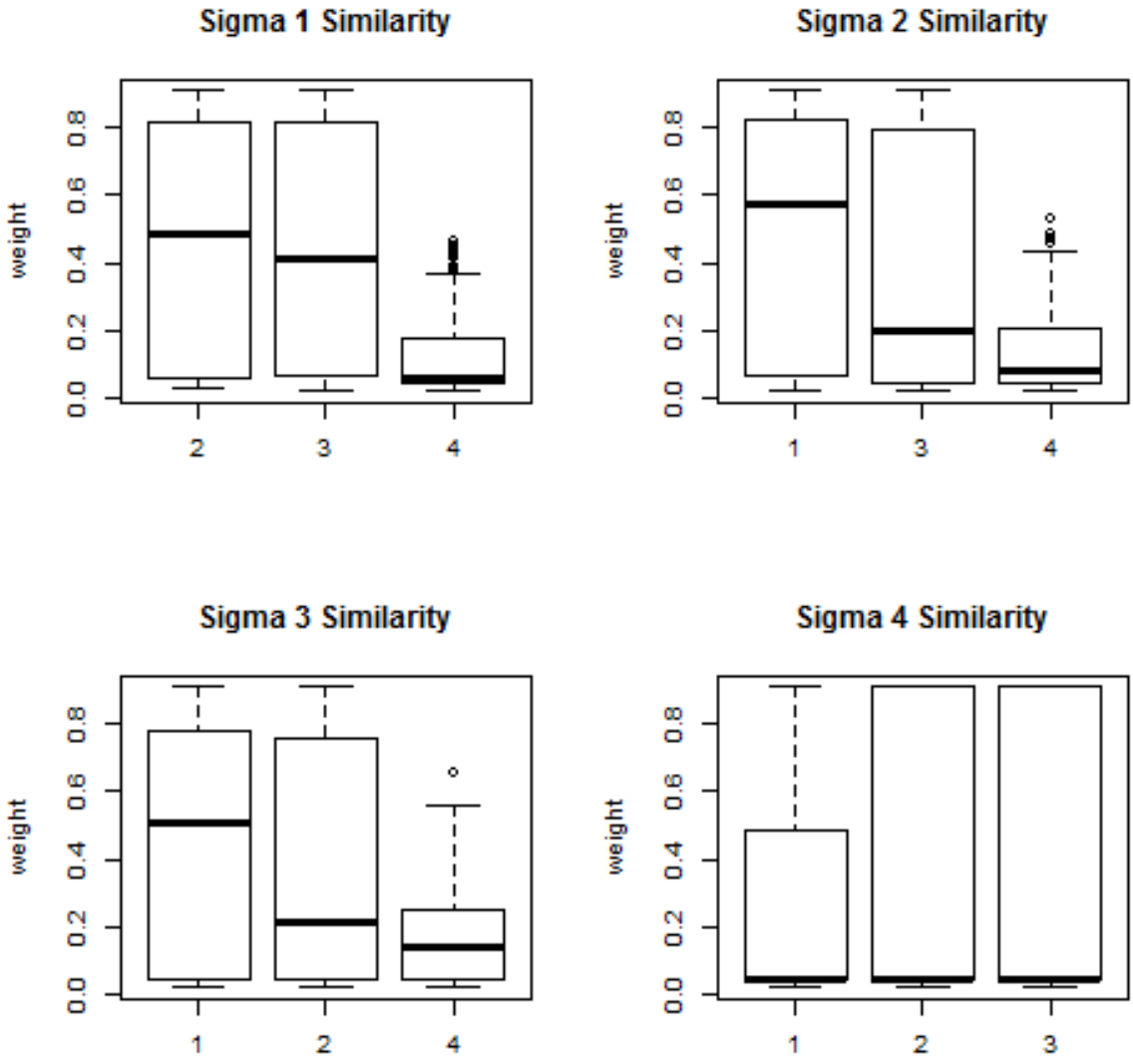

Figure B.11. Case T-3: KL Similarity Performance

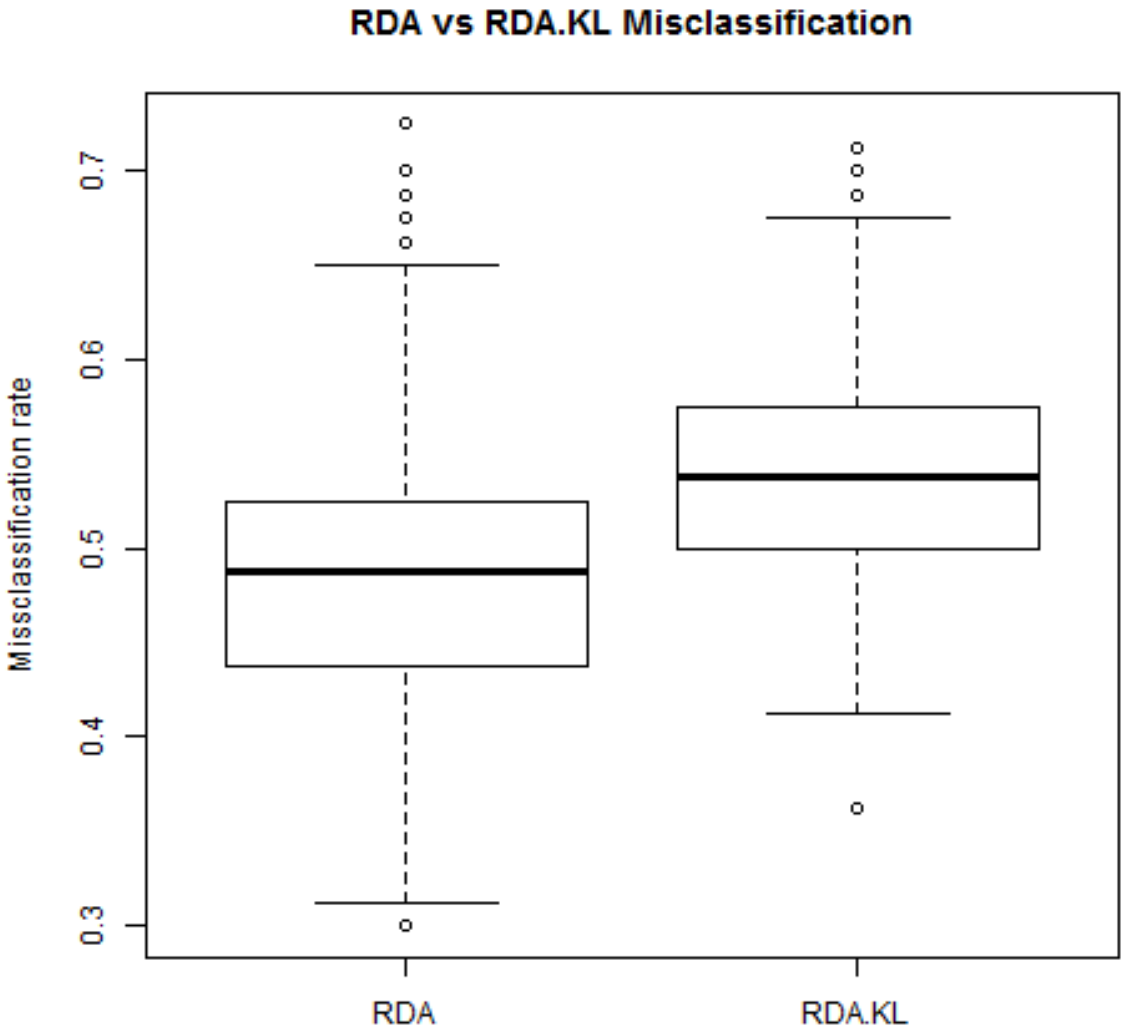

Figure B.12. Case T-3: Misclassification Rate

BIBLIOGRAPHY

Abdel-Aty, S. H. (1954), "Approximate formulae for the percentage points and probability integral of the non central χ^2 distribution," *Biometrika*, 41, 538–540.

Anderson, T. (1971), *The Statistical Analysis of Time Series*, Wiley Series in Probability and Statistics, Wiley.

Berg, A., Paparoditis, E., and Politis, D. N. (2010), "A bootstrap test for time series linearity," *Journal of Statistical Planning and Inference*, 140, 3841 – 3857.

Berg, A. and Politis, D. (2009), "Higher-order accurate polyspectral estimation with flat-top lag-windows," *Annals of the Institute of Statistical Mathematics*, 61, 477–498.

Bregman, L. (1967), "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," {*USSR*} *Computational Mathematics and Mathematical Physics*, 7, 200 – 217.

Brillinger, D. R. (1965), "An introduction to polyspectra," *Ann. Math. Statist.*, 36, 1351–1374.

— (1969), "Asymptotic properties of spectral estimates of second order," *Biometrika*, 56, 375–390.

Brillinger, D. R. and Rosenblatt, M. (1967), "Asymptotic theory of estimates of k-th order spectra," in *Spectral Analysis Time Series (Proc. Advanced Sem., Madison, Wis., 1966)*, John Wiley, New York, pp. 153–188.

Bühlmann, P. (1997), "Sieve bootstrap for time series," *Bernoulli*, 3, 123–148.

— (2002), "Bootstraps for time series," *Statist. Sci.*, 17, 52–72.

Cai, Z., Fan, J., and Yao, Q. (2000), "Functional-coefficient regression models for nonlinear time series," *J. Amer. Statist. Assoc.*, 95, 941–956.

Chan, K. and Tong, H. (1986), "A note on certain integral equations associated with nonlinear timne series analysis," *Probability Theory and Related Fields*, 73, 153–159.

Chang, K. and Tong, H. (1986), "A note on certain integral equations associated with nonlinear time series analysis," *Probability Theory and Related Fields*, 73, 153–159.

Chen, R. and Liu, L. (1993), "Functional coefficient autoregressive models," *Journal of the American Statistical Association*, 421, 298–308.

D'Agostino, R. B. and Stephens, M. A. (eds.) (1986), *Goodness-of-fit techniques*, New York, NY, USA: Marcel Dekker, Inc.

Efron, B. and Tibshirani, R. J. (1993), *An introduction to the bootstrap*, vol. 57 of *Monographs on Statistics and Applied Probability*, New York: Chapman and Hall.

Eguchi, S. and Copas, J. (2006), "Interpreting KullbackLeibler divergence with the NeymanPearson lemma," *Journal of Multivariate Analysis*, 97, 2034 – 2040, ¡ce:title¿Special Issue dedicated to Prof. Fujikoshi¡/ce:title¿.

Friedman, J. H. (1989), "Regularized Discriminant Analysis," *Journal of the American Statistical Association*, 84, pp. 165–175.

Haggan, V. and Ozaki, T. (1980), "Amplitude-dependent exponential autoregressive model fitting for nonlinear random vibrations," Amsterdam: North-Holland, pp. 57–71.

Harvill, J. L. (1999), "Testing time series linearity via goodness-of-fit methods," *Journal of Statistical Planning and Inference*, 75, 331–341.

Harvill, J. L. and Newton, H. J. (1995), "Saddlepoint approximations for the difference of order statistics," *Biometrika*, 82, 226–231.

Harvill, J. L. and Ray, B. K. (2005), "A note on multi-step forecasting with functional coefficient autoregressive models," *International Journal of Forecasting*, 21, 717–727.

— (2006), "Functional coefficient autoregressive models for vector time series," *Comput. Statist. Data Anal.*, 50, 3547–3566.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The elements of statistical learning*, Springer Series in Statistics, New York: Springer, 2nd ed., data mining, inference, and prediction.

Hinich, M. J. (1982), "Testing for Gaussianity and linearity of a stationary time series," *J. Time Ser. Anal.*, 3, 169–176.

Hinich, M. J., Mendes, E. M., and Stone, L. (2005), "Detecting Nonlinearity in Time Series: Surrogate and Bootstrap Approaches." *Studies in Nonlinear Dynamics and Econometrics*, 9, 1 – 13.

Jahan, N. and Harvill, J. L. (2008), "Bispectral-based goodness-of-fit tests of Gaussianity and linearity of stationary time series," *Comm. Statist. Theory Methods*, 37, 3216–3227.

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

Künsch, H. R. (1989), "The jackknife and the bootstrap for general stationary observations," *Ann. Statist.*, 17, 1217–1241.

Kuo, B.-C. and Landgrebe, D. (2002), "A covariance estimator for small sample size classification problems and its application to feature extraction," *Geoscience and Remote Sensing, IEEE Transactions on*, 40, 814 –819.

Parzen, E. (1957), "On consistent estimates of the spectrum of a stationary time series," *Ann. Math. Statist.*, 28, 329–348.

— (1961a), "An appproach to time series analysis," *Ann. Math. Statist.*, 32, 951–989.

— (1961b), "Mathematical considerations in the estimation of spectra," *Technometrics*, 3, 167–190.

Priestley, M. B. (1981), *Spectral Analysis and Time Series*, vol. 1 and 2, London: Academic Press.

Robinson, J. A. (2009), "Covariance estimation in full- and reduced-dimensionality image classification," *Image and Vision Computing*, 27, 1062 – 1071.

Rosenblatt, M. and Ness, v. (1965), "Estimation of the Bispectrum," *Ann. Math. Stat.*, 420–436.

Saxena, K. M. L. and Alam, K. (1982), "Estimation of the noncentrality parameter of a chi squared distribution," *Ann. Statist.*, 10, 1012–1016.

Schumway, R. H. and Stoffer, D. S. (2011), *Time Series Analysis and Its Applications: with R Examples*, New York: Springer, 3rd ed.

Subba Rao, T. and Gabr, M. M. (1980), *An Introduction to Bispectral Analysis and Bilinear Time Series*, New York: Springer-Verlag.

Subba Rao, T. and Gabr, M. M. (1980), "A test for linearity of stationary time series," *J. Time Ser. Anal.*, 1, 145–158.

Tong, H. (1983), *Threshold Models in Nonlinear Time Series Analysis: Lecture Notes in Statistics*, New York: Springer.

— (1990), *Nonlinear Time Series: A Dynamical Systems Approach*, UK: Oxford University Press.

— (1993), *Non-linear Time Series: A Dynamical Systems Approach*, Non-Linear Time Series, Oxford University Press on Demand.

Van Ness, J. W. (1966), "Asympotic normality of bispectral estimates," *Ann. Math. Statist.*, 37, 1257–1272.

Vemuri, B., Liu, M., Amari, S.-I., and Nielsen, F. (2011), "Total Bregman Divergence and Its Applications to DTI Analysis," *Medical Imaging, IEEE Transactions on*, 30, 475–483.