2021

Bickel-Rosenblatt Test Based on Tilted Estimation for Autoregressive Models & Deep Merged Survival Analysis on Cancer Study Using Multiple Types of Bioinformatic Data

Yan Su
Wright State University

Follow this and additional works at: https://corescholar.libraries.wright.edu/etd_all

Part of the Physical Sciences and Mathematics Commons

Repository Citation
Su, Yan, "Bickel-Rosenblatt Test Based on Tilted Estimation for Autoregressive Models & Deep Merged Survival Analysis on Cancer Study Using Multiple Types of Bioinformatic Data" (2021). Browse all Theses and Dissertations. 2624.
https://corescholar.libraries.wright.edu/etd_all/2624

This Dissertation is brought to you for free and open access by the Theses and Dissertations at CORE Scholar. It has been accepted for inclusion in Browse all Theses and Dissertations by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.
BICKEL-ROSENBLATT TEST BASED ON TILTED ESTIMATION FOR AUTOREGRESSIVE MODELS & DEEP MERGED SURVIVAL ANALYSIS ON CANCER STUDY USING MULTIPLE TYPES OF BIOINFORMATIC DATA

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

by

YAN SU

M.S., Wright State University, 2017
M.B.A., Wright State University, 2012
B.Mgt., Dalian Jiaotong University, People's Republic of China, 2010

2021
Wright State University

______________________________
Shuxia Sun, Ph.D.
Dissertation Co-Director

______________________________
Zheng Xu, Ph.D.
Dissertation Co-Director

______________________________
Ivan Medvedev, Ph.D.
Program Director, Interdisciplinary Applied Science and Mathematics Program

______________________________
Barry Milligan, Ph.D.
Vice Provost for Academic Affairs
Dean of the Graduate School

Committee on Final Examination

______________________________
Weizhen Wang, Ph.D.

______________________________
Yang Liu, Ph.D.

______________________________
Joseph Houpt, Ph.D.
This dissertation includes two topics, Bickel-Rosenblatt test based on tilted density estimation for autoregressive models and deep merged survival analysis on cancer study using multiple types of bioinformatic data. In the first topic study, we consider the goodness of fit test the error density of linear and nonlinear autoregressive models using tilted kernel density estimation based on residuals. Bickel-Rosenblatt test statistic is based on the integrated square error of non-parametric error density estimation and a smoothed version of the parametric fit of the density. It is shown that the new type of Bickel-Rosenblatt test statistics behaves asymptotically the same as the one with conventional estimators based on true unobservable errors. We show technique details, simulation studies and real data analysis to present the performance of the new test statistic.

The second topic is about deep survival analysis on cancer study. For cancer survival prediction, we propose to use deep merged survival networks with network layers at high levels merged together for better integration of information from multiple heterogeneous data sets to improve prediction accuracy. We conducted simulation studies to compare our proposed method with other methods in the literature under a range of scenarios. We conducted real data analysis based on breast cancer TCGA data to illustrates the advantage of our method over literature methods, and the advantage of using multiple data sets over using only one data set.

The dissertation is organized as follows. The topic of Bickel-Rosenblatt test for autoregressive models will be presented in two sub-topics, linear and nonlinear autoregressive models. Each topic will be presented in one chapter. In Chapter 1 and 2, we delve deeper into the goodness of fit test of errors in autoregressive time series mod-
els using tilted density estimation. Chapter 3 will explore a new deep survival analysis method on cancer study. For each topic, we first introduce the topic with the literature review, next outline our contributions, such as main results or methods and simulation studies and last discuss further work. The reference list will be provided at the end.
# Contents

1 Bickel-Rosenblatt Test Based on Tilted Density Estimation for Linear Autoregressive Models 1
   1.1 Introduction ................................. 1
      1.1.1 Motivation .................................. 2
      1.1.2 Literature Review .......................... 3
   1.2 Bickel-Rosenblatt Test Statistic .................. 5
   1.3 Tilted Density Estimators ........................ 7
      1.3.1 Main Result ................................ 9
   1.4 Goodness-of-fit Tests in Time Series Models ........ 12
      1.4.1 Error Density Estimation in Autoregressive Models .... 12
      1.4.2 Main Result ................................ 15
   1.5 Simulation Study ................................ 16
      1.5.1 The Goodness-of-fit Test in IID Cases .......... 17
      1.5.2 The Goodness-of-fit Test in IID Cases for $m \geq 4$ .... 19
      1.5.3 The Goodness-of-fit Test on the Error Density of AR(1) Model .. 23
      1.5.4 A Real Data Example ........................ 26
   1.6 Technical Details ............................... 31
      1.6.1 Proof of Theorem 1 .......................... 39
      1.6.2 Proof of Theorem 2 .......................... 40

2 Bickel-Rosenblatt Test Based on Tilted Density Estimation for Nonlinear Autoregressive Models 43
   2.1 Introduction .................................... 43
      2.1.1 Motivation .................................. 44
      2.1.2 Literature Review .......................... 44
   2.2 Statement of The Problem ......................... 46
   2.3 Main Result .................................... 49
   2.4 Empirical Studies ............................... 51
      2.4.1 Simulation Studies .......................... 52
      2.4.2 A Real Data Example ........................ 55
   2.5 Technical Details ............................... 60
      2.5.1 Auxiliary Results ........................... 61
3 Deep Merged Survival Analysis on Cancer Study Using Multiple Types of Bioinformatic Data

3.1 Introduction ......................................................... 73
  3.1.1 Literature Review ........................................... 76

3.2 Materials and Methods ........................................... 79
  3.2.1 Architecture of Deep Merged Survival Neural Networks ..... 80
  3.2.2 Objective Function .......................................... 82
  3.2.3 Evaluation Metrics .......................................... 82

3.3 Simulation Studies ............................................... 83
  3.3.1 Four Simulation Scenarios .................................. 84
  3.3.2 Cancer Survival Prediction Methods ......................... 86
  3.3.3 Simulation Results .......................................... 88

3.4 Real Data Analysis ............................................... 97
  3.4.1 TCGA BRCA Data Sets ...................................... 97
  3.4.2 Data Preprocessing ......................................... 98
  3.4.3 Model Construction ......................................... 101
  3.4.4 Real Data Application Results ............................ 103

3.5 Discussion ......................................................... 109

Bibliography .......................................................... 111
## List of Figures

1.1 Density Function Estimators with IID Random Variables, $n = 100$. .......... 17  
1.2 Compare Power Between BR Test Based on the Conventional Density Estimator and BR Test Based on the Tilted Density Estimator. ...................... 19  
1.3 Scenario 1: Density Function Estimators with Sample Size $n = 100$. .......... 21  
1.4 Scenario 2: Density Function Estimators with Sample Size $n = 500$. ........... 22  
1.5 Error Density Function Estimators, $n = 100$. ................................. 24  
1.6 Compare Power Between BR Test Based on the Conventional Density Estimator and BR Test Based on the Tilted Density Estimator for AR(1) Model with $\phi = 0.5$. .................................................. 26  
1.7 Time Series Plot for Earthquake Data .............................................. 27  
1.8 PACF of Earthquake Data. ............................................................. 28  
1.9 Histogram of Residuals for Earthquake Data ................................. 29  
1.10 Error Density Estimators for Earthquake Data ............................... 30  
2.1 Power Comparison Between BR tests with Conventional Density and Tilted Density Estimators ................................................................. 54  
2.2 Density Function Estimators with Sample Size $n = 100$ ........................ 55  
2.3 Time Series Plot for INDPRO Data .................................................. 56  
2.4 Histogram of Residuals for INDPRO Data ........................................ 59  
2.5 Error Density Estimators for INDPRO Data ....................................... 60  
3.1 Architecture of General Multi-bioinformatic Survival Analysis Network. ... 81  
3.2 Simulation Study: The Architecture of Deep Merged Survival Neural Network for the Simulated Data Survival Analysis .............................. 87  
3.3 Plots of Different Combinations of Sample Size $n$ and Feature Size $p$ Based on Data Simulated Under Scenario 1 (Sparse Causal Predictors Scenario) ....... 90  
3.4 Plots of Different Combinations of Sample Size $n$ and Feature Size $p$ Based on Data Simulated Under Scenario 2 (Predetermined Neural Network Scenario) ................................. 92  
3.5 Plots of Different Combinations of Sample Size $n$ and Feature Size $p$ Based on Data Simulated Under Scenario 3 (Auto-encoder Scenario) ............ 94  
3.6 Plots of Different Combinations of Sample Size $n$ and Feature Size $p$ Based on Data Simulated Under Scenario 4 (Principal Component Scenario) ... 96
3.7 Number of Patients in Different Bioinformatics Data Sets ............... 101
3.8 Architecture of Deep Merged Survival Neural Network on Breast Cancer Survival Analysis ......................................................... 103
3.9 Model Performance Using Different Combinations of Bioinformatics Data 106
3.10 Model Performance Using Different Combinations of Bioinformatics Data. The evaluation metric is the C-index. ......................... 107
3.11 Model Performances Using Different Methods ............................ 108
3.12 Performances of Different methods Using Different Combinations of Bioinformatics Data ......................................................... 109
## List of Tables

1.1 iid case: m=2, Empirical Size and Empirical Power .......................... 18  
1.2 iid case: m=3, Empirical Size and Empirical Power .......................... 18  
1.3 iid case: m = 4, 5, 6, Empirical Size under $H_0$ ........................... 20  
1.4 Scenario 1: ISE with Sample Size $n = 100$. ................................. 21  
1.5 Scenario 2: ISE with Sample Size $n = 500$ ................................. 22  
1.6 Empirical Sizes under $H_0$ with m=2 ....................................... 25  
1.7 Empirical Power under $H_1$ with m=2 ....................................... 25  
1.8 Empirical Size under $H_0$ with m=3 ....................................... 25  
1.9 Empirical Power under $H_1$ with m=3 ....................................... 26  
1.10 Earthquake Data: ISE Based on Different m .............................. 29  
2.1 m=2, Empirical Size and Empirical Power ................................. 53  
2.2 m=3, Empirical Size and Empirical Power ................................. 54  
2.3 ISE with Sample Size $n = 100$ .............................................. 55  
2.4 INDPRO Data: AIC Based on AR(p) ........................................ 57  
2.5 INDPRO Data: ISE Based on Different m .................................... 60  
3.1 Details of Architecture of Neural Network Models .......................... 86  
3.2 Average Concordance Score for Different Methods Based on Data Simu- 
     lated Under Scenario 1 (Sparse Causal Predictors Scenario). Sample size 
     is $n$ and feature size is $p$. ............................................. 89  
3.3 Average Concordance Score for Different Methods Based on Data Simu- 
     lated Under Scenario 2 (Predetermined Neural Network Scenario). Sample 
     size is $n$ and feature size is $p$. ........................................ 91  
3.4 Average Concordance Score for Different Methods Based on Data Simu- 
     lated Under Scenario 3 (Auto-encoder Scenario). Sample size is $n$ and fea-
     ture size is $p$. ......................................................... 93  
3.5 Average Concordance Score for Different Methods Based on Data Simu- 
     lated Under Scenario 4 (Principal Component Scenario). Sample size is $n$
     and feature size is $p$. ................................................ 95  
3.6 Number of Variables Before and After Feature Filtering ............... 100  
3.7 Model Performance Using Different Combinations of Bioinformatics Data 
     Sets. The evaluation metric is the C-index. ............................ 104
Acknowledgment

I am deeply grateful for everyone who helped me along this journey. I especially want to thank you my advisors Dr. Sun and Dr. Xu. They have given me great instructions and encouragement throughout the study. I would like to thank you my dissertation committee members, Dr. Wang, Dr. Liu and Dr. Houpt for their time and support. I would like to thank Dr. Cheng for the helpful advice on the study of Bickel-Rosenblatt test for linear autoregressive models. I also would like to thank Dr. Sharma for helping me install packages and solve issues when I use HPC cluster. Lastly, I would like to thank my family for their support through my Ph.D. study.
Bickel-Rosenblatt Test Based on Tilted Density Estimation for Linear Autoregressive Models

1.1 Introduction

The goodness-of-fit tests of the error distribution in time series models is very important and the empirical process method plays an essential role in the goodness of fit techniques. In order to overcome the limitation of empirical process method, many researchers considered Bickel-Rosenblatt test instead of tests using empirical process method. The advantage of Bickel-Rosenblatt test lies in the fact that it has asymptotic normality and performs better than the Cramer-von Mises test and Kolmogorov-Smirnov test (based on the empirical process method) in detecting sharp peak alternatives.

This study considers goodness-of-fit tests based on tilted density estimators. Motivated by recent developments in the area of data perturbation, we introduce a new type of Bickel-Rosenblatt test statistic by applying tilted kernel density estimators. In this Chapter, we first investigate the asymptotic behaviors of our Bickel-Rosenblatt type test statistic on density functions in the setup of independently identically distributed (iid) cases and show that, under both the null hypothesis and fixed alternatives, the test
statistic, based on the tilted density estimator, has the same limiting distributions as the one based on conventional kernel density estimators. Next, we extend both theoretical results to the case of goodness-of-fit tests on error distributions in linear autoregressive models and show that the Bickel-Rosenblatt test statistic based on tilted density function estimators from residuals has the same asymptotic normal distributions as the one with conventional estimators based on true unobservable errors. Simulation studies and a real-world example are presented to explore the empirical performances of this new test statistic. From simulation studies, we also find that the new Bickel-Rosenblatt test based on the tilted density estimator has higher power than the test based on the conventional estimator.

The remaining of this chapter is organized as follows. In Section 1.1, we present the motivation and the literature review of the topic. In Section 1.2, we introduce the goodness-of-fit test of density functions and restate the limiting distributions of the Bickel-Rosenblatt test statistic. In Section 1.3, we define the new high-order tilted kernel density estimator proposed by Doosti and Hall [1] and Doosti et al. [2] before investigating the limiting distribution of its corresponding Bickel-Rosenblatt test statistic. The goodness-of-fit test based on the tilted density estimator in linear autoregressive models is discussed in Section 1.4. We then present simulation studies in Section 1.5 and provide technical details of the main results in Section 1.6.

1.1.1 Motivation

Goodness-of-fit tests are often used in time series models to determine whether the errors follow a certain distribution. In autoregressive time series models, goodness-of-fit tests based on residual empirical processes have been extensively studied by researchers, including Boldin [3], Pierce [4], and Koul ([5], [6]). Although the empirical process method (e.g. the Cramer-von Mises test, and the Kolmogorov-Smirnov test) is a common technique for goodness-of-fit test problems, parameter estimators may affect
the limiting process, making it difficult to calculate critical values of a test (cf. Durbin [7]). Bickel and Rosenblatt [8] introduced a goodness-of-fit test based on the integrated squared error of the kernel density estimator and showed that its test statistic is asymptotically normally distributed. The limiting distribution of Bickel-Rosenblatt test statistic is not affected by any estimators of nuisance parameters. Later, Lee and Na [9] and Cheng and Sun [10] extended this test to the goodness of fit on error distribution for linear autoregressive models, and nonlinear autoregressive models, respectively. They showed that the normal limiting distribution of the test statistic does not depend on unknown parameters in the model, which makes it fairly easy to find the critical values. Doosti and Hall [1] and Doosti et al. [2] introduced tilted kernel density estimators and proved that this new type of estimator outperforms the conventional kernel density estimator.

In this work, we propose goodness-of-fit tests with the iid setup for density functions and with the setup for error distribution of linear autoregressive models by constructing a Bickel-Rosenblatt type test statistic based on the tilted kernel density estimator. Our simulation results indicate that the new Bickel-Rosenblatt test based on the tilted density estimator outperforms the test based on the conventional estimator.

1.1.2 Literature Review

In autoregressive time series models, the goodness of fit test based on the residual empirical process has been extensively studied by many researchers. Boldin [3] first proposed the goodness of fit hypothesis test using the residual in the empirical process method for errors in stationary autoregressive models. The shortcoming of the empirical process method is that the parameter estimators could affect the limiting process, and the limiting process is not a standard Brownian bridge (cf. Durbin [7]). Hence, in practice, it is hard to calculate the critical values of a test based on the empirical process. Bickel and Rosenblatt [8] proposed a test based on the integrated squared error between
The non-parametric kernel density estimator and true density. The advantage of the Bickel-Rosenblatt test statistics is that its limiting distribution does not depend on any estimators of nuisance parameter. In detecting sharp peak alternative (cf. Ghosh and Huang [11]), Bickel-Rosenblatt test has better performance than Cramer-von Mises test and Kolmogorov-Smirnov test, which are common tests of empirical process method. Takahata and Yoshihara [12] extended the Bickel-Rosenblatt test to dependent processes and they showed that the limiting distribution of the test statistic is the same as in i.i.d. case. However, Bickel-Rosenblatt test loses powers when the observations are highly correlated. Lee and Na [9] investigated Bickel-Rosenblatt test on error density of linear autoregressive time series models and established the asymptotic normality for the test statistic under the null hypothesis \( H_0 : f = f_0 \). Bachmann and Dette [13] presented an alternative proof of a Bickel-Rosenblatt result under the null hypothesis with substantially weaker assumptions than those imposed by Bickel and Rosenblatt[8] and established the limiting distribution of the test statistic under fixed alternative for the goodness-of-fit test on errors in linear autoregressive models. Cheng and Sun [10] extended the results of Lee and Na [9] and Bachmann and Dette [13] to nonlinear autoregressive models.

In terms of the kernel density estimation, it is well known that under the conditions listed in Section 2.2, the optimal convergence rate of the conventional kernel density estimator defined in Equation (2.1) is of order \( O_p(n^{-4/5}) \). This rate of convergence cannot be improved further unless we apply high-order kernel methods by allowing the kernel function \( K \) to take some negative values—in which case it is no longer a valid density function—and requiring that \( f \) be many times continuously differentiable. However, high-order kernel methods may produce density estimators that assume negative values and suffer from spurious oscillations. Some methods have been proposed to correct the negativity of high-order kernel density estimators. See Hall and Murison [14], among others. The tilting-based approach in density function estimation was first pro-
posed by Grenander [15] for enforcing constraints. Owen ([16], [17], [18]) introduced the empirical likelihood method to find probability weights under a set of constraints. See Chen [19], Zhang [20], Müller [21], Schick et al. [22] and Chiang and Tan [23] for more related work. In addition to the empirical likelihood method, other distance measures have been proposed to find the probability weights. We refer to Hall and Presnelt [24], Hall and Huang [25], and Carroll et al. [26] for some details. In order to overcome the shortcomings of high-order density estimators, Doosti and Hall [1] proposed new non-parametric high-order density estimators based on data perturbation by use of tilting or data sharpening. They argued that these new estimators are guaranteed to be positive and more accurate than conventional kernel density estimators. In a following work, Doosti et al. [2] further investigated the tilted nonparametric kernel density estimator and showed that it has a faster convergence rate than a conventional kernel density estimator with a positive kernel.

Our contribution in this work is to construct the Bickel-Rosenblatt test based on tilted error density estimator in linear autoregressive models and show that the test statistic is asymptotically normal distributed. We first derive the limiting distributions of the tilted version of the Bickel-Rosenblatt test statistic (cf. Theorem 1). Secondly, we explore the goodness-of-fit test (about errors in linear autoregressive models) based on tilted density estimators and show that the corresponding Bickel-Rosenblatt test statistic has the same limiting distributions as the goodness-of-fit test statistic based on conventional kernel density estimators (cf. Theorem 2).

### 1.2 Bickel-Rosenblatt Test Statistic

For the sake of completeness, we first revisit the Bickel-Rosenblatt test statistic for the goodness-of-fit test. Let $X_1, X_2, ..., X_n$ denote $n$ iid random variables with unknown density function $f$. The conventional kernel density estimator for $f$ with kernel function $K$
and bandwidth $h$ is thus defined as

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right). \quad (1.1)$$

Consider the following goodness-of-fit test

$$H_0 : f = f_0, \quad \text{vs.} \quad H_1 : f \neq f_0,$$

where $f_0$ is a prescribed probability density function. The Bickel-Rosenblatt test statistic, $T_n$, is defined as

$$T_n = nh^{1/2} \int (f_n(x) - (K_h * f_0)(x))^2 \, dx, \quad (1.2)$$

where $(K_h * f_0)(x) = \int \frac{1}{h} K\left(\frac{x-u}{h}\right) f_0(u) \, du$.

Here, $T_n$ is based on the integrated squared error of the conventional kernel density estimator. In their seminal paper, Bickel and Rosenblatt [8] investigated the global performance of the kernel density estimator and showed that under $H_0$, the test statistic $T_n$ is asymptotically normally distributed. More specifically, if the density function $f$ is twice continuously differentiable with a bounded second derivative, if the bandwidth $h$ satisfies $h \to 0$ and $nh^4 \to \infty$, and if the kernel function $K$ (which is also a density function) is continuous with compact support and symmetric about 0, the Bickel-Rosenblatt test statistic has the following asymptotic properties:

(i) Under the null hypothesis $H_0 : f = f_0$, as $n \to \infty$,

$$T_n - h^{-\frac{1}{2}} \int K^2(t) \, dt \frac{d}{d} N(0, \tau^2), \quad (1.3)$$

where $\tau^2 = 2 \int f_0^2(x) \, dx \int (\int K(t)K(s-t) \, dt)^2 \, ds$. 
(ii) Under the alternative $H_1 : f \neq f_0$, as $n \rightarrow \infty$,

$$(nh)^{-\frac{1}{2}} T_n - n^{\frac{1}{2}} \int (K_h \ast (f - f_0))^2(x) d t \overset{d}{\rightarrow} N(0, 4\sigma^2),$$

where $\sigma^2 = Var((f - f_0)(X_1))$.

Note that the first result under $H_0$ was established by Bickel and Rosenblatt [8] (cf. Theorem 4.1) and that the second under $H_1$ was done by Bachmann and Dette [13] (cf. Theorem 3.1). Property (i) can be used to find the rejection region for the test and Property (ii) allows us to calculate the power of the goodness-of-fit test. Our main contribution in this work is to extend the above results to cases with tilted kernel density estimation for both testing density functions in iid setups and for testing error distributions in linear autoregressive models.

### 1.3 Tilted Density Estimators

We define the tilted kernel density estimator with the same setup as the conventional kernel density estimator in Equation (2.1):

$$f_n^*(x|h,p) = \frac{1}{h} \sum_{i=1}^{n} p_i K\left(\frac{x - X_i}{h}\right),$$

where $p = (p_1, \cdots, p_n)$ is a vector of probability weights satisfying $\sum_{i=1}^{n} p_i = 1$. The conventional kernel density estimator $f_n$ is a special case of $f_n^*$ with $p_1 = \cdots = p_n = 1/n$. Next, we define the integrated squared error of $f_n^*$ as

$$ISE(h, p) = \int \left(f_n^*(x|h,p) - f(x)\right)^2 dx = \int f_n^*(x|h,p)^2 dx - 2 \int f_n^*(x|h,p) f(x) dx + \int f(x)^2 dx.$$
Ideally, we want to choose tuning parameters $h$ and $p$ so that $ISE(h, p)$ is minimized. Doosti et al. [2] proposed the following cross-validation criterion

$$CV(h, p) = \int f^*(x|h, p)^2 dx - \frac{2}{n} \sum_{i=1}^{n} f^*_{-i}(X_i|h, p),$$

(1.7)

where $f^*_{-i}$ is the leave-one-out version of $f^*_n$ by deleting $X_i$ from the original data, i.e.

$$f^*_{-i}(x|h, p) = \frac{1}{h} \sum_{j: j \neq i} p_j K\left(\frac{x - X_j}{h}\right).$$

The authors showed that under appropriate conditions,

$$\sup_{p \in P} \left| \frac{1}{n} \sum_{i=1}^{n} f^*_{-i}(X_i|h, p) - \int f^*(x|h, p)f(x)dx - Q_n \right| = O_p((nh^{-1} + h^4)n^{-c}) = o_p(1) \quad (1.8)$$

uniformly in $h \in [n^{-1/5-c_1}, n^{-1/5+c_1}]$ and in $p \in P$, where $P$ is a class of probability measures satisfying some appropriate conditions, $c$ and $c_1$ are positive constants, and $m = n^{c}$ with $c \in (0, 1/2]$. Here, the quantity $Q_n$ does not depend on $h$ or $p$. Results (1.7) and (1.8) imply that one can choose $p$ and $h$ so that $ISE(h, p) = o_p(n^{-4/5})$, while the rate of convergence of conventional kernel estimators with positive kernels $K$ is $O_p(n^{-4/5})$. Thus, tilted kernel density estimators in general outperform conventional kernel density estimators.

It is worth noting that these $n$ probability weights may only take $m < n$ distinct values, say $q_1, \cdots, q_m$. Doosti et al. [2] proposed a sparse interpolation algorithm for choosing $h$ and $p$ by minimizing $CV(h, p)$: we pick $i_1 < \cdots < i_m$ between 1 and $n$ so that $C_1 n/m \leq i_j - i_{j-1} \leq C_2 n/m$ for constants $C_1, C_2 > 0$ and set $p_i = q_j$ for all $X(i)$ with $i_{j-1} < i \leq i_j$. Here $X(i)$ is the $i$-th order statistic of $X_1, \cdots, X_n$. As such, the tuning parameters under this algorithm are $h, q_1, \cdots, q_m$. We refer to Doosti et al. [2] for more details about the algorithm.
Motivated by the better performance of tilted density estimators, we would like to consider using this type of estimator to construct the goodness-of-fit test on density functions. One of our contributions in this work is to establish the limiting distribution for the Bickel-Rosenblatt type test statistic based on the tilted kernel density estimator in Equation (1.5) to test

$$H_0 : f = f_0, \quad \text{vs.} \quad H_1 : f \neq f_0,$$

where $f$ is the unknown marginal density function for random sample $X_1, \cdots, X_n$, and $f_0$ is a prescribed density function.

### 1.3.1 Main Result

In this subsection, we begin by outlining some basic assumptions about the density $f$, kernel function $K$, bandwidth $h$, and probability weights $p$.

**A1.** $f$ is two-times continuously differentiable with bounded first and second derivatives, and $f^2$ is integrable.

**A2.** Conditions for the kernel function:

(K1) $K$ is continuous, symmetric about 0 with compact support, and

$$\int K(t) dt = 1.$$

(K2) $K$ is three-times differentiable, $K'''$ is bounded, and

$$\int |K^{(j)}(t)| dt < \infty, \quad j = 1, 2, 3, \quad \int |K^{(j)}|^2 dt < \infty, \quad j = 1, 2. \quad (1.10)$$

**A3.** $h \to 0$ and $nh^4 \to \infty$ as $n \to \infty$. 

9
A4. The probability weights $p_1, \ldots, p_n$ satisfy

$$\delta_n := \sum_{i=1}^{n} (p_i - n^{-1})^2 \leq C_0 n^{-1} h / \sqrt{\log n}, \quad \max\{p_1, \ldots, p_n\} \leq C / n,$$

for some positive constants $C_0, C$.

**Remark 1** We would like to point out that A1, (K1), and A3 are regular assumptions about the density function and the kernel in conventional kernel nonparametric density estimation and that the choice of the bandwidth $h = O(n^{-1/5})$ yields the best rate of convergence for $f_n$. Similar conditions are used in tilted kernel density estimation. For example, Doosti et al. [2] suggested that one select the bandwidth in interval $[n^{-1/5} - c_1, n^{-1/5} + c_1]$ for some small positive constant $c_1 \leq 1/30$. In the setup of conventional error density estimation, an additional condition (K2) is needed, thus, $K$ satisfies condition A2. Similar conditions are used by Lee and Na [9], Bachmann and Dette [13], and Cheng and Sun [10].

**Remark 2** In addition to conditions A1-A3, data-tilting introduces probability weights $p_1, \ldots, p_n$ as more tuning parameters. Our assumption (P), $\max_{1 \leq i \leq n} |p_i - n^{-1}| \leq C_0 n^{-1} h / \sqrt{\log n}$, is weaker than Doosti and Hall [1] but stronger than Doosti et al. [2]. Doosti and Hall [1] used condition $p_i = n^{-1} (1 + h^2 d(X_i))$. In a following work, Doosti et al. [2] proposed a sparse interpolation algorithm, which reduces the number of tuning parameters to $m + 1$ by imposing the condition that those probability weights have $m$ distinct values, say $q_1, \ldots, q_m$. They used assumption $\max_{1 \leq i \leq n} |p_i - n^{-1}| \leq h^2 n^{c_2} / \sqrt{\log n}$, for some
constant $c_2$ satisfying $2c_1 + c_2 < \frac{1}{10}$, which leads to

$$\max_{1 \leq i \leq n} |p_i - n^{-1}| \leq h^2 n^{c_2} < n^{-2/5+2c_1} \cdot n^{1/10-2c_1} \sqrt{\log n} = n^{-3/10} / \sqrt{\log n}.$$  

We are now in a position to state the first main result.

**Theorem 1**  If assumptions $A1$-$A4$ hold, the Bickel-Rosenblatt test statistic for the goodness-of-fit test on density by use of the tilted kernel density estimator

$$T_n^* = nh^{1/2} \int (f_n^*(x) - (K_h * f_0)(x))^2 dx,$$  

(1.11)

has the following asymptotic properties:

(i) Under the null hypothesis $H_0 : f = f_0$,

$$T_n^* - h^{-1/2} \int K^2(t) dt \overset{d}{\to} N(0, \tau^2),$$  

(1.12)

where $\tau^2 = 2 \int f_0^2(x) dx \int (\int K(t)K(s-t)dt)^2 ds$.

(ii) Under the alternative $H_1 : f \neq f_0$, as $n \to \infty$,

$$(nh)^{-1/2} T_n^* - n^{1} \int (K_h * (f - f_0))^2(x) dt \overset{d}{\to} N(0, 4\sigma^2),$$  

(1.13)

where $\sigma^2 = Var[(f - f_0)(X_1)]$.

Theorem 1 indicates that the goodness-of-fit test statistic based on the tilted kernel density estimator has the same limiting distributions as the one based on the conventional kernel density estimator. Theorem 1 (i) allows us to find the critical values for
rejecting $H_0$, and Theorem 1 (ii) makes it possible to calculate the power of the test under certain alternatives.

### 1.4 Goodness-of-fit Tests in Time Series Models

In this section, we focus on error density estimation of the first-order stationary autoregressive model $AR(1)$. The results can be easily extended to general stationary autoregressive models $AR(p)$. Our goal is to investigate asymptotic behaviors of the Bickel-Rosenblatt test statistic based on tilted error density estimators.

#### 1.4.1 Error Density Estimation in Autoregressive Models

Let $\{X_j\}, j = 1, 2, ...$ be a zero-mean random sequence from the following first autoregressive model,

$$X_j = \phi X_{j-1} + \epsilon_j, \quad j = 1, 2, ..., n,$$

(1.14)

where $\{\epsilon_j\}$ are iid random variables with an unknown common density $f$. The sequence is stationary if $|\phi| < 1$. We aim to conduct a goodness-of-fit test, such as the Gaussian, on $\epsilon_j$, with a prescribed density function $f_0$. Let $\hat{\phi}$ be an estimator of $\phi$, such that

$$\hat{\phi} - \phi = O_p(a_n),$$

where $a_n$ is a normalizing constant that is equal to $n^{-1/2}$ if $|\phi| < 1$. The $j$-th residual resulting from the fitted autoregressive model $\hat{\epsilon}_j$ can be written as

$$\hat{\epsilon}_j = X_j - \hat{\phi}X_{j-1}.$$
Using these residuals, we can construct the conventional kernel error density estimator as follows

\[
\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left( \frac{x - \hat{\epsilon}_i}{h} \right). \tag{1.15}
\]

Let \( f_n \) be the kernel density estimator based on unobservable errors, i.e.,

\[
f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left( \frac{x - \epsilon_i}{h} \right). \tag{1.16}
\]

Then, the Bickel Rosenblatt test statistics based on \( f_n \) and \( \hat{f}_n \) are defined as

\[
T_n = nh^{1/2} \int |f_n(x) - (K_h * f_0)(x)|^2 dx, \tag{1.17}
\]

\[
\hat{T}_n = nh^{1/2} \int |\hat{f}_n(x) - (K_h * f_0)(x)|^2 dx, \tag{1.18}
\]

where

\[
K_h(x) = \frac{1}{h} K\left( \frac{x}{h} \right), \quad (K_h * f_0)(x) = \int \frac{1}{h} K\left( \frac{x - u}{h} \right) f_0(u) du.
\]

Note that \((K_h * f_0)\) is the smoothed version of the null error density function. Next, we define the tilted version of the above density estimators by introducing tilted parameters \( p_1, \ldots, p_n \).

\[
f^*_n(x) = \frac{1}{h} \sum_{i}^{n} p_i K\left( \frac{x - \epsilon_i}{h} \right), \quad \hat{f}^*_n(x) = \frac{1}{h} \sum_{i}^{n} p_i K\left( \frac{x - \hat{\epsilon}_i}{h} \right). \tag{1.19}
\]
where \( f^*_n(x) \) and \( \hat{f}^*_n(x) \) correspond to \( f_n \) and \( \hat{f}_n \), respectively. Note that \( p_1 = \cdots = p_n = n^{-1} \) gives the conventional kernel error density estimators \( f_n \) and \( \hat{f}_n \). Thus, the test statistics based on the integrated squared error for \( f^*_n \) and \( \hat{f}^*_n(x) \) are defined, respectively, as

\[
T^*_n = n h^{1/2} \int |f^*_n(x) - (K_h * f_0)(x)|^2 dx,
\]

(1.20)

\[
\hat{T}^*_n = n h^{1/2} \int |\hat{f}^*_n(x) - (K_h * f_0)(x)|^2 dx.
\]

(1.21)

There exist four versions of density estimators, \( f_n, f^*_n, \hat{f}_n, \) and \( \hat{f}^*_n \), ordered corresponding to Bickel-Rosenblatt type test statistics \( T_n, T^*_n, \hat{T}_n, \) and \( \hat{T}^*_n \). The first two versions are included here for the sake of developments of the latter two because \( f_n \) and \( f^*_n \) are constructed from unobservable errors in the autoregressive models. In the setup of error density estimation, if assumptions \( \textbf{A1-A4} \) hold, by Theorem 4.1 of Bickel and Rosenblatt [8] and our Theorem 1 stated in Section 1.3,

\[
T_n - h^{-1/2} \int K^2(t) dt \xrightarrow{d} N(0, \tau^2),
\]

(1.22)

\[
T^*_n - h^{-1/2} \int K^2(t) dt \xrightarrow{d} N(0, \tau^2),
\]

(1.23)

where \( \tau^2 = 2 \int f_0^2(x) dx \int (\int K(t) K(s - t) d t)^2 ds, (K_h * f_0)(x) = \frac{1}{h} \int K(\frac{x - u}{h}) f_0(u) du. \)

The asymptotic behaviors of \( \hat{T}_n \) and \( \hat{T}^*_n \) are of primary interest. In this study, we extend the results of Lee and Na [9] and Bachmann and Dette [13] to the setup with tilted density estimators introduced by Doosti and Hall [1] and Doosti et al. [2].
1.4.2 Main Result

We first restate the problem of the goodness-of-fit test on the error density function \( f \),

\[
H_0 : f = f_0 \quad \text{vs.} \quad H_1 : f \neq f_0,
\]

where \( f \) is the unknown density function for errors in the first-order autoregressive model (1.14) and \( f_0 \) is a prescribed density. Our theorem below indicates that, under assumptions A1-A4, the test statistic \( \hat{T}^*_n \) has the same limiting normal distributions as \( T_n \) under the null hypothesis and fixed alternative.

**Theorem 2** Let the linear autoregressive time series model (1.14) satisfy \(|\phi| < 1\) and \(|\hat{\phi} - \phi| = O_p(n^{-1/2})\). If all conditions A1-A4 hold, the Bickel-Rosenblatt test statistic for the goodness-of-fit test on error distribution based on the tilted kernel density estimator (constructed from residuals)

\[
\hat{T}^*_n = nh^{1/2} \int (\hat{f}^*_n(x) - (K_h * f_0)(x))^2 \, dx
\]

has the following asymptotic properties:

(i) Under the null hypothesis \( H_0 : f = f_0 \),

\[
\hat{T}^*_n - h^{-1/2} \int K^2(t) \, dt \xrightarrow{d} \mathcal{N}(0, \tau^2),
\]

where \( \tau^2 = 2 \int f_0^2(x) \, dx \int (\int K(t)K(s-t) \, dt)^2 \, ds \).
(ii) Under the alternative $H_1: f \neq f_0$, as $n \to \infty$,

$$
(nh)^{-\frac{1}{2}} \hat{T}_n - n^\frac{1}{2} \int (K_h \ast (f - f_0))^2(x) dt \xrightarrow{d} N(0, 4\sigma^2), \quad (1.27)
$$

where $\sigma^2 = \text{Var}[(f - f_0)(X_1)]$.

1.5 Simulation Study

In this section, firstly, we use simulated data to examine the empirical performances of goodness of fit tests based on tilted density estimators relating to Theorem 1 with the iid setup and Theorem 2 with linear autoregressive models. We also compare the power of the tests based on the tilted density estimator and the conventional density estimator. Secondly, we use a real data example to illustrate the practical implementation of the Bickel-Rosenblatt test in linear autoregressive models. In all cases, we use the following smooth kernel satisfying assumptions $A2$:

$$
K(x) = \frac{e^{-1/(1-x^2)}}{0.444} I(-1 < x < 1) \quad (1.28)
$$

We would like to point out that the interpolation algorithm by Doosti et al. [2] is used to select bandwidths and probability weights. The goodness-of-fit test based on iid cases will be discussed in Section 1.5.1 and 1.5.2. In Section 1.5.3, we examine the goodness-of-fit test on the error density with AR(1) models. The real data analysis will be demonstrated in Section 1.5.4.
1.5.1 The Goodness-of-fit Test in IID Cases

To compare the performances of different density estimators, we choose $f_0$ as the density of the standard normal distribution. Figure 1.1 shows the plots of density function estimation based on iid random variables with $n = 100$. The red curve is the true density, the blue curve is the conventional density estimator, and the green and black denote tilted density estimators with two different weights ($m = 2$) and three different weights ($m = 3$), respectively. The integrated squared error (ISE) for the tilted estimators are 0.0038 ($m = 2$) and 0.0026 ($m = 3$), which are smaller than the ISE for conventional estimator, 0.0090.

![Figure 1.1: Density Function Estimators with IID Random Variables, n = 100.](image)

Red curve presents the true density function, which is the standard normal distribution. Black curve is the tilted density estimator with three distinct probability weights, which is $m = 3$.

We now present the results of goodness-of-fit tests for the hypotheses:

$$H_0 : f = f_0 \ vs. \ H_1 : f \neq f_0$$  \hspace{1cm} (1.29)
where $f_0$ is the density function of the standard normal distribution. We use the same fixed alternative as in Lee and Na [9], $f_1(x) = \zeta g(\zeta x)$, where $g(x) = (1-\alpha)f_0(x) + \alpha(1/\beta)f_0(x/\beta)$, $\zeta = 1 - \alpha + \alpha\beta^2$ and $\alpha = 0.1, \beta = 3$. We choose varying sample sizes, $n = 100, 200, 500, \text{and } 1000$, and use significance level 0.05. Under the null hypothesis, the empirical size is computed as rejection proportion out of 2000 repetitions, while the empirical power is calculated from 1000 repetitions under the alternative hypothesis. We reject the null hypothesis if the absolute value of test statistic $\hat{t}^* = |\hat{T}^* - h^{-1/2}\int K^2(x)dx| > z_{0.025}$.

Simulation results are presented in Table 1.1 and Table 1.2. As the sample size increases, the empirical size gets closer to the significant level 0.05, and the empirical power approaches 1. Figure 1.2 presents the power of Bickel-Rosenblatt test based on the conventional density estimator and the tilted density estimator. We can see that the Bickel-Rosenblatt test based on the tilted density estimator has higher power than the test based on the conventional density estimator.

<table>
<thead>
<tr>
<th></th>
<th>n=100</th>
<th>n=200</th>
<th>n=500</th>
<th>n=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Empirical Size</strong></td>
<td>0.035</td>
<td>0.040</td>
<td>0.042</td>
<td>0.044</td>
</tr>
<tr>
<td><strong>Empirical Power</strong></td>
<td>0.671</td>
<td>0.914</td>
<td>0.976</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 1.1: iid case: $m=2$, Empirical Size and Empirical Power

<table>
<thead>
<tr>
<th></th>
<th>n=100</th>
<th>n=200</th>
<th>n=500</th>
<th>n=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Empirical Size</strong></td>
<td>0.039</td>
<td>0.043</td>
<td>0.045</td>
<td>0.047</td>
</tr>
<tr>
<td><strong>Empirical Power</strong></td>
<td>0.651</td>
<td>0.925</td>
<td>0.995</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 1.2: iid case: $m=3$, Empirical Size and Empirical Power
1.5.2 The Goodness-of-fit Test in IID Cases for $m \geq 4$

In the previous section, we only present the simulation results for $m = 2$ and $m = 3$ with various sample sizes. In this section, our goal is to investigate the empirical behavior of the goodness-of-fit test with more probability weights scenarios when $m$ is greater than 3. Table 1.3 shows the empirical sizes for $m = 4$, 5 and 6. As we may see, for $m \geq 4$, the empirical size is greater than the significant level of 0.05 with sample sizes $n = 100$, 200, 500 and 1000.
<table>
<thead>
<tr>
<th></th>
<th>n=100</th>
<th>n=200</th>
<th>n=500</th>
<th>n=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>m = 4</td>
<td>0.045</td>
<td>0.077</td>
<td>0.077</td>
<td>0.083</td>
</tr>
<tr>
<td>m = 5</td>
<td>0.053</td>
<td>0.076</td>
<td>0.089</td>
<td>0.110</td>
</tr>
<tr>
<td>m = 6</td>
<td>0.066</td>
<td>0.076</td>
<td>0.093</td>
<td>0.108</td>
</tr>
</tbody>
</table>

Table 1.3: iid case: $m = 4, 5, 6$, Empirical Size under $H_0$

The performance of the goodness-of-fit test with $m \geq 4$ is likely related to the accuracy of the tilted density estimators. We now compare the accuracy of six versions of the density estimators: the conventional kernel density estimator, and the tilted density estimators with $m = 2, 3, 4, 5,$ and $6$. We present two simulated Scenarios with sample sizes 100 and 500, respectively.

**Scenario 1.** Figure 1.3 shows the plots of those six density estimators, and Table 1.4 summarizes their corresponding ISE’s. For $n = 100$, the tilted estimator with $m = 2$ or $3$ gives the smallest ISE and hence, the result of the goodness-of-fit test with $m = 2, 3$ is more reliable, which is consistent with results from Tables 1.1 to 1.3. The absolute value of test statistics for these two cases are $\hat{t}_{m=2}^* = 0.1357$ and $\hat{t}_{m=3}^* = 0.0308$. We fail to reject the null hypothesis with critical value 1.028.
Figure 1.3: Scenario 1: Density Function Estimators with Sample Size $n = 100$.

Compare the accuracy of estimation between tilted density estimators and the conventional density estimator. Among the density estimators, the tilted density estimator with $m = 3$ has the highest accuracy.

<table>
<thead>
<tr>
<th>m</th>
<th>ISE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional estimator</td>
<td>0.0056</td>
</tr>
<tr>
<td>m = 2</td>
<td>0.0045</td>
</tr>
<tr>
<td>m = 3</td>
<td>0.0038</td>
</tr>
<tr>
<td>m = 4</td>
<td>0.0048</td>
</tr>
<tr>
<td>m = 5</td>
<td>0.0059</td>
</tr>
<tr>
<td>m = 6</td>
<td>0.0067</td>
</tr>
</tbody>
</table>

Table 1.4: Scenario 1: ISE with Sample Size $n = 100$.

**Scenario 2.** Figure 1.4 displays plots of density estimators with sample size 500. From Table 1.5, we may see that $m = 2$ yields the smallest ISE, 0.00021. With $\hat{t}_{m=2}^* = 0.5741$, we fail to reject the null hypothesis at the level of 0.05.
Figure 1.4: Scenario 2: Density Function Estimators with Sample Size $n = 500$.

For this scenario, the tilted density estimator with $m = 2$ has the highest accuracy and the conventional density estimator has better performance than the tilted density estimators with $m \geq 3$.

<table>
<thead>
<tr>
<th>$m$</th>
<th>ISE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional estimator</td>
<td>0.00037</td>
</tr>
<tr>
<td>$m = 2$</td>
<td>0.00021</td>
</tr>
<tr>
<td>$m = 3$</td>
<td>0.00051</td>
</tr>
<tr>
<td>$m = 4$</td>
<td>0.00046</td>
</tr>
<tr>
<td>$m = 5$</td>
<td>0.00039</td>
</tr>
<tr>
<td>$m = 6$</td>
<td>0.00048</td>
</tr>
</tbody>
</table>

Table 1.5: Scenario 2: ISE with Sample Size $n = 500$

We would like to point out that, the empirical investigation here only gives a heuristic method to select $m$ for goodness-of-fit tests based on tilted density estimators. Theoretical developments of optimal choice of smoothing parameters including $m$, band-
width, and probability weights are currently under investigation.

Note that, the computation time of one time simulation for \( n = 100 \) is approximately 22 seconds. As the sample size increases, the computation time increases. In addition, based on the cross validation criterion, the optimal selection probability weight \( p_i \) of tilted density estimation could be very close to \( 1/n \), which is the weight of conventional density estimation. To distinct probability weights \( p_i \), we set a constrain in simulation that the difference between first two probability weights is greater than \( 10^{-7} \).

### 1.5.3 The Goodness-of-fit Test on the Error Density of AR(1) Model

With linear autoregressive models, we follow the same framework as in Section 1.5.1. We consider the AR(1) model,

\[
X_j = \phi X_{j-1} + \epsilon_j, \quad j = 1, 2, \ldots, n,
\]

where the coefficient \( \phi \) is estimated from simulated data by least squares principle,

\[
\hat{\phi} = \frac{\sum_{j=1}^{n} X_j X_{j-1}}{\sum_{j=1}^{n} X_{j-1}^2}
\]

We first compare the performances of error density estimators with \( \phi = 0.5 \) and \( n = 100 \). In Figure 1.5, the red curve is for the true error density, the blue, the conventional error density estimator with \( ISE = 0.0012 \), the green, the tilted density estimator for \( m = 2 \) with \( ISE = 0.0008 \), and the black, the tilted error density estimator for \( m = 3 \) with \( ISE = 0.0006 \).
$f_0(x)$ presents the standard normal distribution, which is predetermined error density function. Others are error density estimators. The tilted error density estimator with $m = 3$ has the higher accuracy.

To examine the performance of the goodness-of-fit test on the error density, we consider the hypotheses,

$$H_0 : f = f_0 \ vs. \ H_1 : f \neq f_0$$

where $f_0$ and $f_1(x)$ are the error densities prescribed as in Section 1.5.1 with the iid setup. We consider AR(1) models with $\phi = 0.1, 0.5, 0.9$ for each sample size. The simulation results based on $m = 2$ and $m = 3$ are summarized in Table 1.6 and Table 1.8 for empirical sizes, and Table 1.7 and Table 1.9 for empirical powers. It can be seen that, as the sample size increases, the empirical size gets closer to the significant level and the coefficients are irrelevant to the trend. For the empirical power, Table 1.7 and Table 1.9 indicate that as data size increases, the empirical power increases, and the trend is not much affected by the coefficient values. In addition, we compare power of the test
based on the tilted density estimator and the conventional density estimator and find
Bickel-Rosenblatt test based on tilted density estimator has high power (Figure 1.6). It is
consistent with the results from iid set up.

<table>
<thead>
<tr>
<th>φ</th>
<th>n=100</th>
<th>n=200</th>
<th>n=500</th>
<th>n=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.036</td>
<td>0.038</td>
<td>0.042</td>
<td>0.045</td>
</tr>
<tr>
<td>0.5</td>
<td>0.034</td>
<td>0.043</td>
<td>0.043</td>
<td>0.046</td>
</tr>
<tr>
<td>0.9</td>
<td>0.038</td>
<td>0.039</td>
<td>0.042</td>
<td>0.045</td>
</tr>
</tbody>
</table>

Table 1.6: Empirical Sizes under $H_0$ with m=2

<table>
<thead>
<tr>
<th>φ</th>
<th>n=100</th>
<th>n=200</th>
<th>n=500</th>
<th>n=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.709</td>
<td>0.936</td>
<td>0.998</td>
<td>1.000</td>
</tr>
<tr>
<td>0.5</td>
<td>0.714</td>
<td>0.945</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.9</td>
<td>0.708</td>
<td>0.937</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 1.7: Empirical Power under $H_1$ with m=2

<table>
<thead>
<tr>
<th>φ</th>
<th>n=100</th>
<th>n=200</th>
<th>n=500</th>
<th>n=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.036</td>
<td>0.040</td>
<td>0.044</td>
<td>0.047</td>
</tr>
<tr>
<td>0.5</td>
<td>0.036</td>
<td>0.042</td>
<td>0.044</td>
<td>0.048</td>
</tr>
<tr>
<td>0.9</td>
<td>0.037</td>
<td>0.040</td>
<td>0.045</td>
<td>0.048</td>
</tr>
</tbody>
</table>

Table 1.8: Empirical Size under $H_0$ with m=3
Table 1.9: Empirical Power under $H_1$ with $m=3$

<table>
<thead>
<tr>
<th>$\phi$</th>
<th>$n=100$</th>
<th>$n=200$</th>
<th>$n=500$</th>
<th>$n=1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.644</td>
<td>0.933</td>
<td>0.993</td>
<td>1.000</td>
</tr>
<tr>
<td>0.5</td>
<td>0.672</td>
<td>0.931</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.9</td>
<td>0.672</td>
<td>0.934</td>
<td>0.995</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Figure 1.6: Compare Power Between BR test Based on the Conventional Density Estimator and BR Test Based on the Tilted Density Estimator for AR(1) Model with $\phi = 0.5$.

The test based on tilted density estimators is much powerful than the test based on the conventional density estimator.

1.5.4 A Real Data Example

We present data analyses with a real data example.

Example. We use the number of occurrences of yearly worldwide magnitude 7+ earthquakes from 1916 to 2020 (https://www.usgs.gov/programs/earthquake-hazards/lists-maps-and-statistics).
First, we check the time series plot for the annual frequency of earthquakes of the past 105 years and then recenter the data around the mean. Figure 1.7 shows that the annual number of earthquakes oscillates around the center and the amplitude of the curve is almost the same over time. The Augmented Dickey-Fuller test is employed to check whether or not the data are from a stationary time series. The p-value of the test is 0.01, which indicates that we can fit a stationary time series model to the earthquake data.

The partial autocorrelation plot in Figure 1.8 suggests that we fit the an AR(3) model to the earthquake data set.
There are notable partial autocorrelations for lags 1 and 3. Third-order autoregression model can interpret the data.

The fitted time series model is \( \hat{X}_i = 0.29X_{i-1} + 0.063X_{i-2} + 0.135X_{i-3} \). And the estimated standard deviation of the noises is 3.8. Therefore, we take \( f_0 \) as the density function of the normal distribution with a mean of zero and a standard deviation of 3.8. Figure 1.9 presents a histogram of residuals from the fitted model. To choose an appropriate value for \( m \), we summarize ISE’s with different choices of \( m \) in Table 1.10. The conventional estimator has ISE 0.0016, and the tilted density estimator with \( m = 4 \) has the smallest ISE, 0.0006. As such, we use the goodness-of-fit test based on the tilted density estimator with \( m = 4 \). The critical value for the test is 0.5237, and the absolute value of the test statistic is \( \hat{t}_{m=4}^* = 0.1946 \). We fail to reject the null hypothesis that the errors in the re-centered AR(3) model are normally distributed with mean 0 and a standard deviation 3.8. Figure 1.10 displays the plots of density estimators, the prescribed density \( f_0 \) in red, the conventional estimator in blue, and also the tilted density estimators with different \( m \). We also perform the Kolmogorov–Smirnov test which results in the same conclusion with p-value \( = 0.8709 \).
Table 1.10: Earthquake Data: ISE Based on Different m

<table>
<thead>
<tr>
<th>m</th>
<th>ISE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional estimator</td>
<td>0.0016</td>
</tr>
<tr>
<td>m = 2</td>
<td>0.0015</td>
</tr>
<tr>
<td>m = 3</td>
<td>0.0007</td>
</tr>
<tr>
<td>m = 4</td>
<td>0.0006</td>
</tr>
<tr>
<td>m = 5</td>
<td>0.0008</td>
</tr>
<tr>
<td>m = 6</td>
<td>0.0009</td>
</tr>
</tbody>
</table>

Figure 1.9: Histogram of Residuals for Earthquake Data
To choose the optimal estimator, the tilted density estimator with $m = 4$ has the best performance among the estimators.
1.6 Technical Details

In this section, we first establish three lemmas which may be of interest on their own. The first lemma will be used in the development of the asymptotic normality of the Bickel-Rosenblatt test statistic based on tilted density estimators. It will be used in the proof of Theorem 1 by letting $W_i = X_i$ and in the proof of Theorem 2 by letting $W_i = \epsilon_i$ for $i = 1, \cdots n$. Lemma 2 gives the squared $L_2$-distance between the tilted density estimator and the conventional kernel density estimator. Lemma 3 is established in the setup of error density estimation in linear autoregressive models. It gives the bound of the squared $L_2$-distance between the tilted kernel error density estimator constructed from residuals in the fitted model and the tilted density estimator based on unobservable errors.

**Lemma 1** Let $W_1, \cdots, W_n$ be iid with density function $f$, then, under Assumptions A1-A3,

(i) $\int E\left(K\left(\frac{x-W_i}{h}\right)^2\right) dx = O(h)$,

$\int \left[ E\left(K\left(\frac{x-W_i}{h}\right)\right) \right]^2 dx = O(h^2)$,

(ii) $\int E\left(K^{(j)}\left(\frac{x-W_i}{h}\right)^2\right) dx = O(h)$, for $j = 1, 2$,

(iii) $\int \left[ E\left(K^{(j)}\left(\frac{x-W_i}{h}\right)\right) \right]^2 dx = O(h^2)$, for $j = 1, 2$.

**Proof.** Let $[-B, B]$ be the compact support of $K$. By assumptions A1 and A2, the function $K^2(\frac{x-w}{h})f(w)$ is integrable on $[-B, B] \times (-\infty, \infty)$. Then, by applying substitutions and
Fubini’s theorem, we have

\[
\int E\left[K\left(\frac{x-W_i}{h}\right)\right]^2 \, dx = \int \left( \int K^2\left(\frac{x-w}{h}\right) f(w) \, dw \right) \, dx
\]

\[
= \int \int K^2\left(\frac{x-w}{h}\right) f(w) \, dw \, dx
\]

\[
= h \int \int K^2(y) f(x-hy) \, dy \, dx
\]

\[
= h \int \left( \int f(x-hy) \, dx \right) K^2(y) \, dy
\]

\[
= h \int \left( \int f(t) \, dt \right) K^2(y) \, dy
\]

\[
= h \int K^2(y) \, dy
\]

\[
= O(h),
\]

here assumptions A2 implies \( \int K^2(y) \, dy < \infty \).

Likewise, by Hölder’s inequality and Fubini’s theorem, we get

\[
\int \left[ E\left[K\left(\frac{x-W_i}{h}\right)\right] \right]^2 \, dx = \int \left[ \int K\left(\frac{x-w}{h}\right) f(w) \, dw \right]^2 \, dx
\]

\[
= h^2 \int \left[ \int K(y) f(x-hy) \, dy \right]^2 \, dx
\]

\[
\leq h^2 \int \left[ \left( \int K^2(y) \, dy \right) \left( \int f^2(x-hy) \, dy \right) \right] \, dx
\]

\[
= h^2 \left( \int K^2(y) \, dy \right) \left( \int f^2(x-hy) \, dy \right) \, dx
\]

\[
= h^2 \left( \int K^2(y) \, dy \right) \left( \int f^2(t) \, dt \right) \, dy
\]

\[
= O(h^2),
\]

because \( f^2 \) is integrable which gives

\[
\int \left( \int f^2(t) \, dt \right) \, dy = \int_{-B}^{B} \left( \int f^2(t) \, dt \right) \, dy < \infty.
\]

We complete the proof of (i). Proofs of (ii) and (iii) can be done by following the same
Lemma 2  Let $f_n$ and $f^*_n$ be defined as in (1.1) and (1.5), respectively. If assumptions $A1$-$A4$ hold, we have the following result

$$
\int (f^*_n(x) - f_n(x))^2 dx = O_p(\delta_n h^{-1}) = O_p(n^{-1}/\sqrt{\log n}),
$$

(1.33)

where $\delta_n = \sum_{i=1}^n (p_i - n^{-1})^2 = O(n^{-1}h/\sqrt{\log n})$.

Proof. Note that $Var(f^*_n - f_n) = E(f^*_n - f_n)^2$ as $E(f^*_n - f_n) = 0$. By Lemma 1 (i)

$$
E \int (f^*_n - f_n)^2 dx = \int Var \left( \frac{1}{h} \sum_{i=1}^n \left( p_i - \frac{1}{n} \right) K \left( \frac{x - X_i}{h} \right) \right) dx
$$

$$
= \frac{1}{h^2} \int \sum_{i=1}^n \left( p_i - \frac{1}{n} \right)^2 Var \left( K \left( \frac{x - X_i}{h} \right) \right) dx
$$

$$
\leq \frac{1}{h^2} \sum_{i=1}^n \left( p_i - \frac{1}{n} \right)^2 \int E \left( K \left( \frac{x - X_i}{h} \right) \right)^2 dx
$$

$$
= \frac{1}{h^2} \sum_{i=1}^n \left( p_i - \frac{1}{n} \right)^2 \cdot O(h)
$$

$$
\leq O(\delta_n h^{-1}),
$$

(1.34)

which, together with Markov’s inequality, completes the proof of Lemma 2. Note that assumption $A4$ implies that $\delta_n \leq C_0 n^{-1}h/\sqrt{\log n}$, and hence, $O(\delta_n h^{-1}) = O_p(n^{-1}h/\sqrt{\log n})$.

Lemma 3  If the estimator $\hat{\phi}$ in the linear autoregressive model (1.14) satisfies $|\hat{\phi} - \phi| = O_p(n^{-1/2})$ and assumptions $A1$-$A4$ hold, then

$$
\int (f^*_n(x) - f^*_n(x))^2 dx = O_p(n^{-2}h^{-4}).
$$

(1.35)

Proof. For any positive integer $s$, any two sets of real numbers $a_1, \cdots, a_s$ and $b_1, \cdots, b_s$, framework.
we have

\[
\left( \sum_{i=1}^{s} a_i \right)^2 \leq s \sum_{i=1}^{s} a_i^2, \quad \left( \sum_{i=1}^{s} [a_i + b_i] \right)^2 \leq 2 \left( \sum_{i=1}^{s} a_i \right)^2 + 2 \left( \sum_{i=1}^{s} b_i \right)^2.
\]  

(1.36)

We will repeatedly apply the above results in this section. As stated in Section 1.3.1, the probability weights \( p_1, \cdots, p_n \) have \( m \) distinct values and \( p_i = p_i \) for \( i \)'s between \( i_{j-1} \) and \( i_j, \ j = 1, \cdots, m \). Thus, \( q_{\text{max}} = \max\{q_1, \cdots, q_m\} = \max\{p_1, \cdots, p_n\} \leq C/n \) by assumption A4. This, together with Equation (1.36), leads to

\[
\int \left( \hat{f}_n^*(x) - f_n^*(x) \right)^2 dx = \int \left( \frac{1}{h} \sum_{i=1}^{n} p_i K\left( \frac{x - \hat{\epsilon}_i}{h} \right) - \frac{1}{h} \sum_{i=1}^{n} p_i K\left( \frac{x - \epsilon_i}{h} \right) \right)^2 dx \\
= \int \left( \frac{1}{h} \sum_{i=1}^{n} p_i \left( K\left( \frac{x - \hat{\epsilon}_i}{h} \right) - K\left( \frac{x - \epsilon_i}{h} \right) \right) \right)^2 dx \\
= h^{-2} \int \left( \sum_{i=1}^{n} p_i \left( K\left( \frac{x - \hat{\epsilon}_i}{h} \right) - K\left( \frac{x - \epsilon_i}{h} \right) \right) \right)^2 dx.
\]

(1.37)

By Taylor’s expansion, for \( k = 1, \cdots, n \),

\[
K\left( \frac{x - \hat{\epsilon}_i}{h} \right) - K\left( \frac{x - \epsilon_i}{h} \right) \\
= \frac{\epsilon_i - \hat{\epsilon}_i}{h} K'\left( \frac{x - \epsilon_i}{h} \right) + \frac{(\epsilon_i - \hat{\epsilon}_i)^2}{2h^2} K''(\delta_{ix}) \\
= \frac{(\hat{\phi} - \phi) X_i}{h} K'\left( \frac{x - \epsilon_i}{h} \right) + \frac{(\hat{\phi} - \phi)^2 X_i^2}{2h^2} K''(\delta_{ix}),
\]

(1.38)

where \( \delta_{ix} \) is a quantity between \( (x - \epsilon_i)/h \) and \( (x - \hat{\epsilon}_i)/h \). Then, we have

\[
\int \left( \sum_{i=1}^{n} p_i \left( K\left( \frac{x - \hat{\epsilon}_i}{h} \right) - K\left( \frac{x - \epsilon_i}{h} \right) \right) \right)^2 dx \\
\leq 2 \int \left\{ \sum_{i=1}^{n} p_i \left( \frac{(\hat{\phi} - \phi) X_i}{h} K'\left( \frac{x - \epsilon_i}{h} \right) \right) \right\}^2 dx + 2 \int \left\{ \sum_{i=1}^{n} p_i \left( \frac{(\hat{\phi} - \phi)^2 X_i^2}{2h^2} K''(\delta_{ix}) \right) \right\}^2 dx \\
\leq \frac{2(\hat{\phi} - \phi)^2}{h^2} \int \left\{ \sum_{i=1}^{n} p_i X_i K'\left( \frac{x - \epsilon_i}{h} \right) \right\}^2 dx + \frac{(\hat{\phi} - \phi)^4}{2h^4} \int \left\{ \sum_{i=1}^{n} p_i X_i^2 K''(\delta_{ix}) \right\}^2 dx.
\]

(1.39)
Next, we employ the same technique as in Lee and Na [9] to derive the following equation by use of the fact that, under linear autoregressive models as in (1.14), the sequence \( \{X_j\}_{j \geq 1} \) is stationary; \( \epsilon_i \) is independent of \( X_1, \cdots, X_{i-1} \) for \( i = 1, \cdots, n \); and \( \epsilon_1, \cdots, \epsilon_n \) are iid. Note that by assumption \( A4 \), \( \max_{1 \leq i \leq n} p_i \leq C/n \)

\[
\Delta_n = \sum_{i=1}^{n} p_i^2 \leq C^2 n^{-1}
\]

Thus, we have

\[
E\left\{ \sum_{i=1}^{n} p_i X_{i-1} \left[ K'\left( \frac{X_i - \epsilon_i}{h} \right) - E\left( K'\left( \frac{X_i - \epsilon_i}{h} \right) \right) \right]^2 \right\} 
= E\left\{ \sum_{i=1}^{n} p_i^2 X_{i-1}^2 \left[ K'\left( \frac{X_i - \epsilon_i}{h} \right) - E\left( K'\left( \frac{X_i - \epsilon_i}{h} \right) \right) \right]^2 \right\} 
+ 2\left\{ \sum_{1 \leq i < l \leq n} p_i p_l E\left( X_{i-1} X_{l-1} \left[ K'\left( \frac{X_i - \epsilon_i}{h} \right) - E\left( K'\left( \frac{X_i - \epsilon_i}{h} \right) \right) \right] \cdot E\left[ K'\left( \frac{X_i - \epsilon_i}{h} \right) - E\left( K'\left( \frac{X_i - \epsilon_i}{h} \right) \right) \right] \right) \right\}
= E\left\{ \sum_{i=1}^{n} p_i^2 X_{i-1}^2 \left[ K'\left( \frac{X_i - \epsilon_i}{h} \right) - E\left( K'\left( \frac{X_i - \epsilon_i}{h} \right) \right) \right]^2 \right\} 
= \sum_{i=1}^{n} p_i^2 E(X_{i-1}^2) \cdot E\left[ K'\left( \frac{X_i - \epsilon_i}{h} \right) - E\left( K'\left( \frac{X_i - \epsilon_i}{h} \right) \right) \right]^2 
= M_2 \Delta_n E\left[ K'\left( \frac{X_i - \epsilon_i}{h} \right) - E\left( K'\left( \frac{X_i - \epsilon_i}{h} \right) \right) \right]^2 
\leq M_2 C^2 n^{-1} E\left( K'\left( \frac{X_1 - \epsilon_1}{h} \right) \right)^2 ,
\]

(1.40)

where, \( M_2 \) is the second moment of the stationary sequence \( \{X_j\}_{j \geq 1} \) in model (1.14). We used the fact that \( \epsilon_i \) is independent of \( X_{i-1}, \epsilon_i, \) and \( X_{i-1} \), which results in 0 for the second summation in the first equation.
Then, Equation (1.40) and Lemma 1 (ii) lead to

\[
E \int \left\{ \sum_{i=1}^{n} p_{i} X_{i-1} \left[ K'\left( \frac{X - \epsilon_{i}}{h} \right) - E\left( K'\left( \frac{X - \epsilon_{i}}{h} \right) \right) \right] \right\}^2 dx \\
\leq M_2 C^2 n^{-1} \int E\left( K'\left( \frac{X - \epsilon_1}{h} \right) \right)^2 dx \\
= O(n^{-1} h),
\]

(1.41)

which, together with Markov’s inequality, yields

\[
\int \left\{ \sum_{i=1}^{n} p_{i} X_{i-1} \left[ K'\left( \frac{X - \epsilon_{i}}{h} \right) - E\left( K'\left( \frac{X - \epsilon_{i}}{h} \right) \right) \right] \right\}^2 dx = O_p(n^{-1} h). \tag{1.42}
\]

It therefore follows that

\[
\int \left\{ \sum_{i=1}^{n} p_{i} X_{i-1} K'\left( \frac{X - \epsilon_{i}}{h} \right) \right\}^2 dx \\
\leq 2 \int \left\{ \sum_{i=1}^{n} p_{i} X_{i-1} \left[ K'\left( \frac{X - \epsilon_{i}}{h} \right) - E\left( K'\left( \frac{X - \epsilon_{i}}{h} \right) \right) \right] \right\}^2 dx \\
+ 2 \left( \sum_{i=1}^{n} p_{i} X_{i-1} \right)^2 \int \left[ E\left( K'\left( \frac{X - \epsilon_1}{h} \right) \right) \right]^2 dx 
\]

(1.43)

\[
= O_p(n^{-1} h) + O_p(n^{-1}) \cdot O(h^2) \\
= O_p(n^{-1} h).
\]

In the above derivation, we applied Lemma 1 (iii) and the property of stationary zero-mean AR(1) model:

\[
E \left( \sum_{i=1}^{n} p_{i} X_{i-1} \right)^2 = \sum_{i=1}^{n} p_i^2 E X_{i-1}^2 + 2 \sum_{1 \leq i < j \leq n} p_i p_j \phi^{(i-j)} \leq M_2 \Delta_n + 2C^2 n^{-2} \sum_{1 \leq i < j \leq n} \phi^{(i-j)} = O(n^{-1}).
\]

We now evaluate the second integration in Equation (1.39). Note that we may replace
\( X_{i-1} \) with \( X_i^2 \) and \( K'(\frac{x-\epsilon_i}{h}) \) with \( K''(\frac{x-\epsilon_i}{h}) \) in Equation (1.40) and obtain

\[
E\left\{ \sum_{i=1}^{n} p_i X_i^2 \left[ K''(\frac{x-\epsilon_i}{h}) - E\left[K''(\frac{x-\epsilon_i}{h})\right]\right]\right\}^2 \leq M_4 \Delta_n \left( K''(\frac{x-\epsilon_1}{h}) \right)^2,
\]

(1.44)

where \( M_4 \) is the finite fourth moment of the stationary sequence \( \{X_j\}_{j \geq 1} \) in AR(1) model (1.14). This, together with Lemma 1 (ii), gives

\[
E \int \left\{ \sum_{i=1}^{n} p_i X_i^2 \left[ K''(\frac{x-\epsilon_i}{h}) - E\left[K''(\frac{x-\epsilon_i}{h})\right]\right]\right\}^2 dx
\]

\[
\leq M_4 \Delta_n \int \left( K''(\frac{x-\epsilon_1}{h}) \right)^2 dx
\]

\[
= O(n^{-1} h),
\]

implying

\[
\int \left\{ \sum_{i=1}^{n} p_i X_i^2 \left[ K''(\frac{x-\epsilon_i}{h}) - E\left[K''(\frac{x-\epsilon_i}{h})\right]\right]\right\}^2 dx = O_p(n^{-1} h).
\]

(1.46)

Again, we may replace \( X_{i-1} \) with \( X_i^2 \) and \( K'(\frac{x-\epsilon_i}{h}) \) with \( K''(\frac{x-\epsilon_i}{h}) \) in Equation (1.43) and use Lemma 1 (iii) and Equation (1.44) to get

\[
\int \left\{ \sum_{i=1}^{n} p_i X_i^2 K''(\frac{x-\epsilon_i}{h}) \right\}^2 dx
\]

\[
\leq 2 \int \left\{ \sum_{i=1}^{n} p_i X_i^2 \left[ K''(\frac{x-\epsilon_i}{h}) - E\left[K''(\frac{x-\epsilon_i}{h})\right]\right]\right\}^2 dx
\]

\[
+ 2 \left( \sum_{i=1}^{n} p_i X_i^2 \right)^2 \int \left[ E\left(K''(\frac{x-\epsilon_1}{h})\right) \right]^2 dx
\]

\[
= O_p(n^{-1} h) + O_p(1) \cdot O(h^2)
\]

\[
= O_p(h^2),
\]

here, we use the fact that \( E\left(\sum_{i=1}^{n} p_i X_i^2\right) = M_2 \sum_{i=1}^{n} p_i = M_2 \), which implies that \( \sum_{i=1}^{n} p_i X_i^2 = O_p(1) \).
We now have

\[
\int \left\{ \sum_{i=1}^{n} p_i X_{i-1}^2 K''(\delta_i x) \right\}^2 \, dx \\
\leq 2 \int \left\{ \sum_{i=1}^{n} p_i X_{i-1}^2 \left[ K''(\delta_i x) - K''(\frac{x - \epsilon_i}{h}) \right] \right\}^2 \, dx + 2 \int \left\{ \sum_{i=1}^{n} p_i X_{i-1}^2 K''(\frac{x - \epsilon_i}{h}) \right\}^2 \, dx \\
\leq 2 \left\{ \sum_{i=1}^{n} p_i^2 x_i^4 \right\} \sum_{i=1}^{n} \int \left[ K''(\delta_i x) - K''(\frac{x - \epsilon_i}{h}) \right]^2 \, dx + O_p(h^2) \\
\leq 2 \left\{ \sum_{i=1}^{n} p_i^2 \cdot O_p(1) \right\} \cdot O \left\{ \sum_{i=1}^{n} \left| \hat{\epsilon}_i - \epsilon_i \right| \right\}^2 + O_p(h^2) \\
= O_p(n^{-1}) \cdot O_p(h^{-2}) + O_p(h^2) \\
= O_p(h^2) .
\]

if \( nh^4 \rightarrow \infty \) as \( n \rightarrow \infty \). In the above derivation, we used \( \sum_{i=1}^{n} |\hat{\epsilon}_i - \epsilon_i|^2 = |\hat{\phi} - \phi|^2 \sum_{i=1}^{n} X_{i-1}^2 = O_p(1) \) and applied assumption A2, which requires that \( K \) be compactly supported with a bounded third derivative \( K''' \). Finally, we obtain the following result by combining Equations (1.37), (1.39), (1.43) and (1.48) and using the condition \( |\hat{\phi} - \phi| = O_p(n^{-1/2}) \),

\[
\int (\hat{f}_n^*(x) - f_n^*(x))^2 \, dx \leq h^{-2} \left[ \frac{2(\hat{\phi} - \phi)^2}{h^2} \cdot O_p(n^{-1}h) + \frac{(\hat{\phi} - \phi)^4}{2h^4} \cdot O_p(h^2) \right] \\
= h^{-2} \left[ O_p(n^{-1}h^{-2}) \cdot O_p(n^{-1}h) + O_p(n^{-2}h^{-4}) \cdot O_p(h^2) \right] \\
= O_p(n^{-2}h^{-3} + n^{-2}h^{-4}) \\
= O_p(n^{-2}h^{-4}) .
\]

We complete the proof of Lemma 3.
1.6.1 Proof of Theorem 1

We begin by evaluating the difference between \( T^*_n \) and \( T_n \)

\[
T^*_n - T_n = \frac{nh^{1/2}}{2} \int (f_n^*(x) - K_h * f_0(x))^2 \, dx - nh^{1/2} \int (f_n(x) - K_h * f_0(x))^2 \, dx
\]

\[
= nh^{1/2} \int (f_n^*(x) - f_n(x))^2 \, dx + 2nh^{1/2} \int (f_n^*(x) - f_n(x))(f_n(x) - K_h * f_0(x)) \, dx.
\]  

(1.50)

Note that if assumptions A1-A3 hold, then, under \( H_0 : f = f_0 \)

\[
\int (f_n(x) - K_h * f_0(x))^2 \, dx = O_p\left( \frac{1}{nh} \right),
\]  

(1.51)

but under \( H_1 : f \neq f_0 \), we have

\[
\int (f_n(x) - K_h * f_0(x))^2 \, dx = O_p(1).
\]  

(1.52)

Proof of Part (i).

Under \( H_0 : f = f_0 \), by the Cauchy-Schwarz inequality, Lemma 2, and Equations (1.50) and (1.51), we have

\[
|T^*_n - T_n| \leq nh^{1/2} \int (f_n^*(x) - f_n(x))^2 \, dx
\]

\[
+ 2nh^{1/2} \left( \int (f_n^*(x) - f_n(x))^2 \, dx \right)^{1/2} \left( \int (f_n(x) - K_h * f_0(x))^2 \, dx \right)^{1/2}
\]

\[
= nh^{1/2} \cdot O_p(\delta_n h^{-1}) + 2nh^{1/2} \cdot O_p(\delta_n^{1/2} h^{-1/2}) \cdot O_p(n^{-1/2} h^{-1/2})
\]

\[
= O_p(nh^{-1/2} \delta_n) + O_p(n^{1/2} h^{-1/2} \delta_n^{1/2})
\]

\[
= O_p(n^{1/2} h^{-1/2} \delta_n^{1/2})
\]

\[
= o_p(1),
\]  

(1.53)
if $\delta_n \leq C_0 n^{-1} h / \sqrt{\log n}$. Theorem 1 (i) follows from Equations (1.3), (1.53), and Slutsky’s theorem.

**Proof of Part (ii)**

Under $H_1 : f \neq f_0$, by Lemma 2 and Equation (1.52), we have

\[
|T_n^* - T_n| \leq n h^{1/2} \int (f_n^* (x) - f_n(x))^2 dx \\
+ 2 n h^{1/2} \left( \int (f_n^* (x) - f_n(x))^2 dx \right)^{1/2} \left( \int (f_n(x) - K_n * f_0(x))^2 dx \right)^{1/2} \\
= nh^{1/2} \cdot O_p (\delta_n h^{-1}) + 2 nh^{1/2} \cdot O_p (\delta_n^{1/2} h^{-1/2}) \cdot O_p (1) \\
= O_p (n \delta_n^{1/2}),
\]

which leads to

\[
(n h)^{-1/2} (T_n^* - T_n) = (n h)^{-1/2} \cdot O_p (n \delta_n^{1/2}) = O_p (n^{1/2} h^{-1/2} \delta_n^{1/2}) = o_p (1).
\]

This, together with Equation (1.4) and Slutsky’s theorem, yields Theorem 1 (ii).

We complete the proof of Theorem 1.

**1.6.2 Proof of Theorem 2**

In this section, we focus on the goodness-of-fit test about the error density function of first-order autoregressive models. Here, we use $T_n$ and $T_n^*$ to denote Bickel-Rosenblatt test statistics for the conventional kernel error density estimator and its tilted version based on unobservable errors in the model. As such, all previously established results for the density estimators apply. As defined in Section 1.4, $\hat{T}_n$ and $\hat{T}_n^*$ are versions of $T_n$ and $T_n^*$ based on residuals from the fitted models, respectively.
Proof of Part (i).

Under $H_0 : f = f_0$, by the triangle inequality, Equation (1.51) and Lemmas 2 and 3, we obtain

$$\left| \hat{T}_n - T_n^* \right| \leq nh^{1/2} \int (\hat{f}_n^*(x) - f_n^*(x))^2 \, dx$$

$$+ 2nh^{1/2} \left( \int (\hat{f}_n^*(x) - f_n^*(x))^2 \, dx \right)^{1/2} \left( \int (f_n^*(x) - K_h \ast f_0(x))^2 \, dx \right)^{1/2}$$

$$\leq nh^{1/2} \int (\hat{f}_n^*(x) - f_n^*(x))^2 \, dx + 2nh^{1/2} \left( \int (\hat{f}_n^*(x) - f_n^*(x))^2 \, dx \right)^{1/2}$$

$$\cdot \left[ \left( \int (f_n^*(x) - f_n(x))^2 \, dx \right)^{1/2} + \left( \int (f_n(x) - K_h \ast f_0(x))^2 \, dx \right)^{1/2} \right]$$

$$= nh^{1/2} \cdot O_p(n^{-2}h^{-4}) + 2nh^{1/2} \cdot O_p(n^{-2}h^{-4})^{1/2}$$

$$\cdot \left[ O_p(\delta_n h^{-1})^{1/2} + O_p(n^{-1}h^{-1})^{1/2} \right]$$

$$= O_p(n^{-1}h^{-7/2}) + O_p(h^{-3/2}) \cdot \left[ O_p(\delta_n^{1/2} h^{-1/2}) + O_p(n^{-1/2}h^{-1/2}) \right]$$

$$= O_p(n^{-1}h^{-7/2}) + O_p(\delta_n^{1/2} h^{-2}) + O_p(n^{-1/2}h^{-2})$$

$$= O_p(n^{-1/2}h^{-2} + \delta_n^{1/2} h^{-2})$$

$$= O_p(n^{-1/2}h^{-2})$$

$$= o_p(1),$$

if $\delta_n \leq C_0 n^{-1}h/\sqrt{\log n}$, $h \to 0$, and $nh^4 \to \infty$ as $n \to \infty$.

Thus, Slutsky's theorem and Equation (1.56) imply that, under $H_0$, $\hat{T}_n^*$ has the same asymptotic normal distributions as $T_n^*$, and hence, Theorem 2 (i) follows from Theorem 1 (i).

Proof of Part (ii)
Under $H_1: f \neq f_0$, by Equation (1.52) and Lemmas 2 and 3, we have

$$|\hat{T}_n^* - T_n^*| \leq nh^{1/2} \int (\hat{f}_{n}^*(x) - f_{n}^*(x))^2 dx$$

$$+ 2nh^{1/2} \left( \int (\hat{f}_{n}^*(x) - f_{n}^*(x))^2 dx \right)^{1/2} \left( \int (f_{n}^*(x) - K_h * f_0(x))^2 dx \right)^{1/2}$$

$$\leq nh^{1/2} \int (\hat{f}_{n}^*(x) - f_{n}^*(x))^2 dx + 2nh^{1/2} \left( \int (\hat{f}_{n}^*(x) - f_{n}^*(x))^2 dx \right)^{1/2}$$

$$\cdot \left[ \left( \int (f_{n}^*(x) - f_{n}(x))^2 dx \right)^{1/2} + \left( \int (f_{n}(x) - K_h * f_0(x))^2 dx \right)^{1/2} \right]$$

$$= nh^{1/2} \cdot O_p(n^{-2}h^{-4}) + 2nh^{1/2} \cdot O_p(n^{-2}h^{-4})^{1/2}$$

$$\cdot \left[ O_p(\delta_n h^{-1})^{1/2} + O_p(1) \right]$$

$$= O_p(n^{-1}h^{-7/2}) + O_p(h^{-3/2}) \cdot [O_p(1)]$$

$$= O_p(h^{-3/2}),$$

which implies that

$$(nh)^{-1/2}(\hat{T}_n^* - T_n^*) = (nh)^{-1/2} \cdot O_p(h^{-3/2}) = O_p(n^{-1/2}h^{-2}) = o_p(1). \quad (1.58)$$

Finally, Theorem 2 (ii) follows from Theorem 1 (ii), Equation (1.58), and Slutsky’s theorem.
Bickel-Rosenblatt Test Based on Tilted Density Estimation for Nonlinear Autoregressive Models

2.1 Introduction

This chapter considers the goodness-of-fit test on the error distribution in nonlinear autoregressive time series models. We propose a new type of the Bickel-Rosenblatt test by use of the tilted density estimator based on residuals from the fitted model and show that, under appropriate conditions, the corresponding test statistic has an asymptotic normal distribution. Moreover, our simulation studies indicate that this new type of test is more powerful than the one based on the conventional kernel error density estimator. A real data example is also presented to demonstrate the practical implementation of this goodness-of-fit test.

The rest of the chapter is organized as follows. In Section 2.1, we show motivation and literature review of this topic. In Section 2.2, we state the problem of interest and propose the new type of Bickel-Rosenblatt test statistic based on the tilted error density estimator. The main result is given in Section 2.3. We present simulation studies in Section 2.4 and provide technical details of the theoretical development in Section 2.5.
In Section 2.6, we briefly discuss two possible directions for future investigation.

## 2.1.1 Motivation

In Chapter 1, we present the motivation of the study of the new Bickel-Rosenblatt test based on tilted density estimation for linear autoregressive model and show advantages of the new test statistic. Many researchers are working on modeling economic and financial time series data. Even though linear autoregressive models are very commonly studied in academic and applied in research, there are still some unexplained aspects for financial and economic data.

Therefore, in this chapter, we extend the new type Bickel-Rosenblatt test based on tilted density estimation for linear autoregressive model to nonlinear autoregressive model.

## 2.1.2 Literature Review

Underlying distributions of random errors play an essential role in statistical inferences of autoregressive models. Goodness-of-fit tests are often used in time series models to determine whether the random errors follow a prescribed distribution, such as Gaussianity. In autoregressive time series models, goodness-of-fit tests based on residual empirical processes have been extensively studied by many researchers, including Boldin [3], Pierce [4] and Koul ([5], [6]). However, with the empirical process-based method, it is sometimes challenging to calculate critical values of a test, as the parameter estimators may affect the limiting process (cf. Durbin [7]). Another type of goodness-of-fit test is the Bickel-Rosenblatt type test, which is based on the integrated squared error of a non-parametric kernel error density estimator. Some of the advantages of the Bickel-Rosenblatt test statistic are that its limiting distribution does not depend on any estimators of nuisance parameters and that, in the present of sharp peak alternatives (cf. Ghosh and Huang [11]), it outperforms the Cramer-von Mises test and Kolmogorov-
Smirnov test, both of which are based on empirical processes.

In this study, we propose a new type of Bickel-Rosenblatt test on errors in nonlinear autoregressive models and show that the corresponding test statistic has an asymptotic normal distribution. Moreover, our simulation studies indicate that this new type of goodness-of-fit test constructed from the tilted error density estimator is more powerful than that based on the conventional kernel error density estimator. With advances in stochastic processes, statistical inferences for nonlinear autoregressive models have become increasingly important, and we need more techniques for model assessments. For example, smooth transition autoregressive (STAR) models consist of a popular family of nonlinear autoregressive models. The normality assumption for random errors is needed if the specification tests are derived by Lagrange multiplier type tests (cf. Terasvirta [27]). The goodness-of-fit test of an error density involves two steps: estimating the error density function based on residuals from the fitted model and constructing the test statistic based on this estimator. A sensible density estimator likely results in a reasonable goodness-of-fit test. This work is partly motivated by two recent papers, Doosti and Hall [1] and Doosti et al. [2], in which the authors investigated the tilted kernel density estimator and showed by both theoretical developments and simulation studies that this new type of estimator outperforms the conventional kernel density estimator.

Bickel and Rosenblatt [8] first introduced the goodness-of-fit test based on the integrated squared error of the nonparametric density estimator and showed that its test statistic is asymptotically normally distributed. Lee and Na [9] applied the Bickel-Rosenblatt type statistic to the goodness-of-fit test on error distribution in linear autoregressive models. They showed that, under the null hypothesis, the limiting distribution of the test statistic does not depend on unknown parameters in the model, making it relatively easy to find the critical values. Bachmann and Dette [13] established the limiting distribution of the test statistic under a fixed alternative, which makes it possible to evaluate the power of the test. Cheng and Sun [10] generalized the existing results to nonlin-
ear autoregressive models. In chapter 1, we first used the tilted error density estimator to construct the goodness-of-fit test on errors in linear autoregressive models. In this work, we investigate goodness-of-fit tests by use of the tilted error density estimator in nonlinear autoregressive models.

Our contribution in this work is to construct the Bickel-Rosenblatt test based on tilted error density estimator in nonlinear autoregressive models and show that the test statistic is asymptotically normal distributed. We explore the goodness-of-fit test (about errors in nonlinear autoregressive models) based on tilted density estimators and show that the corresponding Bickel-Rosenblatt test statistic has the same limiting distributions as the goodness-of-fit test statistic based on conventional kernel density estimators (cf. Theorem 3).

2.2 Statement of The Problem

Let \( X = \{X_i\} \) be a strictly stationary stochastic process described by the following nonlinear autoregressive model

\[
X_i = \alpha_\theta(X_{i-1}, X_{i-2}, \ldots, X_{i-p}) + \epsilon_i, \quad (2.1)
\]

where \( \alpha_\theta \) is a family of measurable functions from \( \mathbb{R}^p \to \mathbb{R} \) with \( \theta = (\theta_1, \theta_2, \ldots, \theta_q)' \in \Theta \subset \mathbb{R}^q \). Additionally, \( \{\epsilon_i\} \) are i.i.d. random errors with an unknown density function \( f \), and \( X_{i-1}, \ldots, X_{i-p} \) are independent of \( \epsilon_i \), for \( i \geq 1 \). Our goal is to conduct the following hypothesis test on the unknown error density \( f \),

\[
H_0 : f = f_0 \quad \text{vs.} \quad H_1 : f \neq f_0 , \quad \tag{2.2}
\]

where \( f_0 \) is a prescribed error density function.
We first construct a tilted kernel estimator for the error density function from residuals. Let \( \hat{\theta} \) be an estimator of \( \theta \), satisfying the law of the iterated logarithm, as stated by assumption A3 in the next section. From the fitted model, we obtain the following residuals:

\[
\hat{e}_i = X_i - \alpha_\hat{\theta}(X_{i-1}, X_{i-2}, \ldots, X_{i-p}), \quad i = 1, \cdots, n.
\] (2.3)

The tilted error density estimator is defined as

\[
\hat{f}_n^*(x) = \frac{1}{h} \sum_{i=1}^{n} p_i K\left( \frac{x - \hat{e}_i}{h} \right),
\] (2.4)

where \( p_1, \cdots, p_n \) are the probability weights satisfying \( \sum_{i=1}^{n} p_i = 1 \), and \( p_1 = \cdots = p_n = 1/n \) yields the conventional kernel density estimator. The tilted density estimator differs from the conventional kernel estimator in that, instead of using the uniform weight for all observations, a set of probability weights \( p = (p_1, \cdots, p_n)^T \) are introduced for finer tuning. The integrated squared error of \( \hat{f}_n^* \) is defined as:

\[
ISE(h, p) = \int (\hat{f}_n^*(x|h, p) - f(x))^2 \, dx.
\] (2.5)

Theoretically, we want to choose tuning parameters \( h \) and \( p_1, \cdots, p_n \) so that \( ISE(h, p) \) is minimized, which is unrealistic because the error density is unknown. Doosti et al. [2] proposed the following cross-validation criterion:

\[
CV(h, p) = \int \hat{f}_n^*(x|h, p)^2 \, dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_n^*(\hat{e}_i|h, p).
\] (2.6)
where \( \hat{f}_{-i}^* \) is the leave-one-out version of \( \hat{f}_n^* \) by deleting \( \hat{e}_i \) from the original data, i.e.

\[
\hat{f}_{-i}^*(x|h, p) = \frac{1}{h} \sum_{j: j \neq i} p_j K\left(\frac{x - \hat{e}_j}{h}\right).
\]

The authors showed that under appropriate conditions, \( CV(h, p) \) is a good approximation of \( ISE(h, p) \). For practical implementation, these \( n \) probability weights may only take \( m < n \) distinct values, say \( q_1, \ldots, q_m \). In this study, we employ the sparse interpolation algorithm proposed by Doosti et al. [2] to choose \( h \) and probability weights by minimizing \( CV(h, p) \): we pick \( i_1 < \cdots < i_m \) between 1 and \( n \) so that \( C_1 n/m \leq i_j - i_{j-1} \leq C_2 n/m \) for constants \( C_1, C_2 > 0 \), and we set \( p_i = q_j \) for all \( \hat{e}(i) \) with \( i_{j-1} < i \leq i_j \). Here \( \hat{e}(i) \) is the \( i \)-th order statistic of \( \hat{e}_1, \ldots, \hat{e}_n \). As such, the tuning parameters under this algorithm are \( h, q_1, \ldots, q_m \). We refer to Doosti et al. [2] for more details about this algorithm.

We then define our new type of Bickel-Rosenblatt test statistic for the above goodness-of-fit test as

\[
\hat{T}_n^* = nh^{1/2} \int \left| \hat{f}_n^*(x) - (K_h * f_0)(x) \right|^2 dx., \quad (2.7)
\]

where,

\[
K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right), \quad (K_h * f_0)(x) = \int \frac{1}{h} K\left(\frac{x - u}{h}\right) f_0(u) du. \quad (2.8)
\]

Our focus here is to investigate the goodness-of-fit test on the error density by establishing the limiting distribution of \( \hat{T}_n^* \) and conducting simulation studies to evaluate the empirical performance of this new type of test in the setup of nonlinear autoregressive models.

As convention, we assume throughout this study that, unless otherwise specified, limits are taken as \( n \to \infty \).
2.3 Main Result

We begin this section by introducing some basic assumptions on the nonlinear autoregressive model (2.1) and on the tilted kernel density estimator defined in (2.4).

**A1.** Let \( U \subset \Theta \subset \mathbb{R}^q \) be an open neighborhood of \( \theta \). We assume that for all \( y \in \mathbb{R}^p \), \( \theta = (\theta_1, \theta_2, ..., \theta_q) \in U \), \( s, t = 1, ..., q \),

\[
\left| \frac{\partial}{\partial \theta_s} \alpha_\theta(y) \right| \leq M_1(y), \quad \left| \frac{\partial^2}{\partial \theta_s \partial \theta_t} \alpha_\theta(y) \right| \leq M_2(y),
\]

where \( EM_1^4(X_{i-1}, ..., X_{i-p}) < \infty \) and \( EM_2^4(X_{i-1}, ..., X_{i-p}) < \infty \) for \( i \geq 1 \).

**A2.** Let \( Y_{is} = \frac{\partial}{\partial \theta_s} \alpha_\theta(X_i, ..., X_{i-p}) \), where \( s \in \{1, ..., q\} \) and \( i \in \{1, ..., n\} \).

\[
\sum_{i=1}^{n} Y_{is}^2 = O_p(n^\beta), \text{ where } 0 < \beta < 1.
\]

**A3.** There exists a positive constant \( C_0 \) such that

\[
\limsup_{n \to \infty} \sqrt{n \log \log(n)} \sum_{s=1}^{q} (\hat{\theta}_s - \theta_s)^2 \leq C_0.
\]

**B1.** \( f, |f'|, \) and \( |f''| \) are uniformly bounded.

**B2.** \( K \) is a symmetric, compactly supported kernel with bounded second and third derivatives, and

\[
\int K(t)dt = 1. \quad (2.10)
\]

**B3.** For bandwidth \( h, n \to \infty, h \to 0 \), and \( nh^4 \to \infty \).
For $p$, the probability weights $p_1, \cdots, p_n$ satisfy

$$\delta_n := \sum_{i=1}^{n} (p_i - n^{-1})^2 \leq C_1 n^{-1} h / \sqrt{\log n}, \quad \max\{p_1, \cdots, p_n\} \leq C_2 / n$$

for some positive constants $C_1, C_2$.

**Remark 3** Assumptions $A_1$ to $A_3$ are imposed on the family of nonlinear autoregressive models. Similar conditions are used in Cheng and Sun [10]. Assumptions $B_1$ to $B_3$ are usual conditions on kernel density estimation. With regard to data tilting, probability weights $p_1, \cdots, p_n$ are introduced as more tuning parameters. Assumption $B_4$ guarantees that $\hat{f}_n^*$ is not too different from the conventional kernel estimator. This condition is weaker than that used by Doosti and Hall [1] but stronger than that of Doosti et al. [2].

Doosti and Hall [1] used condition $p_i = n^{-1}(1 + h^2 d(X_i))$. In a following work, Doosti et al. [2] proposed a sparse interpolation algorithm, which reduces the number of tuning parameters to $m + 1$ by imposing the condition that those probability weights have $m$ distinct values, say $q_1, \cdots, q_m$. They used assumption $\max_{1 \leq i \leq n} |p_i - n^{-1}| \leq h^2 n^{c_2} / \sqrt{\log n}$, for some constant $c_2$ satisfying $2c_1 + c_2 < \frac{1}{10}$, which leads to

$$\max_{1 \leq i \leq n} |p_i - n^{-1}| \leq h^2 n^{c_2} < n^{-2/5+2c_1} n^{1/10-2c_1} / \sqrt{\log n} = n^{-3/10} / \sqrt{\log n}.$$

We are now in a position to state the main result.

**Theorem 3** In the setup of the nonlinear autoregressive time series model (2.1), suppose that all assumptions $A_1$ to $A_3$ and $B_1$ to $B_4$ hold. Then, if bandwidth $h$ satisfies $n^{1-\beta} h^2 / (\log \log n) \rightarrow$
∞, the new Bickel-Rosenblatt test statistic (constructed from residuals),

$$\hat{T}_n^* = nh^{1/2} \int (\hat{f}_n^*(x) - (K_h \ast f_0)(x))^2 \, dx,$$  \hspace{1cm} (2.11)

has the following asymptotic properties:

(i) Under the null hypothesis $H_0 : f = f_0$,

$$\hat{T}_n^* - h^{-1/2} \int K^2(t) \, dt \overset{d}{\rightarrow} N(0, \tau^2),$$  \hspace{1cm} (2.12)

where $\tau^2 = 2 \int f_0^2(x) \, dx \int (\int K(t)K(s-t) \, dt) \, ds$.

(ii) Under the alternative $H_1 : f \neq f_0$, as $n \to \infty$,

$$\left(nh^{-1/2} \hat{T}_n^* - n^{1/2} \int (K_h \ast (f - f_0))^2(x) \, dt \right) \overset{d}{\rightarrow} N(0, 4\sigma^2),$$  \hspace{1cm} (2.13)

where $\sigma^2 = Var[(f - f_0)(X_1)]$.

Theorem 3 indicates that the goodness-of-fit test statistic based on the tilted kernel density estimator has the same limiting distributions as the one based on the conventional kernel density estimator. Theorem 3 (i) allows us to find the critical values for rejecting $H_0$, and Theorem 3 (ii) makes it possible to calculate the power of the test under certain alternatives.

### 2.4 Empirical Studies

In this section, we first simulate data from a prescribed nonlinear autoregressive model and then check the empirical performance of goodness-of-fit tests based on tilted den-
sity estimators relating to Theorem 3. Next, we analyze a real-world dataset to demonstrate the practical implementation of the new Bickel-Rosenblatt test with nonlinear autoregressive models.

Throughout this section, we use the kernel function

$$K(x) = \frac{e^{-1/(1-x^2)}}{0.444} I(-1 < x < 1).$$

(2.14)

We use the sparse interpolation algorithm given by Doosti et al. [2] to choose \( h \) and probability weights by minimizing \( CV(h, p) \).

### 2.4.1 Simulation Studies

For the simulation, we use the following logistic smooth transition autoregressive (LSTAR) model of order 2 (cf. Terasvirta [27]):

$$X_t = 1.80X_{t-1} - 1.06X_{t-2} + (0.02 - 0.9X_{t-1} + 0.795X_{t-2})\left(1 + e^{-100(X_{t-1} - 0.02)}\right)^{-1} + \epsilon_t,$$

(2.15)

where \( \epsilon_t \sim NID(0, 1) \). Coefficients in the model are estimated from simulated data by the conditional least squares method (cf. Klimko and Nelson [28]). We consider the following goodness-of-fit test:

$$H_0 : f = f_0 \; \text{vs.} \; H_1 : f \neq f_0,$$

(2.16)

where \( f_0 \) is the density function of the standard normal distribution, and the fixed alternative, \( f_1 \) (cf. Lee and Na [9]), is taken as

$$f_1(x) = \zeta g(\zeta x),$$
where \( g(x) = (1-\alpha) f_0(x) + \alpha(1/\beta) f_0(x/\beta); \ \zeta = 1 - \alpha + \alpha \beta^2; \) and \( \alpha = 0.1, \beta = 3. \) In the setup of linear autoregressive models, in chapter 1, we showed that the Bickel-Rosenblatt test, constructed from the tilted density estimator, is a reliable and useful method for the goodness-of-fit test when \( m = 2 \) and 3. For \( m \geq 4 \), the empirical size can be greater than the significance level, 0.05. Therefore, in this study, we focus on the performance of the goodness-of-fit test with \( m = 2 \) and 3.

To examine the performance of our test statistic, we simulate data from model (2.15) with different sample sizes \( n = 100, 200, 500, \) and \( 1000, \) and we use 0.05 as the specified significance level. The absolute value of the test statistic is \( \hat{t}^* = |\hat{T}^*_n - h^{-1/2} \int K^2(x) dx|. \) The null hypothesis will be rejected if

\[
\hat{t}^* = |\hat{T}^*_n - h^{-1/2} \int K^2(x) dx| > z_{0.025} \tau = 1.028.
\]

The empirical size is calculated from the rejected number out of 2000 repetitions, and the empirical power is calculated from 1000 repetitions under the alternative. Simulation results for \( m = 2 \) and 3 are presented in Table 2.1 and Table 2.2, respectively. Note that as the sample size increases, the empirical size also increases—getting close to 0.05—and the empirical power approaches 1. Moreover, Figure 2.1 indicates that our new Bickel-Rosenblatt test based on the tilted error density estimator is more powerful than the test based on the conventional density estimator.

<table>
<thead>
<tr>
<th>( n=100 )</th>
<th>( n=200 )</th>
<th>( n=500 )</th>
<th>( n=1000 )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Empirical Size</strong></td>
<td>0.040</td>
<td>0.042</td>
<td>0.043</td>
</tr>
<tr>
<td><strong>Empirical Power</strong></td>
<td>0.721</td>
<td>0.945</td>
<td>0.998</td>
</tr>
</tbody>
</table>

Table 2.1: \( m=2 \), Empirical Size and Empirical Power
Next, we compare the performance of six error density estimators: the conventional kernel density estimator and tilted density estimators with $m = 2, 3, 4, 5, \text{ and } 6$. Here, we present the simulation results for the sample size $n = 100$.

Figure 2.2 displays the plots of different density estimators, and Table 2.3 summarizes the corresponding ISEs. In this example, the tilted estimator with $m = 3$ gives the smallest ISE, 0.00060, which is consistent with the simulation result in chapter 1 in the setup of linear autoregressive models. We conclude that the result of the goodness-of-fit tests with $m = 2$ and 3 are more reliable. The absolute values of the test statistic for these two cases are $\hat{t}_{m=2}^* = 0.4550$ and $\hat{t}_{m=3}^* = 0.5487$, respectively. We fail to reject the null hypothesis with critical value 1.028.
Figure 2.2: Density Function Estimators with Sample Size $n = 100$

<table>
<thead>
<tr>
<th>m</th>
<th>ISE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional estimator</td>
<td>0.00153</td>
</tr>
<tr>
<td>m = 2</td>
<td>0.00092</td>
</tr>
<tr>
<td>m = 3</td>
<td>0.00060</td>
</tr>
<tr>
<td>m = 4</td>
<td>0.00080</td>
</tr>
<tr>
<td>m = 5</td>
<td>0.00084</td>
</tr>
<tr>
<td>m = 6</td>
<td>0.00088</td>
</tr>
</tbody>
</table>

Table 2.3: ISE with Sample Size $n = 100$

2.4.2 A Real Data Example

We use U.S. data of the Industrial Production Index (INDPRO) from 2010 to 2020 (https://fred.stlouisfed.org/series/INDPRO) to demonstrate how to implement our proposed goodness-of-fit test. We first check the time series plot for the monthly
and seasonally adjusted industrial production index. Figure 2.3 shows that the industrial production index rose from 2010 to 2013, collapsed between 2013 and 2016, and then recovered until another sudden collapse in May 2020.

Smooth transition autoregressive (STAR) models are often used to fit this type of data. A STAR model of order \( p \) is defined as

\[
X_t = \pi_{10} + \pi_1' W_t + (\pi_{20} + \pi_2' W_t) F(X_{t-d}) + \epsilon_t, \tag{2.17}
\]

where \( \epsilon_t \sim N(0, \sigma^2_\epsilon) \), \( W_t = (X_{t-1}, \ldots, X_{t-p})' \), and \( \pi_j = (\pi_{j1}, \ldots, \pi_{jp})' \) for \( j = 1, 2 \). Here, \( d \) is called the delay parameter, and \( F \), bounded between 0 and 1, is called the smooth transition function. We will consider two families of STAR models, the logistic smooth transition autoregressive (LSTAR) model with the transition function

\[
F(X_{t-d}) = \left(1 + e^{-\gamma(X_{t-d}-c)}\right)^{-1},
\]

and the exponential smooth transition autoregressive (ESTAR) model with the transition function
function

\[ F(X_{t-d}) = 1 - e^{-\gamma (X_{t-d} - c)^2}, \]

where \( c \) is the threshold value and \( \gamma > 0 \) determines the speed and smoothness of the transition. To select an appropriate model for the INDPRO dataset, we follow the procedure proposed by Terasvirta [27].

**Step 1.** Fix the delay parameter \( d \) between 1 and \( D \). We fit a linear autoregressive model of order \( p \):

\[ X_t = \beta_0 + \beta_1 X_{t-1} + \ldots + \beta_p X_{t-p} + \epsilon_t, \quad t \in [1, T] \]

We choose AR(3) as the best autoregressive model because it has the smallest Akaike Information Criterion (AIC) in Table 2.4.

<table>
<thead>
<tr>
<th>( p )</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>452.60</td>
</tr>
<tr>
<td>2</td>
<td>443.24</td>
</tr>
<tr>
<td>3</td>
<td>435.21</td>
</tr>
<tr>
<td>4</td>
<td>437.20</td>
</tr>
<tr>
<td>5</td>
<td>438.88</td>
</tr>
</tbody>
</table>

**Step 4.** Approximate the transition function \( F(X_{t-d}) \) in the STAR model with a third
order polynomial to obtain the following auxiliary autoregressive model:

\[ X_t = \alpha_0 + \alpha_1 W_t + \sum_{j=1}^{p} \alpha_{2j} X_{t-j} X_{t-d} + \sum_{j=1}^{p} \alpha_{3j} X_{t-j} X_{t-d}^2 + \sum_{j=1}^{p} \alpha_{4j} X_{t-j} X_{t-d}^3 + \epsilon_t. \] (2.18)

We then use a Lagrange multiplier (LM) type test introduced by Luukkonen et al. [29] to test the linearity against the STAR model alternative with different delay parameters. More specifically, we test

\[ H_0: \alpha_{2j} = \alpha_{3j} = \alpha_{4j} = 0. \]

The test with \( d = 2 \) results in the smallest p-value, \( 9.695 \times 10^{-8} \). Therefore, the linear model \( X_t = \alpha_0 + \alpha_1 W_t + \epsilon_t \) is not adequate, and a STAR model with the delay parameter \( d = 2 \) is more appropriate. **Step 3.** Decide between ESTAR models and LSTAR models via a series of F tests within full model (2.18)

\[ H_{01}: \alpha_{4j} = 0, \quad j = 1, 2, 3; \]
\[ H_{02}: \alpha_{3j} = 0 \mid \alpha_{4j} = 0, \quad j = 1, 2, 3; \]
\[ H_{03}: \alpha_{2j} = 0 \mid \alpha_{3j} = \alpha_{4j} = 0, \quad j = 1, 2, 3. \]

Among all the above tests, \( H_{02} \) gives the smallest p-value. Thus, according to the rule by Terasvirta [27], we choose ESTAR model of order \( p = 3 \) with delay parameter \( d = 2 \) as our final model to fit the data.

The fitted ESTAR(3) model is listed below:

\[ X_t = -870.32 + 4.42 X_{t-1} + 7.01 X_{t-2} - 1.84 X_{t-3} \]
\[ + (879.47 - 3.38 X_{t-1} - 7.37 X_{t-2} + 2.06 X_{t-3})(1 - e^{-184.62(X_{t-2} - 101.28)^2}). \]
The estimated standard deviation of the noises is 0.65. Therefore, we take $f_0$ as the density function of the normal distribution with a mean of zero and a standard deviation of 0.65. Figure 2.4 presents a histogram of residuals from the fitted model. To choose an appropriate value for $m$, we summarize ISE’s with different choices of $m$ in Table 2.5. The conventional estimator has ISE 0.0086, and the tilted density estimator with $m = 2$ has the smallest ISE, 0.0077. As such, we use the goodness-of-fit test based on the tilted density estimator with $m = 2$. The critical value for the test is 1.275, and the absolute value of the test statistic is $\hat{t}_{m=2} = 0.176$. We fail to reject the null hypothesis and conclude that the random errors in the ESTAR (3) model for the INDPRO data follow a normal distribution. Note that Figure 2.5 displays the plots of density estimators, with the prescribed density $f_0$ in red, the conventional estimator in blue, and tilted density estimators with different choices of $m$.

![Histogram of Residuals](image)

Figure 2.4: Histogram of Residuals for INDPRO Data
### Table 2.5: INDPRO Data: ISE Based on Different m

<table>
<thead>
<tr>
<th>m</th>
<th>ISE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional estimator</td>
<td>0.0086</td>
</tr>
<tr>
<td>m = 2</td>
<td>0.0077</td>
</tr>
<tr>
<td>m = 3</td>
<td>0.0109</td>
</tr>
<tr>
<td>m = 4</td>
<td>0.0113</td>
</tr>
<tr>
<td>m = 5</td>
<td>0.0113</td>
</tr>
<tr>
<td>m = 6</td>
<td>0.0112</td>
</tr>
</tbody>
</table>

Figure 2.5: Error Density Estimators for INDPRO Data

#### 2.5 Technical Details

We first present four auxiliary lemmas that will be used to establish the main result. Lemma 1, which is of independent interest, gives the limiting distribution of the goodness-of-fit test on the unknown density function with an i.i.d. setup. Please refer to Theorem 1
in chapter 1 for more details. Lemmas 5 and 6 are proved here to assist the development of Lemma 7, which evaluates the squared $L_2$ distance between $\hat{f}_n^*$, the tilted error density estimator based on residuals, and $f_n^*$, the tilted estimator based on unobservable errors.

### 2.5.1 Auxiliary Results

**Lemma 4** Under assumptions $B_1$ to $B_4$, the test statistic constructed by use of the tilted kernel density estimator $f_n^*$ from the unobservable errors $\epsilon_1, \cdots, \epsilon_n$,

$$T_n^* = n h^{1/2} \int (f_n^*(x) - (K_h*f_0)(x))^2 \, dx,$$

has the following asymptotic properties:

(i) Under the null hypothesis $H_0 : f = f_0$,

$$T_n^* - h^{-\frac{1}{2}} \int K^2(t) \, dt \overset{d}{\to} N(0, \tau^2),$$

where $\tau^2 = 2 \int f_0^2(x) \, dx \int (\int K(t)K(s-t) \, dt)^2 \, ds$.

(ii) Under a fixed alternative $H_1 : f \neq f_0$,

$$(nh)^{-\frac{1}{2}} T_n^* - n^{\frac{3}{2}} \int (K_h*f_0)^2(x) \, dx \overset{d}{\to} N(0, 4\sigma^2),$$

where $\sigma^2 = Var[(f-f_0)(X_1)]$.

**Proof.** See the proof of Theorem 1 in chapter 1.
Lemma 5  Under assumptions A1 to A3, we have the following:

\[ \sum_{i=1}^{n} (\epsilon_i - \hat{\epsilon}_i)^2 = O_p\left(n^{-\beta+1}\log\log n\right). \]  

(2.22)

Proof. By Taylor’s expansion, there exists a random quantity \( \lambda \in (0, 1) \) and \( \theta^* = \theta + \lambda(\hat{\theta} - \theta) \) such that

\[ \epsilon_i - \hat{\epsilon}_i = \alpha \hat{\theta} (X_{i-1}, X_{i-2}, \ldots, X_{i-p}) - \alpha \theta (X_{i-1}, X_{i-2}, \ldots, X_{i-p}) \]

\[ = \sum_{s=1}^{q} (\hat{\theta}_s - \theta_s) Y_{is} + \frac{1}{2} \sum_{s=1}^{q} \sum_{t=1}^{q} (\hat{\theta}_s - \theta_s)(\hat{\theta}_t - \theta_t) Z_{ist}, \]  

(2.23)

where

\[ Y_{is} = \frac{\partial}{\partial \theta_s} \alpha \theta (X_{i-1}, \ldots, X_{i-p}), \quad Z_{ist} = \frac{\partial^2}{\partial \theta_s \partial \theta_t} \alpha \theta^* (X_{i-1}, \ldots, X_{i-p}), \quad s, t = 1, \ldots, q. \]

Thus, we have

\[ (\epsilon_i - \hat{\epsilon}_i)^2 \leq 2\left\{ \sum_{s=1}^{q} (\hat{\theta}_s - \theta_s) Y_{is} \right\}^2 + \frac{1}{2} \left\{ \sum_{s=1}^{q} \sum_{t=1}^{q} (\hat{\theta}_s - \theta_s)(\hat{\theta}_t - \theta_t) Z_{ist} \right\}^2 \]

\[ \leq 2q \sum_{s=1}^{q} (\hat{\theta}_s - \theta_s)^2 Y_{is}^2 + \frac{1}{2} q^2 \sum_{s=1}^{q} \sum_{t=1}^{q} (\hat{\theta}_s - \theta_s)^2(\hat{\theta}_t - \theta_t)^2 Z_{ist}^2. \]  

(2.24)

For \( 1 \leq i \leq n \) and \( 1 \leq s, t \leq q \), we obtain from assumption A1 that

\[ E(Z_{ist}^2) \leq (EM_2^4(X_{i-1}, \ldots, X_{i-p}))^{1/2} < \infty, \]

which, together with Markov’s inequality, leads to

\[ \sum_{i=1}^{n} Z_{ist}^2 = O_p(n). \]  

(2.25)
By assumption $A3$, if $n$ is sufficiently large, we get
\[
||\hat{\theta} - \theta||_\infty \leq ||\hat{\theta} - \theta||_2 \leq C_0(n^{-1/2}(\log \log n)^{1/2}). \tag{2.26}
\]

Then, by Equations (2.24)-(2.26), we have
\[
\sum_{i=1}^{n} (\epsilon_i - \hat{\epsilon}_i)^2 \leq 2q \sum_{i=1}^{n} \sum_{s=1}^{q} (\hat{\theta}_s - \theta_s)^2 Y_{is}^2 + \frac{q^2}{2} \sum_{i=1}^{n} \sum_{s=1}^{q} (\hat{\theta}_s - \theta_s)^2 (\hat{\theta}_t - \theta_t)^2 Z_{ist}^2 \tag{2.27}
\]
\[
\leq 2C_0^2 q n^{-1} \log(\log n) \sum_{i=1}^{n} Y_{is}^2 + \frac{q^2}{2} C_0^4 n^{-2} (\log \log n)^2 \sum_{i=1}^{n} Z_{ist}^2
\]
\[
= O_p\left(n^{-1+\beta \log \log n}\right).
\]

In the above derivation, we also used assumption $A2$, $\sum_{i=1}^{n} Y_{is}^2 = O_p(n^\beta)$.

**Lemma 6** Under the assumptions of Theorem 3, for $s, t = 1, \cdots, q$ and $\theta \in U$, we have

(i) $\int \left\{ \sum_{i=1}^{n} p_i Y_{is} K'\left(\frac{x-\epsilon_i}{h}\right) \right\}^2 dx = O_p(n^{-1} h + n^{-1+\beta} h^2)$,

(ii) $\int \left\{ \sum_{i=1}^{n} p_i Y_{is}^2 K''\left(\frac{x-\epsilon_i}{h}\right) \right\}^2 dx = O_p(n^{-1} h + h^2) = O_p(h^2),$

(iii) $\int \left\{ \sum_{i=1}^{n} p_i Z_{ist} K'\left(\frac{x-\epsilon_i}{h}\right) \right\}^2 dx = O_p(n^{-1} h + h^2) = O_p(h^2),$

(iv) $\int \left\{ \sum_{i=1}^{n} p_i Z_{ist}^2 K''\left(\frac{x-\epsilon_i}{h}\right) \right\}^2 dx = O_p(n^{-1} h + h^2) = O_p(h^2)$.

**Proof** Using the result that
\[
\left( \sum_{i=1}^{k} a_i \right)^2 \leq k \sum_{i=1}^{k} a_i^2, \quad k \geq 2, \tag{2.28}
\]
we have

\[
\begin{align*}
& \int \left\{ \sum_{i=1}^{n} p_i Y_{is} K'\left( \frac{x - \epsilon_i}{h} \right) \right\}^2 dx \\
\leq & \ 2 \int \left\{ \sum_{i=1}^{n} p_i Y_{is} \left[ K'\left( \frac{x - \epsilon_i}{h} \right) - E K'\left( \frac{x - \epsilon_i}{h} \right) \right] \right\}^2 dx + 2 \int \left\{ \sum_{i=1}^{n} p_i Y_{is} E K'\left( \frac{x - \epsilon_i}{h} \right) \right\}^2 dx.
\end{align*}
\]

Note that

\[
E \int \left\{ \sum_{i=1}^{n} p_i Y_{is} \left[ K'\left( \frac{x - \epsilon_i}{h} \right) - E K'\left( \frac{x - \epsilon_i}{h} \right) \right] \right\}^2 dx
= \int E \left\{ \sum_{i=1}^{n} p_i^2 Y_{is}^2 \left[ K'\left( \frac{x - \epsilon_i}{h} \right) - E K'\left( \frac{x - \epsilon_i}{h} \right) \right] \right\}^2 dx
+ 2 \int E \sum_{i<j} p_i Y_{is} Y_{js} \left[ K'\left( \frac{x - \epsilon_i}{h} \right) - E K'\left( \frac{x - \epsilon_i}{h} \right) \right] \left[ K'\left( \frac{x - \epsilon_j}{h} \right) - E K'\left( \frac{x - \epsilon_j}{h} \right) \right] dx.
\]

The model assumption that \( X_{i-1}, \ldots, X_{i-p} \) are independent of \( \epsilon_i \) for \( i \geq 1 \) implies that \( \epsilon_i \) is independent of \( Y_{is} \), as the latter is a function of \( X_{i-1}, \ldots, X_{i-p} \). Moreover, for \( j > i \), \( \epsilon_j \) is independent of \( \epsilon_i; X_{i-1}, \ldots, X_{i-p}; \) and \( X_{j-1}, \ldots, X_{j-p} \). Therefore, the last term in the above equation is 0, which leads to

\[
E \int \left\{ \sum_{i=1}^{n} p_i Y_{is} \left[ K'\left( \frac{x - \epsilon_i}{h} \right) - E K'\left( \frac{x - \epsilon_i}{h} \right) \right] \right\}^2 dx
= \int \left\{ \sum_{i=1}^{n} p_i^2 E(Y_{is}^2) \cdot E \left[ K'\left( \frac{x - \epsilon_i}{h} \right) - E K'\left( \frac{x - \epsilon_i}{h} \right) \right]^2 \right\} dx
\leq \int \left\{ \sum_{i=1}^{n} p_i^2 E(Y_{is}^2) \cdot E \left[ K'\left( \frac{x - \epsilon_i}{h} \right) \right]^2 \right\} dx
= \sum_{i=1}^{n} p_i^2 E(Y_{is}^2) \int E \left[ K'\left( \frac{x - \epsilon_i}{h} \right) \right]^2 dx.
\]

Assumption A1 implies that

\[
EY_{1s}^2 \leq EM_1^2 \leq (EM_1^4)^{1/2} < \infty, \quad s = 1, \ldots, q.
\]
Next, by assumption \( B4 \), Equations (2.30)-(2.32), and Lemma 4.2 of Cheng and Sun [10], we have

\[
E \int \left\{ \sum_{i=1}^{n} p_i Y_{i} \left[K'\left(\frac{x-\epsilon_i}{h}\right) - EK'\left(\frac{x-\epsilon_i}{h}\right)\right] \right\}^2 dx
\]

\[\leq C_2^2 n^{-1} (EM_1^4)^{1/2} \int E[K'\left(\frac{x-\epsilon_1}{h}\right)]^2 dx\]

\[= O(n^{-1} h),\]

which, together with Markov’s inequality, yields

\[
\int \left\{ \sum_{i=1}^{n} p_i Y_{i} \left[K'\left(\frac{x-\epsilon_i}{h}\right) - EK'\left(\frac{x-\epsilon_i}{h}\right)\right] \right\}^2 dx = O_p(n^{-1} h). \tag{2.34}
\]

Assumption \( A2 \), assumption \( B4 \), and Lemma 4.2 of Sun and Cheng [10] produce

\[
\int \left\{ \sum_{i=1}^{n} p_i Y_{i} EK'\left(\frac{x-\epsilon_i}{h}\right) \right\}^2 dx \leq \left\{ \sum_{i=1}^{n} p_i^2 \right\} \left\{ \sum_{i=1}^{n} Y_{i}^2 \right\} \int \left\{ EK'\left(\frac{x-\epsilon_1}{h}\right) \right\}^2 dx = O_p(n^{-1+\beta} h^2). \tag{2.35}
\]

Thus, result (i) of Lemma 6 follows from Equations (2.29), (2.34), and (2.35).

Similarly, we can show that

\[
\int \left\{ \sum_{i=1}^{n} p_i Z_{ist} \left[K'\left(\frac{x-\epsilon_i}{h}\right) - EK'\left(\frac{x-\epsilon_i}{h}\right)\right] \right\}^2 dx = O_p(n^{-1} h), \tag{2.36}
\]
and, using assumption $A1$ and Lemma 4.2 of Cheng and Sun [10],

\[
\int \left\{ \sum_{i=1}^{n} p_i Z_{ist} EK'\left(\frac{x - \epsilon_i}{h}\right) \right\}^2 dx 
\leq \left\{ \sum_{i=1}^{n} p_i^2 \right\} \left\{ \sum_{i=1}^{n} Z_{ist}^2 \right\} \int \left\{ EK'\left(\frac{x - \epsilon_1}{h}\right) \right\}^2 dx 
\leq C_2^2 n^{-1} \cdot nM_2^2 \cdot O(h^2)
\]

\[= O_p(h^2).\]

Finally,

\[
\int \left\{ \sum_{i=1}^{n} p_i Z_{ist} K'\left(\frac{x - \epsilon_i}{h}\right) \right\}^2 dx 
\leq 2 \int \left\{ \sum_{i=1}^{n} p_i Z_{ist} \left[ K'\left(\frac{x - \epsilon_i}{h}\right) - EK'\left(\frac{x - \epsilon_i}{h}\right) \right] \right\}^2 dx + 2 \int \left\{ \sum_{i=1}^{n} p_i Z_{ist} EK'\left(\frac{x - \epsilon_i}{h}\right) \right\}^2 dx 
= O_p(n^{-1} h + h^2).
\]

Lemma 6 (iii) is proved. Results (ii) and (iv) can be proved by following the same framework used in (iii). We complete the proof of Lemma 6.

**Lemma 7** Suppose that assumptions $A1$ to $A3$ and $B1$ to $B3$ hold. Then, if $n^{1-\beta} h^2 / (\log \log n) \to \infty$, we have

\[
\int (\hat{f}_n^*(x) - f_n^*(x))^2 dx = O_p\left(n^{-2+\beta} h^{-2} \log \log n + n^{-2} h^{-4} (\log \log n)^2\right). \quad (2.39)
\]
Proof. By Taylor’s expansion,

\[
\int (f_n^* (x) - f_n^*(x))^2 \, dx \\
= \int \left\{ \sum_{i=1}^{n} p_i (K_h(x - \hat{\epsilon}_i) - K_h(x - \epsilon_i)) \right\}^2 \, dx \\
= \int \left\{ \frac{1}{h} \sum_{i=1}^{n} p_i \left( \frac{\epsilon_i - \hat{\epsilon}_i}{h} K' \left( \frac{x - \epsilon_i}{h} \right) + \frac{(\epsilon_i - \hat{\epsilon}_i)^2}{2h^2} K''(\gamma_{ix}) \right) \right\}^2 \, dx \\
\leq \frac{2}{h^4} \int \left\{ \sum_{i=1}^{n} p_i (\epsilon_i - \hat{\epsilon}_i) K' \left( \frac{x - \epsilon_i}{h} \right) \right\}^2 \, dx + \frac{1}{2h^6} \int \left\{ \sum_{i=1}^{n} p_i (\epsilon_i - \hat{\epsilon}_i)^2 K''(\gamma_{ix}) \right\}^2 \, dx \\
= \frac{2}{h^4} T_{1n} + \frac{1}{2h^6} T_{2n},
\]

where \( \gamma_{ix} \) is a random quantity between \( \frac{x - \epsilon_i}{h} \) and \( \frac{x - \hat{\epsilon}_i}{h} \).

First, we evaluate \( T_{1n} \). By Equation (2.23),

\[
T_{1n} = \int \left\{ \sum_{i=1}^{n} p_i (\epsilon_i - \hat{\epsilon}_i) K' \left( \frac{x - \epsilon_i}{h} \right) \right\}^2 \, dx \\
= \int \left\{ \sum_{i=1}^{n} p_i \left[ \sum_{s=1}^{q} (\hat{\theta}_s - \theta_s) Y_{is} + \frac{1}{2} \sum_{i=1}^{q} (\hat{\theta}_s - \theta_s)(\hat{\theta}_t - \theta_t) Z_{ist} \right] K' \left( \frac{x - \epsilon_i}{h} \right) \right\}^2 \, dx \\
\leq 2 \int \left\{ \sum_{i=1}^{n} p_i \left[ \sum_{s=1}^{q} (\hat{\theta}_s - \theta_s) Y_{is} \right] K' \left( \frac{x - \epsilon_i}{h} \right) \right\}^2 \, dx \\
+ \frac{1}{2} \int \left\{ \sum_{i=1}^{n} p_i \left[ \sum_{i=1}^{q} (\hat{\theta}_s - \theta_s)(\hat{\theta}_t - \theta_t) Z_{ist} \right] K' \left( \frac{x - \epsilon_i}{h} \right) \right\}^2 \, dx \\
\leq 2q \sum_{s=1}^{q} (\hat{\theta}_s - \theta_s)^2 \int \left\{ \sum_{i=1}^{n} p_i Y_{is} K' \left( \frac{x - \epsilon_i}{h} \right) \right\}^2 \, dx \\
+ \frac{q^2}{2} \sum_{s=1}^{q} (\hat{\theta}_s - \theta_s)^2 (\hat{\theta}_t - \theta_t)^2 \int \left\{ \sum_{i=1}^{n} p_i Z_{ist} K' \left( \frac{x - \epsilon_i}{h} \right) \right\}^2 \, dx,
\]
which, together with Lemma 6 and assumption A3, yields

\[
T_{1n} = \int \left\{ \sum_{i=1}^{n} p_i(\epsilon_i - \hat{\epsilon}_i)K'(\frac{x - \hat{\epsilon}_i}{h}) \right\}^2 dx
\]

\[
\leq 2q \sum_{s=1}^{q} (\hat{\theta}_s - \theta_s)^2 \cdot O_p(n^{-1}h + n^{-1+\beta}h^2) + \frac{q^2}{2} \sum_{s=1}^{q} \sum_{i=1}^{q} (\hat{\theta}_s - \theta_s)^2(\hat{\theta}_i - \theta_i)^2 \cdot O_p(h^2)
\]

\[
\leq 2qC_0n^{-1} \log \log n \cdot O_p(n^{-1}h + n^{-1+\beta}h^2) + \frac{q^2}{2} C_0^4 n^{-2}(\log \log n)^2 \cdot O_p(h^2)
\]

\[
= O_p((n^{-2}h + n^{-2+\beta}h^2) \log \log n).
\]

(2.42)

Next, we consider \(T_{2n}\).

\[
T_{2n} = \int \left\{ \sum_{i=1}^{n} p_i(\epsilon_i - \hat{\epsilon}_i)^2K''(\gamma_{ix}) \right\}^2 dx
\]

\[
= \int \left\{ \sum_{i=1}^{n} p_i[\sum_{s=1}^{q} (\hat{\theta}_s - \theta_s)Y_{is} + \frac{1}{2}\sum_{s=1}^{q} \sum_{i=1}^{q} (\hat{\theta}_s - \theta_s)(\hat{\theta}_i - \theta_i)Z_{is}]^2K''(\gamma_{ix}) \right\}^2 dx
\]

\[
\leq \int \left\{ \sum_{i=1}^{n} p_i[2q \sum_{s=1}^{q} (\hat{\theta}_s - \theta_s)^2Y_{is}^2 + \frac{q^2}{2} \sum_{s=1}^{q} \sum_{i=1}^{q} (\hat{\theta}_s - \theta_s)^2(\hat{\theta}_i - \theta_i)^2Z_{is}]^2K''(\gamma_{ix}) \right\}^2 dx
\]

\[
\leq 8q^2 \int \left\{ \sum_{s=1}^{q} (\hat{\theta}_s - \theta_s)^2 \sum_{i=1}^{n} p_iY_{is}^2K''(\gamma_{ix}) \right\}^2 dx
\]

\[
+ \frac{q^4}{2} \int \left\{ \sum_{s=1}^{q} \sum_{i=1}^{q} (\hat{\theta}_s - \theta_s)^2(\hat{\theta}_i - \theta_i)^2Z_{is}^2]^2K''(\gamma_{ix}) \right\}^2 dx
\]

\[
\leq 8q^3 \sum_{s=1}^{q} (\hat{\theta}_s - \theta_s)^4 \int \left\{ \sum_{i=1}^{n} p_iY_{is}^2K''(\gamma_{ix}) \right\}^2 dx
\]

\[
+ \frac{q^6}{2} \sum_{s=1}^{q} \sum_{i=1}^{q} (\hat{\theta}_s - \theta_s)^4(\hat{\theta}_i - \theta_i)^4 \int \left\{ \sum_{i=1}^{n} p_iZ_{is}^2K''(\gamma_{ix}) \right\}^2 dx.
\]

(2.43)

By Lemmas 5 and 6, and assumptions A1, B2, and B3, when \(n^{1-\beta}h^2/(\log \log n) \to \infty\), we
\[
\int \left\{ \sum_{i=1}^{n} p_i Y_i^2 | K''(\gamma_{i,x})| \right\}^2 dx \\
\leq 2 \int \left\{ \sum_{i=1}^{n} p_i Y_i^2 \left| K''\left(\frac{x - \epsilon_i}{h}\right)\right| \right\}^2 dx + 2 \int \left\{ \sum_{i=1}^{n} p_i Y_i^2 \left| K''(\gamma_{i,x}) - K''\left(\frac{x - \epsilon_i}{h}\right)\right| \right\}^2 dx \\
\leq O_p(h^2) + 2 \sum_{i=1}^{n} p_i^2 Y_i^4 \left\{ \sum_{i=1}^{n} \int \left| K''(\gamma_{i,x}) - K''\left(\frac{x - \epsilon_i}{h}\right)\right|^2 dx \right\} \\
= O_p(h^2) + O_p(n^{-1}) \cdot O\left(\frac{1}{h^2} \sum_{i=1}^{n} (\epsilon_i - \hat{\epsilon}_i)^2 \right) \\
= O_p(h^2) + O_p(n^{-1}) \cdot O_p\left(h^{-2} n^{-1+\beta \log \log n} \right) \\
= O_p\left(h^2 + n^{-2+\beta} h^{-2} \log \log n \right) \\
= O_p(h^2). 
\]

Likewise, by following the same framework, we show that

\[
\int \left\{ \sum_{i=1}^{n} p_i Z_{is}^2 | K''(\gamma_{i,x})| \right\}^2 dx = O_p(h^2). 
\]

By assumption A3 and Equations (2.43)-(2.45), we obtain

\[
T_{2n} = \int \left\{ \sum_{i=1}^{n} p_i (\epsilon_i - \hat{\epsilon}_i)^2 K''(\gamma_{i,x}) \right\}^2 dx \\
\leq O(n^{-2}(\log \log n)^2) \cdot O_p(h^2) + O(n^{-4}(\log \log n)^4) \cdot O_p(h^2) \\
= O_p\left(n^{-2} h^2 (\log \log n)^2 \right). 
\]

Combining Equations (2.40), (2.42), and (2.46) gives

\[
\int (\hat{f}_n^*(x) - f_n^*(x))^2 dx \\
\leq \frac{2}{h^4} \cdot O_p\left((n^{-2} h + n^{-2+\beta} h^2) \log \log n \right) + \frac{1}{2h^6} \cdot O_p\left(n^{-2} h^2 (\log \log n)^2 \right) \\
= O_p\left(n^{-2+\beta} h^{-2} \log \log n + n^{-2} h^{-4} (\log \log n)^2 \right). 
\]
Lemma 7 is proved.

2.5.2 Proof of Theorem 3

Note that Lemma 4 holds if all assumptions of Theorem 3 are satisfied. Thus, under $H_0: f = f_0$, we have

$$\int (f_n^*(x) - Kh \ast f_0(x))^2 dx = O_p\left(\frac{1}{nh}\right), \quad (2.48)$$

but under $H_1: f \neq f_0$, we have

$$\int (f_n^*(x) - Kh \ast f_0(x))^2 dx = O_p(1). \quad (2.49)$$

Proof of Part (i).

Under $H_0: f = f_0$, by Equation (2.48), when $n^{1-\beta} h^2 / (\log \log n) \to \infty$, we have

$$|\hat{T}_n^* - T_n^*| \leq nh^{1/2} \int (\hat{f}_n^*(x) - f_n^*(x))^2 dx$$

$$+ 2nh^{1/2} \left( \int (\hat{f}_n^*(x) - f_n^*(x))^2 dx \right)^{1/2} \left( \int (f_n^*(x) - Kh \ast f_0(x))^2 dx \right)^{1/2}$$

$$= nh^{1/2} \cdot O_p(n^{-2+\beta} h^{-2} \log \log n + n^{-2} h^{-4} (\log \log n)^2)$$

$$+ 2nh^{1/2} \cdot O_p(n^{-2+\beta} h^{-2} \log \log n + n^{-2} h^{-4} (\log \log n)^2)^{1/2} \cdot O_p(n^{-1} h^{-1})^{1/2}$$

$$= O_p(n^{-1+\beta} h^{-3/2} \log \log n + n^{-1} h^{-7/2} (\log \log n)^2)$$

$$+ O_p(n^{-1/2+\beta/2} h^{-1} (\log \log n)^{1/2} + n^{-1/2} h^{-2} \log \log n)$$

$$= o_p(1). \quad (2.50)$$

Therefore, $\hat{T}_n^*$ has the same limiting distribution as $T_n^*$, and, hence, Theorem 3 (i) follows from Lemma 4.
Proof of Part(ii)

Under $H_1: f \neq f_0$, by Equation (2.49), if $n^{1-\beta} h^2 / (\log \log n) \to \infty$, we have

$$|\hat{T}_n^* - T_n^*| \leq nh^{1/2} \int (\hat{f}_n^*(x) - f_n^*(x))^2 \, dx + 2nh^{1/2} \left( \int (\hat{f}_n^*(x) - f_n^*(x))^2 \, dx \right)^{1/2} \left( \int (f_n^*(x) - K_h * f_0(x))^2 \, dx \right)^{1/2}$$

$$= nh^{1/2} \cdot O_p(n^{-2+\beta} h^{-2} \log \log n + n^{-2} h^{-4} (\log \log n)^2)$$

$$+ 2nh^{1/2} \cdot O_p(n^{-2+\beta} h^{-2} \log \log n + n^{-2} h^{-4} (\log \log n)^2)^{1/2} \cdot O_p(1)$$

$$= O_p(n^{\beta/2} h^{-1/2} (\log \log n)^{1/2} + h^{-3/2} \log \log n).$$

Thus,

$$(nh)^{-1/2} |\hat{T}_n^* - T_n^*| = n^{-1/2} h^{-1/2} \cdot O_p(n^{\beta/2} h^{-1/2} (\log \log n)^{1/2} + h^{-3/2} \log \log n)$$

$$= O_p(n^{-1/2+\beta/2} h^{-1} (\log \log n)^{1/2} + n^{-1/2} h^{-2} \log \log n)$$

$$= o_p(1).$$

This implies that $(nh)^{-1/2} \hat{T}_n^*$ has the same asymptotic distribution as $(nh)^{-1/2} T_n^*$, which, together with Lemma 4, leads to Part (ii) of Theorem 3. We complete the proof of the main result in this study.

2.6 Discussion

In this study, we propose a new goodness-of-fit test on the error density function in nonlinear autoregressive models and show that its test statistic based on the integrated squared loss has an asymptotic normal distribution. Further investigation of the rate of convergence is desirable. In our numerical studies, we used a trial-and-error method to select $m$, the number of distinct probability weights, to construct the tilted density estimator. Theoretical development on the optimal choices of $h$, $m$, and probability
weights $q_1, \ldots, q_m$ warrants further investigation.
Deep Merged Survival Analysis on Cancer Study Using Multiple Types of Bioinformatic Data

3.1 Introduction

Cancer is a prevalent and fatal disease. Many people are suffering from cancers and most patients cannot be fully cured under current medical conditions so that the goal is to lengthen the survival time and improve life quality.

Cancer survival prediction based on the patients' historical information, current situation, treatment plan and related information can help physicians to make decisions on optimal and personalized treatment, and provide improved prognosis and monitoring of disease progression. With rapid development in bioinformatic and biomedical technologies, massive heterogeneous data have been collected and accumulated from patients. In addition to the patients' clinical data, more complex bioinformatic data have been collected including the patient's whole-genome genotype data, gene expression data, epigenetic methylation data, copy number data, whole-slide image data, mobile monitoring data, and electronic health record data.

In the era of big data, massive complex heterogeneous data sets have been collected.
Whenever possible, multiple data sets are preferred in statistical analysis to maximize the use of information. In general, the use of more data is believed to have better performance with appropriate methods. However, data collected from multiple sources using different technologies can be complex and heterogeneous. A patient’s traditional clinical data have tens of variables at most which can be simply treated as predictors in a regression analysis. A patient’s whole-genome genotype data can include 10 K to 1 M genetic variables. A patient’s methylation data can include 1 K to 10 K epigenetic variables. A patient’s whole slide image data can include multiple whole slide images of tissue samples with each image recorded as a matrix. A patient’s mobile monitoring of health data can include the monitoring of the patient’s health status and behaviours for every second, i.e. nearly continuous-time monitoring.

These new complex heterogeneous data sets pose challenges to statistical methods. These data should be considered to be the benefit of big data era, not the curse to the researchers. It is crucial to develop statistical methods to better analyze recent-emerged data sets with a new data format such as electronic health record data, image and video data, sequencing data. In addition, it is also crucial to develop methods which can appropriately handle multiple complex heterogeneous data sets.

Data integration is the practice of consolidating data from disparate sources into a single database with ultimate goal of facilitating data access. For cancer patients, heterogeneous data sets from multiple sources have been well integrated. For example, The Cancer Genome Atlas Program (TCGA), have collected over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data for cancer studies and still remains publicly available for anyone in the research community to use. (https://www.cancer.gov/tcga). These data (including clinical, molecular, and imaging data) are well integrated whereas direct combine the variables in a regression model may not work since data are heterogeneous. For example, suppose each patient has 10 clinical variables, 1,000 gene expression variables, and a histological image of tissue sample
with the resolution of 1,000 × 1,000 pixels. Direct concatenation of these variables will obtain a vector of \(10 + 1,000 + 1,000 \times 1,000 = 1,001,010\) variables. It is infeasible to conduct a regression using 1,001,010 predictors as a vector, even for a sparse regression using \(L_1\) penalty. Direct use of modern machine learning methods such as random forest, support vector machine, k-nearest neighbors, and neural networks are expected to have poor performance due to the following two statements. The first statement is that 1,001,010 predictors are too many for many machine learning methods such as random forest and support vector machine. The second statement is that direct concatenation of data from multiple sources ignored heterogeneity in data sets, which makes clinical variables only account for the proportion of \(1/100,101\) in the concatenated vector of predictors, so that machine learning methods taking the whole concatenated vector as the input will ignore the clinical variables.

Instead of direct concatenations of predictors, appropriate methods making use of multiple sources of heterogeneous data can be proposed. Our proposed method is a deep merged neural network, which is a hierarchical modeling of data to handle a large dimension of predictors. To address the issue of integrating multiple sources of data, we use a state-of-the-art model architecture. A chain of concatenated layers is used to handle each data set. Then, the upper layers of these chains are merged together to form a “merge-layer”, and then fully connected layers (FNN) are built upon the merge layer. This state-of-art deep learning network has the capacity of handling multiple heterogeneous data sets, and can combine the information from multiple data sets at an upper level so that the information from multiple data sets will be treated equally.

The use of neural networks with merge layers is originally proposed for regression analysis of continuous responses and classification analysis of binary responses. However, we expect this architecture can be used to analyze other types of responses with adaptations such as specifying a cost function for survival outcome. Then, we intend to study the performance of this model architecture (deep survival neural network plus
merge layer, i.e. deep merged survival neural network) in cancer survival analysis based on multiple heterogeneous data sets.

Although there have been multiple survival studies in cancer survival analysis in the literature, most of these studies use only one type or a few types of bioinformatics data, such as mRNA, micro-RNA, and copy number variation (CNV). In this study, we intend to compare our deep survival neural network model with merged layers with the literature methods of survival analysis to better understand the methodology of cancer survival analysis based on multiple data sets.

The remaining of this article is organized as follow: In Section 3.1, literature review is surveyed. In Section 3.2, the method of deep survival neural networks with merge layer for survival analysis is proposed. In Section 3.3, simulation study is presented. In Section 3.4, real data analysis for breast cancer survival prediction is investigated. Section 3.5 provides the conclusion and discussions.

3.1.1 Literature Review

Survival analysis is a collection of statistical analyses that attempt to model data with the duration of time until to one or more events happen, such as death event and disease occurrence event. For some instances, the events of interest do not occur within the observation time. This is known as censoring. Based on the reason of censoring occurrence, there are three types censoring (cf. Lee and Wang [30], right-censoring (observed survival time \( \leq \) true survival time), left-censoring (observed survival time \( \geq \) true survival time) and interval censoring (events occur during a certain interval). In this study, we only consider right censoring, which is most common scenario in practical problems (Marubini and Valsecchi [31]). The one of the primary goals of survival analysis is to predict survival probability, such as survival function (the probability of the time to event is not earlier than a specific time) (Lee and Wang [30]), and hazard function (the likelihood of the event occurring at a specific time given that no event occurring before) (Dunn and...
Traditional survival analysis and machine learning methods have been studied by many researchers.

Three types of classical statistical methods are widely used to estimate the survival probability, non-parametric, semi-parametric and parametric methods. Kaplan-Meier (KM) (cf. Kaplan and Meier [33]) method is a popular non-parametric method, while the main disadvantage is that as a univariate method, it cannot use all covariates to predict the survival probability. In addition, parametric methods are efficient and accurate approaches under the assumption of the time to event following a particular distribution but in practice, the distribution is unknown. For semi-parametric methods, Cox proportional hazards model (CoxPH) (Cox [34]) is a widely used survival analysis method, which is a hybrid of parametric and non-parametric method (Equation (3.1)). Unlike non-parametric and parametric method, CoxPH involves additional covariates and the underlying distribution of time to the event of interest need not to be specified. The hazard function \( \lambda(t, X_i) \) in CoxPH is defined as,

\[
\lambda(t, X_i) = \lambda_0(t) e^{X_i \beta},
\]

(3.1)

\[
X_i \beta = \sum_{j=1}^{p} \beta_j X_{ij}, \quad X_i \in \mathbb{R}^p,
\]

(3.2)

where \( X_i \) is a vector of covariates and \( \beta \) is coefficient vector for corresponding covariates. The nonparametric part of the model is the baseline hazard function \( \lambda_0(t) \), which an arbitrary non-negative function of time and the parametric part is \( X_i \beta \). Cox [35] proposed to use the partial likelihood to estimate parameters in CoxPH.

As more data collecting, the number of covariates \( p \) could exceed the number of instances \( n \). It is challenging to involve all features, so essential features could be selected via penalty functions, such as Lasso (Tibshirani [36]). Tibshirani [37] applies penalized log-partial likelihood with with penalty, \( l_1 \) regularier \( (\lambda \sum_{p=1}^{p} |\beta_p|) \). Two assumptions for CoxPH are that the hazard functions are proportion over time and the there
are linear relationship between log hazard and each covariate (Equation (3.2)). However, the assumptions could not hold because of a violation of some features and also be over-simplistic in many real applications, non-linear function of \( h(X) \) is needed. Widely used methods allowing a non-linear function of \( h(X) \) includes random forest survival model (Ishwaran et al. [38]), and neural network survival model which model \( h(X) \) as a highly hierarchical model via the use of a sequence of multiple concatenated layers to transform input \( X \) layer by layer to \( h(X) \). The neural network approximation allows non-linearity and flexibility of \( h(X) \) which has shown success in many real applications especially for high-dimensional \( X \).

\[
\begin{align*}
a^{[1]} &= f(W^{[1]} X_i + b^{[2]}) \\
a^{[2]} &= f(W^{[2]} a^{[1]} + b^{[2]}) \\
& \vdots \\
a^{[r-1]} &= f(W^{[r-1]} a^{[r-2]} + b^{[r-1]}) \\
\end{align*}
\]

\[h(X_i) = W^{[r]} a^{[r-1]} + b^{[r]}\]

where \( f \) is non-linear activation functions, \( W \) is weights matrix, \( b \) is bias vector and \( X_i \) is a vector of covariates for an individual.

Nowadays, many researchers study on deep survival neural networks, which is based on non-linear risk function instead of linear Cox regression model. Faraggi and Simon [39] is the first to extend CoxPH to non-linear neural network with a single hidden layer to estimate individual's risk of death. Next, to assess prognostic factor such as estrogen (ER) and progesterone (PgR) receptors on the study of recurrence of breast cancer, Mariani et al. [40] applied both CoxPH and neural network proposed in Faraggi and Simon's work [39] and showed that the predictive value for both methods was relative low. Later, Baesens et al. [41] compared the performance CoxPH with neural network survival analysis in personal loan data and showed that the neural network survival analysis did not
outperform the CoxPH.

With the rapid development in neural network in the past decade, many researchers proposed innovative neural network methods on survival analysis by estimating the risk score of the diseases. Ranganath et al. [42] applied deep survival analysis on the electronic health record (EHR) data to predict the risk score of coronary heart disease and showed the model outperform CoxPH. Ching et al. [43] proposed deep survival analysis approaches to estimate disease risk based on a single bioinformatic data from 10 TCGA RNA-Seq data sets. Katzman et al. [44] introduced DeepSurv model and applied in two real clinical studies, the Worcester Heart Attack Study (WHAS) and the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC). Later, multiple tumor datasets were considered by many researchers in survival analysis. To identifying robust survival subgroups of hepatocellular carcinoma (HCC), Chaudhary et al. [45] used auto-encoder to select features for TCGA liver cancer study based on RNA-seq, miRNA and methylation. Poirion et al. [46] applied auto-encoder to produce new features for risk stratification in bladder cancer. To integrate multiple tumor datasets, Sun et al. [47] applied three deep neural network on gene expression, copy number and clinical data for breast cancer to extract features for each data type and then combine the reduced features with a weighted linear function. Huang et al. [48] introduced a procedure, which was to apply lmQCM algorithm and neural network to reduce features for mRNA and micro-RNA (miRNA), and then to combine some clinical data to predict the risk score based on CoxPH.

3.2 Materials and Methods

The methods for cancer analysis under consideration include regularized Cox regression, random survival analysis, deep survival analysis methods and etc. Researchers believe that deep neural network outperforms among other survival analysis methods.
However the performance of deep survival analysis with more multiple types of bioinformatics datasets need to be investigated. To our best knowledge, only neural networks with model architecture of one sequence of layers are adopted in the literature. There are no merge layers in the model architecture. Data from multiple sources are directly combined together to form a single data set, which is then fed into the neural network with only a sequence of layers. This literature method combines data at the input level, but not at the higher level as our deep survival neural network with merge layers.

Because different bioinformatics datasets can be very heterogeneous, including quite different number of predictors, we propose to use merge layer to combine the variables from different data sets at the higher level in neural network structure. This allows us to extract features first and then combine features at the higher level in neural network structure. Therefore, instead of combining different bioinformatics data at the beginning input level, we introduce a general hierarchical deep survival analysis, which is flexible to reduce features on each bioinformatics data first and then concatenate all the reduced bioinformatics data in the analysis.

### 3.2.1 Architecture of Deep Merged Survival Neural Networks

We propose a new deep survival neural network model, which has concatenated layer in the architecture to integrate the information from each data group. Then, we apply fully connected neural network. In addition, we use adaptive moment estimation (Adam) optimizer, which is gradient-based optimization of stochastic objective functions (Kingma and Ba [49]) and rectified linear unit (ReLU) \( f(x) = \max(0, x) \) as the activation function in the neural network layers. Figure 3.1 shows the architecture of our proposed model.
Figure 3.1: Architecture of General Multi-bioinformatic Survival Analysis Network.

The Figure shows the procedure of our method. First, we clean source data sets via data preprocessing and then use the deep merged survival neural network to predict risk score.

First, we apply data preprocessing, such as normalization, imputation of missing value and feature filtering based on feature size, on each type of data set. Next, each data set is fed into an individual neural network. We concatenate the hidden features from these individual neural network and then feed the concatenated features into a fully connected neural network to predict the response. This architecture (multiple individual neural networks for each individual data set, and then merge the output from each individual neural network as the input for another neural network) allows us to combine information from multiple data sets. Compared with the use of only one single neural network for all data as the inputs, our model allow bioinformatic data contribute to hazard ratios in a relatively equal scale at survival analysis.
3.2.2 Objective Function

The input variable is a specific data set format \((T_i, \delta_i, X_i)\) for individual \(i\), where 
\(T_i\) is observed duration, 
\(\delta_i\) is the event/censoring indicator (1= event, 0=censor),
\(X_i\) is a \(p\)-dimensional covariate vector.

To fit Cox's proportional hazards model (3.1), log partial likelihood function (3.1) is applied (Cox [35]) to estimate the coefficients of CoxPH.

\[
l(\beta) = \log L(\beta) = \sum_{i=1}^{n} \delta_i [x_i^T \beta - \log \{ \sum_{j \in R_i} \exp(x_j^T \beta) \}] 
\]

where \(R(T_i) = \{ j : T_j \geq T_i \} \) is the risk set including the individuals who are 'at risk' for failure at time \(T_i\).

The deep survival neural network model is to predict log-risk function \(h_i\). Therefore, deep survival model uses negative log partial likelihood function as the loss function, i.e.,

\[
l_{\text{cox}} = -\frac{1}{N} \sum_{i, \delta_i = 1} \{ \hat{h}_i - \log \{ \sum_{j : T_j \geq T_i} \exp(\hat{h}_j) \} \}, 
\]

where \(\hat{h}_i\) is the predicted risk score, \(\delta_i = 1\) indicates the occurrence of death events for individual \(i\), \(N\) is total number of death events, and \(T_i\) is survival time for individual \(i\).

3.2.3 Evaluation Metrics

A range of evaluation metrics have been used for model performance evaluation to measure predicted hazard ratio and true survival time. Among these evaluation metrics, the concordance index, or C-index, is one of the most used criteria to evaluate the prediction accuracy for survival analysis (Steck et al. [50]).

C-index represents the global assessment of the model's discrimination power, which refers to the model's ability to correctly provide a reliable ranking of the survival times.
based on the individual risk score. C-index can be computed as the ratio of the number of concordance pairs to the number of comparable pairs using the formula:

\[
C\text{-index} = \frac{\sum_{i \neq j} (1_{T_j < T_i})(1_{h_j > h_i})\delta_j}{\sum_{i \neq j} (1_{T_j < T_i})\delta_j}
\]  

(3.6)

where \(h_i\) is the risk score for observation \(i\) and \(\delta_j\) is the event indicator for observation \(j\). The index function \(1_{T_j < T_i} = 1\) if \(T_j < T_i\) and 0 otherwise. The index function \(1_{h_j > h_i} = 1\) if \(h_j > h_i\) and 0 otherwise. C-index with value near 0.5 means the model is ineffective, and C-index with value over 0.7 indicates a good model.

Comparable pairs are all pairs of patients \(i\) and \(j\), where \(i \neq j\) except the following two situations.

(1). If both \(T_i\) and \(T_j\) are censored, we do not consider this pair in computation due to unobserved time of event occurrence.

(2). If there is one censored data, e.g. \(T_i\) is observed and \(T_j\) is censored, meanwhile, \(T_j < T_i\), we do not consider this pair because we do not know who got disease or was dead first.

Concordant pairs can belong to the following two situations.

(1). If \(T_i\) and \(T_j\) are not censored and \(h_i > h_j\), \(T_i < T_j\), this pair is a concordant pair.

(2). If there is one censored data, e.g. \(T_i\) is observed and \(T_j\) is censored, this pair is a concordant pair when \(T_j > T_i\), \(h_i > h_j\).

In this study, we evaluate and compare the performance of our method with literature methods using C-index criterion via 5-fold cross validation.

### 3.3 Simulation Studies

We conduct simulation studies to evaluate the performance of different methods in cancer survival analysis. We consider a range of sample sizes \((n = 300, 500, 1000, 2000)\) and
a range of feature sizes \( p = 500, 1000, 2000 \) in our simulation studies. We explore four different scenarios in our simulation studies.

Two different bioinformatics data sets, i.e. gene expression data and methylation data, are simulated via software R.4.0.2 with package RUVcorr (simulateGEdata function) and package methylkit (dataSim function). They are considered to be predictors \( X \) in modeling cancer survivals. Then we simulated survival responses \( Y \) under a range of scenarios using different models. We carefully design different scenarios so that we can evaluate model performance under different scenarios to mimic different real situations. We use simsurv R package [51] to generate survival data with censor rate controlled to be around 50% and different true models specified in simulations for different scenarios.

After we simulated survival data, we applied our method and literature methods to analyze the data. The concordance index is used as the evaluation metric in model performance evaluation. For each scenario, we conducted 300 simulations. In each simulation, we conduct 5-fold cross validation, and use the average of 5 calculated concordance index as the average performance of a method in one simulation. Then, we calculate the average for the 300 simulations.

We describe the four simulation scenarios in detail in Section 3.3.1. We specify the compared models and methods in Section 3.3.2. Simulation results are showed in Section 3.3.3.

3.3.1 Four Simulation Scenarios

We simulate survival response \( Y \) based on Cox regression model with its causal features \( Z \), which was simulated based on original predictors \( X \). We design four different scenarios using different models to generate “causal” features \( Z \) based on original predictors \( X \). The specification of the four scenarios are as follows.
1. **Sparse Causal Predictors Scenario.** Randomly select 10 features from each of the two bioinformatics data sets. Then use the 20 features as causal features to simulate survival response \( Y \).

2. **Predetermined Neural Network Scenario.** Feed each simulated data set into a predetermined neural network with neural network parameter values fixed in each simulation and specify the output features from each neural network to be 10 features. Then use the 20 features as causal features to simulate survival response \( Y \). Neural network parameter values are generated from a standard normal distribution.

3. **Auto-encoder Scenario.** We apply auto-encoder neural network to simulate 10 features for each data type as causal features and then simulate survival response \( Y \).

4. **Principal Component Scenario.** Principal component analysis is used to reduce the full features sizes to 10 predictors for each data type and then simulate survival response \( Y \) with a total of 20 causal features.

We use simsurv R package to generate survival responses ([51]). The R package simsurv allows the specification of causal predictors and coefficients to control censored rate in simulating survival responses.

Table 3.1 shows the architecture of models in scenario 2 (Predetermined Neural Network Scenario) and scenario 3 (Auto-encoder Scenario). Auto-encoder neural networks consist of two parts: encoder part and decoder. We specify the number of nodes of hidden layers for each part.
Table 3.1: Details of Architecture of Neural Network Models

<table>
<thead>
<tr>
<th>Models</th>
<th># of Nodes hidden layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predetermined Neural Network Scenario</td>
<td>128,32,10</td>
</tr>
<tr>
<td>Auto-encoder- Encoder</td>
<td>128,64,32,10</td>
</tr>
<tr>
<td>Auto-encoder- Decoder</td>
<td>10,32,64,128</td>
</tr>
</tbody>
</table>

3.3.2 Cancer Survival Prediction Methods

We compare our proposed method, i.e. deep merged survival neural network, with four literature methods taking original features $X$ as the input, i.e. (i) Cox regression, (ii) penalized Cox regression (Cox-LASSO), (iii) Principal Component (PC) Cox regression (Cox PCA), and (iv) Random Survival Forest (RSF) under different scenarios. The first literature method Cox regression is infeasible because the model input $X$ from bioinformatic data sets is high-dimensional. In addition, we also evaluate the performance of Cox regression model taking causal features $Z$ as the input (True-Cox). Because survival response $Y$ is simulated based on causal features ($Z$), which was calculated based on original predictor $X$, we expect the Cox regression taking true causal features $Z$ as the input (True Cox Model) to have good performance. However, in practice, this method (Cox regression taking causal features $Z$ as the input) is not feasible because true causal features are unknown so that True Cox Model is not feasible. The evaluation of True-Cox method may provide reference to the ideal performance, because the estimation method based on the same model used in simulation could show good performance if estimation errors are not big.

The specification of methods are as follows.

1. **True Cox.** It is Cox regression model taking true causal features $Z$ as the inputs.

   This method is expected to have good performance but not feasible in practice.
because true causal features are unknown.

2. **Cox LASSO.** It is a penalized Cox regression model taking the original data $X$ as the input and using a $L_1$ penalty to encourage model sparsity.

3. **Cox PCA.** Top principal components (PCs) of data $X$ are extracted and then feed as the input into a Cox regression model.

4. **Random Survival Forest (RSF).** It is a survival random forest taking original data $X$ as the input and analyze survival response $Y$.

5. **Merged Deep Survival Neural Network (Deep Survival).** It is our proposed method. Individual neural network is adopted for each individual data set, and the variables in the last layer of each individual neural network are merged and then is fed into another neural network.

Figure 3.2 displays the architecture of our proposed deep merged survival neural network for the simulation.

![Figure 3.2: Simulation Study: The Architecture of Deep Merged Survival Neural Network for the Simulated Data Survival Analysis](image)

Figure 3.2: Simulation Study: The Architecture of Deep Merged Survival Neural Network for the Simulated Data Survival Analysis
3.3.3 Simulation Results

To evaluate the performance of different methods, we conduct 300 simulation under 5 scenarios. In each simulation, we conduct 5-fold cross validation. The results for each scenario are presented with the 354 overall mean of C-index. In Table 3.2 to 3.5, we show the results for the sample size of 500, 1000 and 2000. In addition, we plot the results of each scenario in Figure 3.3 to Figure 3.6.

The performance of our method (Survival NN), three feasible methods (Cox LASSO, Cox PCA, Random Survival Forest), and one infeasible method (True-Cox) are shown in Tables and Figures. Note that True Cox machine is not feasible in practice though we expect it show good model performance. In the tables, for each combination of $n$ and $p$, the highest performance scores among the four feasible methods (our proposed deep merged survival neural network, abbreviated as survival NN, Cox Lasso, Cox PCA, and RSF) is bolded.

The results in Scenario 1 (Sparse Causal Predictors Scenario) are shown in Table 3.2 and Figure 3.3). Under this scenario, we found that Cox Lasso has the best performance among the four feasible methods. Our proposed method (Survival NN) performs better than Cox PCA method and Random Survival Forest for most combinations.
<table>
<thead>
<tr>
<th>$p$</th>
<th>$n$</th>
<th>True Cox</th>
<th>Cox Lasso</th>
<th>Cox PCA</th>
<th>RSF</th>
<th>Survival NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td></td>
<td>0.940</td>
<td><strong>0.962</strong></td>
<td>0.692</td>
<td>0.768</td>
<td>0.748</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>0.939</td>
<td><strong>0.960</strong></td>
<td>0.686</td>
<td>0.750</td>
<td>0.756</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>0.938</td>
<td><strong>0.957</strong></td>
<td>0.682</td>
<td>0.737</td>
<td>0.748</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>0.960</td>
<td><strong>0.967</strong></td>
<td>0.664</td>
<td>0.712</td>
<td>0.733</td>
</tr>
<tr>
<td>300</td>
<td></td>
<td>0.944</td>
<td><strong>0.967</strong></td>
<td>0.698</td>
<td>0.798</td>
<td>0.788</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>0.942</td>
<td><strong>0.967</strong></td>
<td>0.697</td>
<td>0.776</td>
<td>0.784</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>0.942</td>
<td><strong>0.964</strong></td>
<td>0.688</td>
<td>0.765</td>
<td>0.767</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>0.966</td>
<td><strong>0.977</strong></td>
<td>0.674</td>
<td>0.740</td>
<td>0.752</td>
</tr>
<tr>
<td>300</td>
<td></td>
<td>0.947</td>
<td><strong>0.969</strong></td>
<td>0.704</td>
<td>0.817</td>
<td>0.831</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>0.944</td>
<td><strong>0.970</strong></td>
<td>0.701</td>
<td>0.796</td>
<td>0.817</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>0.948</td>
<td><strong>0.971</strong></td>
<td>0.691</td>
<td>0.779</td>
<td>0.785</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>0.970</td>
<td><strong>0.980</strong></td>
<td>0.676</td>
<td>0.759</td>
<td>0.763</td>
</tr>
</tbody>
</table>

Table 3.2: Average Concordance Score for Different Methods Based on Data Simulated Under Scenario 1 (Sparse Causal Predictors Scenario). Sample size is $n$ and feature size is $p$. 
For scenario 2 (Table 3.3, Figure 3.4), the model we proposed has the best performance among the other three feasible methods for most combinations. Cox Lasso has better performance than Cox PCA and Random Survival Forest (RSF) in this scenario.
<table>
<thead>
<tr>
<th>$p$</th>
<th>$n$</th>
<th>True Cox</th>
<th>Cox Lasso</th>
<th>Cox PCA</th>
<th>RSF</th>
<th>Survival NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td></td>
<td>0.942</td>
<td>0.704</td>
<td>0.656</td>
<td>0.686</td>
<td><strong>0.705</strong></td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>0.860</td>
<td>0.656</td>
<td>0.626</td>
<td>0.639</td>
<td><strong>0.659</strong></td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>0.836</td>
<td>0.575</td>
<td>0.568</td>
<td>0.558</td>
<td><strong>0.595</strong></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>0.943</td>
<td>0.669</td>
<td>0.634</td>
<td>0.653</td>
<td><strong>0.699</strong></td>
</tr>
<tr>
<td>300</td>
<td>500</td>
<td>0.943</td>
<td>0.729</td>
<td>0.668</td>
<td>0.709</td>
<td><strong>0.735</strong></td>
</tr>
<tr>
<td>500</td>
<td>1000</td>
<td>0.861</td>
<td><strong>0.688</strong></td>
<td>0.636</td>
<td>0.661</td>
<td>0.684</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>0.835</td>
<td>0.600</td>
<td>0.576</td>
<td>0.573</td>
<td><strong>0.663</strong></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>0.944</td>
<td>0.709</td>
<td>0.646</td>
<td>0.680</td>
<td><strong>0.720</strong></td>
</tr>
<tr>
<td>300</td>
<td>500</td>
<td>0.943</td>
<td>0.737</td>
<td>0.676</td>
<td>0.722</td>
<td><strong>0.761</strong></td>
</tr>
<tr>
<td>500</td>
<td>2000</td>
<td>0.864</td>
<td>0.702</td>
<td>0.644</td>
<td>0.677</td>
<td><strong>0.707</strong></td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>0.835</td>
<td>0.618</td>
<td>0.585</td>
<td>0.587</td>
<td><strong>0.644</strong></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>0.945</td>
<td>0.731</td>
<td>0.655</td>
<td>0.699</td>
<td><strong>0.734</strong></td>
</tr>
</tbody>
</table>

Table 3.3: Average Concordance Score for Different Methods Based on Data Simulated Under Scenario 2 (Predetermined Neural Network Scenario). Sample size is $n$ and feature size is $p$. 


In Scenario 3 (Table 3.4, Figure 3.5), our new model has the greatest C-index among other feasible methods, and also Cox Lasso outperforms the other two methods, Cox PCA and Random Survival Forest(RSF).
<table>
<thead>
<tr>
<th>$p$</th>
<th>$n$</th>
<th>True Cox</th>
<th>Cox Lasso</th>
<th>Cox PCA</th>
<th>RSF</th>
<th>Survival NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>300</td>
<td>0.951</td>
<td>0.773</td>
<td>0.748</td>
<td>0.730</td>
<td><strong>0.773</strong></td>
</tr>
<tr>
<td>500</td>
<td>500</td>
<td>0.936</td>
<td>0.758</td>
<td>0.742</td>
<td>0.728</td>
<td><strong>0.763</strong></td>
</tr>
<tr>
<td>1000</td>
<td>1000</td>
<td>0.939</td>
<td>0.731</td>
<td>0.725</td>
<td>0.712</td>
<td><strong>0.745</strong></td>
</tr>
<tr>
<td>2000</td>
<td>2000</td>
<td>0.941</td>
<td>0.712</td>
<td>0.716</td>
<td>0.702</td>
<td><strong>0.734</strong></td>
</tr>
<tr>
<td>300</td>
<td>300</td>
<td>0.958</td>
<td>0.799</td>
<td>0.768</td>
<td>0.751</td>
<td><strong>0.818</strong></td>
</tr>
<tr>
<td>500</td>
<td>500</td>
<td>0.951</td>
<td>0.782</td>
<td>0.757</td>
<td>0.742</td>
<td><strong>0.799</strong></td>
</tr>
<tr>
<td>1000</td>
<td>1000</td>
<td>0.948</td>
<td>0.763</td>
<td>0.743</td>
<td>0.731</td>
<td><strong>0.774</strong></td>
</tr>
<tr>
<td>2000</td>
<td>2000</td>
<td>0.948</td>
<td>0.747</td>
<td>0.729</td>
<td>0.724</td>
<td><strong>0.756</strong></td>
</tr>
<tr>
<td>300</td>
<td>300</td>
<td>0.966</td>
<td>0.809</td>
<td>0.780</td>
<td>0.765</td>
<td><strong>0.857</strong></td>
</tr>
<tr>
<td>500</td>
<td>500</td>
<td>0.937</td>
<td>0.790</td>
<td>0.771</td>
<td>0.755</td>
<td><strong>0.839</strong></td>
</tr>
<tr>
<td>1000</td>
<td>1000</td>
<td>0.955</td>
<td>0.770</td>
<td>0.753</td>
<td>0.742</td>
<td><strong>0.808</strong></td>
</tr>
<tr>
<td>2000</td>
<td>2000</td>
<td>0.956</td>
<td>0.763</td>
<td>0.741</td>
<td>0.740</td>
<td><strong>0.784</strong></td>
</tr>
</tbody>
</table>

Table 3.4: Average Concordance Score for Different Methods Based on Data Simulated Under Scenario 3 (Auto-encoder Scenario). Sample size is $n$ and feature size is $p$. 
In Scenario 4 (Table 3.5, Figure 3.6), Cox PCA regression models perform the best among the other four feasible methods. The Survival NN method perform better than Random Survival Forest while it is difficult to compare the performance new model with the performance of Cox Lasso.
<table>
<thead>
<tr>
<th>$p$</th>
<th>$n$</th>
<th>True Cox</th>
<th>Cox Lasso</th>
<th>Cox PCA</th>
<th>RSF</th>
<th>Survival NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td></td>
<td>0.968</td>
<td>0.883</td>
<td><strong>0.959</strong></td>
<td>0.771</td>
<td>0.872</td>
</tr>
<tr>
<td>500</td>
<td>500</td>
<td>0.932</td>
<td>0.870</td>
<td><strong>0.927</strong></td>
<td>0.784</td>
<td>0.866</td>
</tr>
<tr>
<td>1000</td>
<td>1000</td>
<td>0.938</td>
<td>0.850</td>
<td><strong>0.934</strong></td>
<td>0.774</td>
<td>0.846</td>
</tr>
<tr>
<td>2000</td>
<td>2000</td>
<td>0.936</td>
<td>0.825</td>
<td><strong>0.931</strong></td>
<td>0.750</td>
<td>0.818</td>
</tr>
<tr>
<td>300</td>
<td></td>
<td>0.972</td>
<td>0.896</td>
<td><strong>0.967</strong></td>
<td>0.780</td>
<td>0.906</td>
</tr>
<tr>
<td>500</td>
<td>1000</td>
<td>0.934</td>
<td>0.883</td>
<td><strong>0.931</strong></td>
<td>0.793</td>
<td>0.897</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>0.939</td>
<td>0.863</td>
<td><strong>0.937</strong></td>
<td>0.784</td>
<td>0.875</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>0.943</td>
<td>0.842</td>
<td><strong>0.941</strong></td>
<td>0.762</td>
<td>0.853</td>
</tr>
<tr>
<td>300</td>
<td></td>
<td>0.976</td>
<td>0.898</td>
<td><strong>0.972</strong></td>
<td>0.784</td>
<td>0.931</td>
</tr>
<tr>
<td>500</td>
<td>2000</td>
<td>0.935</td>
<td>0.882</td>
<td><strong>0.934</strong></td>
<td>0.799</td>
<td>0.921</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>0.940</td>
<td>0.862</td>
<td><strong>0.939</strong></td>
<td>0.792</td>
<td>0.902</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>0.951</td>
<td>0.838</td>
<td><strong>0.949</strong></td>
<td>0.768</td>
<td>0.886</td>
</tr>
</tbody>
</table>

Table 3.5: Average Concordance Score for Different Methods Based on Data Simulated Under Scenario 4 (Principal Component Scenario). Sample size is $n$ and feature size is $p$.  

In summary, simulation results indicate that different scenarios have different best models. The methods with the best performance (the best players) are Cox Lasso, Survival NN, Survival NN, Cox PCA, under Scenario 1 (Sparse Causal Predictors Scenario), Scenario 2 (Predetermined Neural Network Scenario), Scenario 3 (Auto-encoder Scenario), Scenario 4 (Principal Component Scenario), respectively.
3.4 Real Data Analysis

For a decade, the Cancer Genome Atlas (TCGA) collected bioinformatic data sets, histological images and patients’ clinical information for 33 different cancer types. We use TCGA breast invasive carcinoma (BRCA) data for survival analysis in this study.

3.4.1 TCGA BRCA Data Sets

There are 1093 patients with BRCA in TCGA BRCA data. Among them, 1085 patients are female. TCGA BRCA data includes multiple different types of bioinformatics data which includes mRNA, miRNA, DNA methylation, Copy Number Variation (CNV), histological images and clinical information.

In our real data application, we consider six specific TCGA BRCA data sets. The specification of the six data sets are

1. mRNA expression values, retrieved from
   TCGA/BRCA/illuminahiseq/rnaseqv2/RSEM/genes/normalized.

2. miRNA expression values, retrieved from
   TCGA/BRCA/illuminahiseq/mirnaseq/miR/gene/expression.

3. DNA methylation beta values, retrieved from Broad GDAC Firehose by the name humanmethylation450.

4. copy number variation, retrieved from National Cancer Institute by the name TCGA program.

5. histological images, provided in Hou et al.’s study [52].

6. clinical and survival information, retrieved from TCGA BRCA database.

Clinical variables considered in our real application includes diagnosis age, race, estrogen receptors (ER) status and progesterone receptors (PR) status. The progesterone
receptor (PR) and estrogen receptor (ER) are essential tissue markers that guide the
treatment of breast cancer (Mouttet et al. [53]). The clinical characteristics of the 1093
patients in TCGA data are (1) diagnostic age (mean = 58; range = 28-90), (2) PR positive
ratio (proportion = 66.76%), and (3) ER positive ratio (proportion = 77.1%).

Regarding survival information, we retrieve overall survival (OS) events and survival
time in days are from TCGA BRCA program. Overall survival events refer to patients’ vital
status, which is either alive or dead. Overall survival time are the length of the time from
the date of diagnosis with the disease such as cancer for patients who were diagnosed
with the disease and still remain alive.

### 3.4.2 Data Preprocessing

Our real data preprocessing includes three steps: (1) data transformation and normal-
ization, (2) feature filtering, and (3) missing value imputation.

**Data Transformation and Normalization**

Different transformation and normalization methods are applied to different types of
bioinformatics data.

The mRNA expression values $X$ were retrieved from TCGA BRAC program. They are
calculated using the software RSEM (RNA-seq by Expectation Maximum). The miRNA
values are RPKM (Reads per Kilobase of transcript per Million). Log2 transformation
with the pseudo-count 1 is applied on both mRNA values and miRNA values, i.e. $g(X) = \log_2(1 + X)$.

DNA methylation is measured by beta values, taking values between 0 and 1. The
beta value of 0 indicates no methylation at a specific locus and 1 indicates CpG methyl-
ated. Copy number variation (CNV) can take 3 possible values: $-1$ refers to the loss with
fewer copies of a gene than normal, 1 refers to the gain with more copies of a gene than
Histopathological image data is retrieved from TCGA BRAC program. These histological images, especially the spatial distribution of cell nuclei shown in the images, are believed to be useful for the diagnosis of cancers. Hou et al. [52] proposed an accurate method to conduct nucleus segmentation based on Whole Slide Tissue images (WSI) from TCGA database. Summary statistics from these segmented images (such as the number of pixels with nuclei) can help in cancer diagnosis. In this study, we adopted Hou et al. [52]’s segmentation method to segment Whole Slide Images (WSI), and then calculate summary statistics (the total number of nuclei pixels, the mean and standard deviation of the number of nucleic pixels and total number of nuclei for each WSI) from segmented data. These image summary statistics (Hou et al. [52] are believed to contain useful information for cancer diagnosis. Then we apply min-max normalization on the image summary statistics to calculate the data of normalized summary statistics for histological images. The formula for min-max normalization is $g(X) = \frac{X - \text{min}(X)}{\text{max}(X) - \text{min}(X)}$.

**Feature Filtering**

Bioinformatics data needs variable filtering in data preprocessing. Variables were ordered based on a metric, and then filtered. Following the literature, we use coefficient of variation (CV) to rank mRNA, miRNA and methylation variables, and maximum entropy to rank copy number variation (CNV) variables. Top 20% variables are stored for future studies and other variables are filtered out.

Table 3.6 shows the number of features before and after marker filtering for the four types of bioinformatics data (mRNA, miRNA, methylation and CNV). For histological image data and clinical information data, because there are only tens of variables, we use all variables in our study, i.e. WSI summary statistics and clinical variables, without feature filtering.
<table>
<thead>
<tr>
<th></th>
<th>mRNA</th>
<th>miRNA</th>
<th>Methylation</th>
<th>CNV</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Before filtering</strong></td>
<td>20253</td>
<td>897</td>
<td>396065</td>
<td>19729</td>
</tr>
<tr>
<td><strong>After filtering</strong></td>
<td>4050</td>
<td>179</td>
<td>79213</td>
<td>3945</td>
</tr>
</tbody>
</table>

Table 3.6: Number of Variables Before and After Feature Filtering

**Missing Value Imputation**

Not all the patients have data in all the five bioinformatics data sets (mRNA, miRNA, methylation, CNV and WSI). There may be some patients with some information missing. We make a Venn diagram in Figure 3.7 to show the number of patients and overlapping of patients in the 5 bioinformatics datasets. There are 571 patients with all the five types of information (mRNA, miRNA, methylation, CNV and WSI) available. We use the 571 patients in our subsequent analysis.

There are missing values for the 571 patients in our study. We filled missing values by the median, i.e. median imputation.
3.4.3 Model Construction

In this sub-section, we show the architecture of our proposed model, which has a merge layer to integrate the information from each data group. Each specific data group is handled by one customized neural network, and the last layers of these neural networks are combined by a merge layer, which is fed into another neural network with a Cox loss function to conduct survival analysis. The Cox loss function is a negative partial likelihood as defined in the methodology section.

Specific model architecture for breast cancer survival analysis combining multiple types of bioinformatics data (mRNA, miRNA, methylation CNV, WSI and clinic information) is shown in Figure 3.8.

In this model, four specific data sets (mRNA, miRNA, methylation and CNV) are handled by four neural networks, with each data set fed into one specifically designed
neural networks. The last layers of the four individual neural networks, clinical variables and whole slide image (WSI) variables are combined together by a merge layer, which is fed into a survival neural network, i.e. the Cox loss function is used.

The benefits of this model architecture are (1) allowing each data set handled by customized designed individual neural network, (2) allowing the integration of information in different data sets to be combined at higher level instead of at the input level so that each data set or each variable group is well represented in our model architecture by a merge layer, and (3) allowing simultaneous estimation of the whole model architecture together.

Based on our constructed model, six different types of bioinformatics data sets, or six different variable groups, i.e. mRNA, miRNA, methylation, copy number variation (CNV), whole slide image (WSI) and clinic variables, are used together in our real data application of TCGA breast cancer survival analysis.
3.4.4 Real Data Application Results

To evaluate the performance of different methods in real data application, we apply 5-fold cross validation for 10 times. The cross-validation used 20% of data as testing data, and 80% of data as training data. In this way, 50 testing data sets, are generated. The concordance score, i.e. C-index, is used as evaluation metric. Mean, median, and standard deviation of 50 C-index, is reported. In addition, we compare the performances of using different combinations of bioinformatics data sets and the performance of using
different methods.

In Table 3.4.4, we compare the performances of methods by using different combinations of bioinformatic data for 571 female patients. Following literature conventions, we always include clinical variables in analysis as control variables. Thus, in all combinations of bioinformatics variable groups, clinical variables are included as the default. Mean, median and standard deviation (SD) of the concordance scores are reported based on the 50 testing data sets.

Among all possible combinations of bioinformatics data sets, or variable groups, we find that the use of all six bioinformatics data sets based on our model achieves the best performance.

There are widely used combination of different bioinformatics variable groups for cancer survival analysis in the literature. Most researchers believe that clinical variables such as age and gender are most important and should be used in analysis as covariates. The mRNA variables are preferred to be used compared with other bioinformatics variable groups (miRNA, copy number variation, methylation, whole slide images). Figure 3.9 displays the performance of widely used combinations of bioinformatic data for cancer survival analysis. It can be found that with more and more data sets or variable groups added, the performance improves on the whole.

Figure 3.10 shows an increasing trend on mean of C-index, which represents accuracy of the model. The upward trend indicates the more bioinformatic data included, the more accuracy of prediction.

Table 3.7: Model Performance Using Different Combinations of Bioinformatics Data Sets. The evaluation metric is the C-index.

<table>
<thead>
<tr>
<th>Combinations</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>mRNA miRNA Methylation CN WSI</td>
<td>0.705</td>
<td>0.706</td>
<td>0.013</td>
</tr>
<tr>
<td>mRNA miRNA Methylation CN</td>
<td>0.698</td>
<td>0.697</td>
<td>0.022</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Combinations</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>mRNA miRNA Methylation WSI</td>
<td>0.685</td>
<td>0.686</td>
<td>0.025</td>
</tr>
<tr>
<td>mRNA miRNA CN WSI</td>
<td>0.691</td>
<td>0.700</td>
<td>0.028</td>
</tr>
<tr>
<td>mRNA Methylation CN WSI</td>
<td>0.696</td>
<td>0.694</td>
<td>0.022</td>
</tr>
<tr>
<td>miRNA Methylation CN WSI</td>
<td>0.694</td>
<td>0.699</td>
<td>0.020</td>
</tr>
<tr>
<td>mRNA miRNA Methylation</td>
<td>0.689</td>
<td>0.687</td>
<td>0.026</td>
</tr>
<tr>
<td>mRNA miRNA CN</td>
<td>0.693</td>
<td>0.695</td>
<td>0.029</td>
</tr>
<tr>
<td>mRNA miRNA WSI</td>
<td>0.678</td>
<td>0.674</td>
<td>0.026</td>
</tr>
<tr>
<td>mRNA Methylation CN</td>
<td>0.696</td>
<td>0.699</td>
<td>0.022</td>
</tr>
<tr>
<td>mRNA Methylation WSI</td>
<td>0.692</td>
<td>0.689</td>
<td>0.016</td>
</tr>
<tr>
<td>mRNA CN WSI</td>
<td>0.684</td>
<td>0.690</td>
<td>0.025</td>
</tr>
<tr>
<td>miRNA Methylation CN</td>
<td>0.694</td>
<td>0.697</td>
<td>0.031</td>
</tr>
<tr>
<td>miRNA Methylation WSI</td>
<td>0.681</td>
<td>0.675</td>
<td>0.024</td>
</tr>
<tr>
<td>miRNA CN WSI</td>
<td>0.614</td>
<td>0.617</td>
<td>0.024</td>
</tr>
<tr>
<td>Methylation CN WSI</td>
<td>0.689</td>
<td>0.685</td>
<td>0.027</td>
</tr>
<tr>
<td>mRNA miRNA</td>
<td>0.676</td>
<td>0.685</td>
<td>0.035</td>
</tr>
<tr>
<td>mRNA Methylation</td>
<td>0.677</td>
<td>0.682</td>
<td>0.040</td>
</tr>
<tr>
<td>mRNA CN</td>
<td>0.688</td>
<td>0.686</td>
<td>0.016</td>
</tr>
<tr>
<td>mRNA WSI</td>
<td>0.670</td>
<td>0.669</td>
<td>0.021</td>
</tr>
<tr>
<td>miRNA Methylation</td>
<td>0.681</td>
<td>0.687</td>
<td>0.023</td>
</tr>
<tr>
<td>miRNA CN</td>
<td>0.615</td>
<td>0.616</td>
<td>0.027</td>
</tr>
<tr>
<td>miRNA WSI</td>
<td>0.602</td>
<td>0.602</td>
<td>0.036</td>
</tr>
<tr>
<td>Methylation CN</td>
<td>0.680</td>
<td>0.671</td>
<td>0.043</td>
</tr>
<tr>
<td>Methylation WSI</td>
<td>0.678</td>
<td>0.679</td>
<td>0.021</td>
</tr>
<tr>
<td>CN WSI</td>
<td>0.580</td>
<td>0.582</td>
<td>0.014</td>
</tr>
</tbody>
</table>
Table 3.7 – Continued from previous page

<table>
<thead>
<tr>
<th>Combinations</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>mRNA</td>
<td>0.671</td>
<td>0.679</td>
<td>0.022</td>
</tr>
<tr>
<td>miRNA</td>
<td>0.598</td>
<td>0.601</td>
<td>0.041</td>
</tr>
<tr>
<td>Methylation</td>
<td>0.676</td>
<td>0.675</td>
<td>0.022</td>
</tr>
<tr>
<td>CN</td>
<td>0.577</td>
<td>0.582</td>
<td>0.015</td>
</tr>
<tr>
<td>WSI</td>
<td>0.638</td>
<td>0.650</td>
<td>0.031</td>
</tr>
</tbody>
</table>

Figure 3.9: Model Performance Using Different Combinations of Bioinformatics Data
Next, we compare the performance of using different methods (our method versus literature methods) given the same dataset. The results of different methods are shown in Figure 3.11. The literature methods we considered are (1) CoxPH, (2) CoxPH + LASSO, (3) Cox + PCA, (4) Deep Survival (2018) (Katzman et al. [44]), (5) Random Survival Forest (RSF, Ishwaran et al. [38]), and (6) our proposed model.

CoxPH is a classical linear Cox proportional hazard model. However, because our bioinformatics data sets includes more than 10K variables in $X$ with the number of patients to be less than 573. CoxPH can not be used. However, other methods based on Cox regression, such as using principal components (PCs) as the input in Cox regression, i.e. Cox+PCA, or $L_1$ penalized cox regression (CoxPH+LASSO) can be used.

In our Cox+PCA method, we first compute top 20 PCs for each bioinformatics data set, and then put these PCs into a Cox regression model. DeepSurv method is proposed by J.L. Katzman et al. [44]. DeepSurv is widely applied to conduct analysis based on
an individual data set or a few bioinformatics data sets. Random Survival Forest was proposed by Ishwaran et al. [38]. It uses random forest for survival analysis.

We report the performance of our proposed model (deep merged survival neural networks, shortened as Survival NN) and the literature methods (CoxPH+LASSO, CoxPH+PCA, Deep Survival (2018) (Katzman et al. [44]), Random Survival Forest) in Figure 3.11. It can be found that our proposed model has the best performance.

![Figure 3.11: Model Performances Using Different Methods](image)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Concordance Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoxPH+Lasso</td>
<td>Mean=0.609</td>
</tr>
<tr>
<td>CoxPH+PCA</td>
<td>Mean=0.508</td>
</tr>
<tr>
<td>DeepSurv(2018)</td>
<td>Mean=0.657</td>
</tr>
<tr>
<td>RSF</td>
<td>Mean=0.618</td>
</tr>
<tr>
<td>Survival NN</td>
<td>Mean=0.705</td>
</tr>
<tr>
<td></td>
<td>Median=0.619</td>
</tr>
<tr>
<td></td>
<td>Median=0.507</td>
</tr>
<tr>
<td></td>
<td>Median=0.662</td>
</tr>
<tr>
<td></td>
<td>Median=0.620</td>
</tr>
<tr>
<td></td>
<td>Median=0.706</td>
</tr>
</tbody>
</table>

Next, we show the performance of using different methods and different combinations of data sets together in one figure. We noticed in the literature, researchers typically only use a few bioinformatics data sets for specific methods. Figure 3.12 shows the model performance using five different methods (CoxPH+LASSO, Cox+PCA, DeepSurv(2018), RSF and our survival NN) and six different combinations of multiple data types (mRNA, mRNA+miRNA, mRNA+miRNA+CN, mRNA+miRNA+Methylation, mRNA+miRNA+Methylation+CN, and mRNA+miRNA+Methylation+CN+wsii)
It can be found that given the same combination of data sets, our method is the best model among our methods with the only exception that when only mRNA is used, RSF shows better performance than our model as indicated in blue line.

All six combinations of data sets and all five methods, the best-performing player is our method using all the bioinformatics data sets (mRNA, miRNA, Methylation, CN, and WSI) under consideration. This shows that the advantage of our method over other methods in conducting analysis based on our proposed innovative deep merged survival neural network with multiple different types of bioinformatics data sets.

Figure 3.12: Performances of Different methods Using Different Combinations of Bioinformatics Data

3.5 Discussion

We propose the deep merged survival neural network method with merge layer which can integrate multiple heterogeneous data to predict survival. In simulation studies, the new deep survival model has more accuracy for prediction in Scenario 2 (Predetermined
Neural Network Scenario) and Scenario 3 (Auto-encoder Scenario). In breast cancer study, our model considers 6 different feature groups (e.g. clinical variables, mRNA, miRNA, epigenetic methylation, Copy Number Variation (CNV), and histological whole slide images), which are widely used feature groups in the literature. In general, the method is good to integrate and balance the information from individual bioinformatic data. In addition, the method is useful to integrate multiple types of bioinformatic data with a large feature size. The results show that our model has good performance compared with other statistical methods in making predictions based on multiple data sets.

By deep merging information based on survival neural networks, we expect our model will be better to integrate multiple heterogeneous complex data such as when multiple data sets have different feature sizes $p$. Our current simulation study is based on multiple data with equal feature size. Future studies should consider simulations based on multiple data sets with unequal feature sizes. Our proposed method, deep merged survival neural network, is a general method to conduct analysis based on multiple complex and heterogeneous data sets. In real data application, we only illustrate its use in breast cancers, whereas the use of our method can be illustrated using other types of cancer data.

The TCGA cancer data has missing values. We fill the missing values by median imputation as widely used bioinformatics data pre-processing for neural networks. Future studies can be made on effect of different missing value imputation methods such as mean imputation on our proposed method.

For future studies, customized individual neural network can be applied to handle different types of data sets. For example, instead of using whole slide image summary statistics, directly take wsi images input and then use convolutional neural network (CNN) to handle it. In addition, recurrent neural network (RNN) can be used to handle DNA sequence data.
Bibliography


[48] Zhi Huang, Xiaohui Zhan, Shunian Xiang, Travis S. Johnson, Bryan Helm, Christina Y. Yu, Jie Zhang, Paul Salama, Maher Rizkalla, Zhi Han, and Kun Huang.


