Spring 5-15-2022

# Contribution to Data Science: Time Series, Uncertainty Quantification and Applications

Dhrubajyoti Ghosh
*Washington University in St. Louis*

WASHINGTON UNIVERSITY IN ST.LOUIS

School of Arts & Sciences
Department of Mathematics and Statistics

Dissertation Examination Committee:
Soumendra Lahiri, Chair
Tucker McElroy, Co-Chair
William Boettcher
Todd Kuffner
Debasish Mondal
Victor Wickerhauser

# Contribution to Data Science: Time Series, Uncertainty Quantification and Applications

by
Dhrubajyoti Ghosh

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2022
St. Louis, Missouri

# Table of Contents

# List of Figures

# List of Tables

# Acknowledgments

The winter of 90’s,
A story began;
With the birth of a boy,
And the tracks he ran.
Sumitra and Goutam,
The quintessential parents,
Nurture the boy,
For the future he rents.
From AK Ghosh to St. Lawrence,
The boy chased a beacon,
Met many great people,
Teachers and friends, thank you everyone,
Tushar, Suman and Arijit,
You deserve a special mention.
The school days went fast,
Like the summer in Winterfell;
Adulthood knocks on door,
It’s time to come out of the shell.
A choice was to be made,
and ISI won the laurel,
Adieu my peace of mind,
Started a lifelong battle.
It was never easy, note it down,
For every success you earn,
A sacrifice mars your gown.
Old leaves fall, Dreams die down,
A life you leave,
and a life you crown.
Friends and teachers,
Thank you all,
Without you,
The boy was sure to fall.
For every tunnel,
there lies a bend,
After 5 years of toil,
ISI life comes to an end.

8000 miles away from home,
A new life begins,
New faces, covered in mask,
Locked doors, and roadside inns.
Lots to learn,
And so little time,
Streams of knowledge,
But nothing to rhyme.
Thank you everyone,
For giving a home away from home,
Words would never suffice,
Thank you, for accepting this mome.
Life teaches us all,
Pulls us down only to be torn,
The boy has to be killed,
For a man needed to be born.

Years of living alone,
The boy finally found his yod;
"Whoever is delighted in solitude,
Is either a wild beast or a god".
5 years was a lot of time,
Lets just keep this short,
Some days it was wonderful,
And for days it was amort.
Love and hate molds us all,
A tale of ice and fire,
The ghosts of past stays by our side,
From Mordor to the peaceful shire.
The end has finally come,
Its time to destroy the ring,
There is no secret ingredient,
It was always you, Alone in the ring.

A big shout out to all my seniors, friends, and juniors who have helped me in my last 5 years. Hazrada, Sahooda, Chakda, Sumanda, Rahulda, Arkopalda, Golda, Sohinidi, Moumitadi, Indrabatidi you guys were the best seniors anyone could have asked for. I will forever remember the lunches with Hazrada, the delicious treats given by Sahooda, the cricket and tennis matches with Chakda, the philosophical and whatnot discussions with Sumanda, the FIFA and chess matches with Rahulda, and the long chats with Arkopalda. Salilda, you were one of my closest friends during my years at NCSU. Thank you both you and Chakda for teaching me a lot of things, which has helped me immensely in my years at St Louis. Samriddha, Debangan, Rohan, Eva, Sohom, Soham, Nabarun, Nabaneet, Sayantan a big thanks to all of you for keeping in touch with me and motivating me even after settling in different parts of the world. Sukanya, thank you for being a great friend, and for always being there for me whenever I needed someone to talk to. Tuhin, Sumit, Subhankar, Indrila, Shubhajit you guys were great juniors, and it is unfortunate that I couldn't spend more time with you guys since moving out of Raleigh. Samhita, I really don't know where to start. You became one of my closest friends in such a short span of time, a friend I really never had. You will always be important to me, and I hope we can maintain our friendship in the years to come.

Prof. Lahiri, thank you for being a wonderful guide. It would never have been possible for me to finish my Ph.D. without your guidance and helpful insights. Thank you for being an emotional support when I was having trouble concentrating on work during my years in St Louis. I am glad I chose you to be my advisor at NCSU. Prof. McElroy, thank you for your inputs and for guiding me during my Ph.D. years. It was a great experience working with both of you, and I sincerely hope to continue doing so in the future. Also, thank you Bill and Rob for our productive discussions in our LAS meetings. Thank you Prof. Wickerhauser, Prof. Mondal, and Prof. Kuffner for agreeing to be on my dissertation committee, and for your recommendations to my applications.

Finally, I would like to thank my parents, Sumitra and Goutam. I am where I am because of you, and any success I achieve in the future will be because of the hard work you put into me during my childhood. I know now how difficult it was for both of you, and no amount of words will ever be enough to thank you for what you have done for me. I am sorry I have to leave you for my career, but I know that I will always keep trying my best to make you the proudest parents ever.

Dhrubajyoti Ghosh

*Washington University in Saint Louis*

*August 2022*

Dedicated to my parents Sumitra and Goutam

## ABSTRACT OF THE DISSERTATION

Contribution to Data Science: Time Series, Uncertainty Quantification and Applications

by

Dhrubajyoti Ghosh

Doctor of Philosophy in Statistics

Washington University in St. Louis, 2022

Professor Soumendra Lahiri, Advisor

Professor Tucker McElroy, Co-Advisor

Time series analysis is an essential tool in modern statistical analysis, with many real data problems having temporal components that need to be studied better to understand the temporal dependence structure in the data. For example, in the stock market, it is of significant importance to identify the ups and downs of the stock prices, for which time series analysis is crucial. Most existing literature on time series deals with linear time series or with Gaussianity assumption. However, there are multiple instances where the time series shows nonlinear trends or when the underlying error structure is non-Gaussian. In such cases, nonlinear time series analysis is essential, which can be achieved using a nonlinear parametric design or nonparametric approaches.

In Chapter 2, we have presented a quadratic prediction technique to improve prediction accuracy when the time series is nonlinear or non-gaussian. We have also quantified the amount of prediction gain we achieve using the quadratic prediction procedure. Furthermore, we have provided a characterization of the processes for which the quadratic prediction will always give a better result than linear prediction in terms of the bispectra of the underlying process. We have provided comprehensive simulation studies and two real data

analyses to substantiate the theoretical results obtained. Chapter 3 deals with polyspectral means, a higher-order version of spectral means, which gives us important insights into a time series under the existence of non-linearity. We have proposed an estimate of the polyspectral mean and derived its asymptotic distribution. We have also proposed a linearity test based on the obtained asymptotic normality result. Finally, we have provided a simulation study and a real-world data analysis to offer possible applications of the polyspectral means in the real-world scenario.

The next part of the thesis deals with real data analysis. Chapter 4 is devoted to an election-prediction algorithm, which utilizes hashtag information and the dynamic network structure in social media data and opinion polls. We proposed two algorithms, one using the network structure (THANOS) and one without (THOS). Both our methods performed better than existing election prediction algorithms. Moreover, the network structure gave a closer prediction for closely fought elections than the one without it. Chapter 5 involves proposing a bot-detection algorithm for social media data. Inorganic accounts, famously known as bots, are used extensively for spreading malicious information and false propaganda, and it is of significant importance to identify them as quickly as possible. We have extracted several temporal and semantic features and used known machine learning algorithms to identify the inorganic accounts.

The final chapter deals with bootstrap in extreme value analysis. Efron's bootstrap is proven to be inconsistent with extreme value theory. [50] showed that m out of n bootstrap works in this particular scenario when $m = o(n)$. that m out of n bootstrap works in this particular scenario when $m = o(n)$. However, there has not been much work on how to find the optimal $m$ in the $m$ out of $n$ bootstrap. In Chapter 6, we propose an optimal choice of m that would minimize the bootstrap's convergence rate. We have given a real-world data analysis using the AQI level of several cities worldwide.

# Chapter 1

# Introduction

Time series analysis is essential in analyzing real data with a temporal component. With continuous monitoring and data collection becoming increasingly common, the need for time series analysis is continually increasing. Time series analysis often comes down to the question of causality: how did the past influence the future? Time series analysis finds application in a variety of sectors: medicine ([32], [84], [49]), weather forecast ([43], [114], [66]), economic forecasting ([33], [125], [78]), and so on. There are numerous literature in time series analysis, most of which works with linear time series or under Gaussianity assumption ([25], [141], [56]). However, there are multiple instances where the time series shows nonlinear trends or when the underlying error structure is non-Gaussian, which demands nonlinear time series analysis ([69], [101], [68]). In such instances, nonlinear time series analysis is essential, which can be achieved using a nonlinear parametric structure or nonparametric approaches ([46]). Nonlinear time series prediction can offer potential accuracy gains over linear methods when the process is nonlinear. In this thesis, we provide a brief overview of our contribution to the field of time series analysis. The first contribution is motivated by the need to improve prediction accuracy when there is non-Gaussianity or non-linearity in the data. Several real-world applications, including stock market data, economic data, and astronomy data, exhibit nonlinear trends in their temporal components. Polyspectral means, higher-order versions of spectral means, give essential insights into the data under nonlinear circumstances. In our second work, we propose an estimate for polyspectral mean and its asymptotic distribution and also used the same to present a test for linearity of time series. In chapter 4, we propose a model THANOS, which incorporates various temporal aspects of social media data to predict the results of a two-party election. Chapter 5 deals with identifying inorganic accounts, popularly known as bots, in social media. Finally, Chapter 6 gives an optimal choice of m in m out n bootstrap for sample extremes.

## 1.1 Time Series Analysis

Chapters 2 and 3 deal with theoretical works on time series analysis. In Chapter 2, We explore the class of quadratic predictors, which directly generalize linear predictors. We show they can be computed using the second, third, and fourth auto-cumulant functions when the time series is stationary. The new formulas for quadratic predictors generalize the normal equations for linear prediction of stationary time series, and hence we obtain quadratic generalizations of the Yule-Walker equations; we explicitly quantify the prediction gains in quadratic over linear methods. We say a stochastic process is *second-order forecastable* if quadratic prediction provides an advantage over linear prediction. One of the critical results of the paper characterizes second-order forecastable processes in terms of the spectral and bispectral densities. We verify these conditions for some popular nonlinear time series models. Numerical results and real data examples presented here show marked improvement over linear predictions in many cases.

Higher-order spectra (or polyspectra), defined as the Fourier Transform of a stationary process' autocumulants, are helpful in the analysis of nonlinear and non-Gaussian processes. Polyspectral means are weighted averages over Fourier frequencies of the polyspectra, and estimators can be constructed from analogous weighted averages of the higher-order periodogram (a statistic computed from the data sample's discrete Fourier Transform). In Chapter 3, We derive the asymptotic distribution of a class of polyspectral mean estimators, obtaining an exact expression for the limit distribution that depends on both the given weighting function and higher-order spectra. Secondly, we use bispectral means to define a new test of the linear process hypothesis. Simulations document the finite sample properties of the asymptotic results. Two applications illustrate our results' utility: we test the linear process hypothesis for a Sunspot time series and the Gross Domestic Product, and we conduct a clustering exercise based on bispectral means with different weight functions. Chapters 2 and 3 is joint work with Prof. Soumendra Lahiri from Washington University in St. Louis and Prof. Tucker McElroy from the U.S. Census Bureau.

## 1.2 Data Analysis

Chapters 4 and 5 mainly focus on real-world data analysis, particularly social media data. In this section, we worked on social media data; the social media platform considered being Twitter due to the ease of data collection. The influence and impact of social media campaigns on democratic elections is a crucial research topic for modern big-data analytics. Since the 2016 U.S. presidential election, much attention has been devoted to retrospectively identifying influence efforts by malign state actors deployed through common social media platforms like Facebook, Twitter, Instagram, etc. Much of this work focused on identifying perpetrators, techniques used to propagate damaging content, and the possible connections between malign external action and domestic individuals or parties. Studies examining how this influence affects elections have received less attention. While several studies in the political and social science literature have questioned the use of social media data for forecasting the results of political races, Chapter 4 proposes a novel modeling approach to predicting electoral outcomes by *combining* public opinion polls and Twitter data and by incorporating key summary features of the network structure of the Twitter data to produce accurate predictions. Application to real data from Ireland's $36^{th}$ amendment referendum and the United States' 2018 Congressional elections exhibits promising results. This work was done in collaboration with Dr. William Boettcher, Dr. Rob Johnston and Dr. Michele Kolb from North Carolina State University, and Prof. Soumendra Lahiri from Washington University in St. Louis.

Inorganic users, or bots, play a significant role in social media activities. Often, these bots are used to propagate malicious propaganda and influence public opinion. These can have substantial impacts on electoral outcomes or military conflicts. Hence, it is vital to identify inorganic accounts as fast as possible. Chapter 5 deals with identifying these inorganic accounts and discusses how we can efficiently identify these accounts through machine learning techniques. We have used two real data sources, one from Botometer and one from Social Sifter Dataset, collected using inorganic accounts from different influence campaigns. This work was done in collaboration with Dr. William Boettcher and Dr. Rob Johnston from North Carolina State University, and Prof. Soumendra Lahiri from Washington University in St. Louis.

## 1.3 Extreme Value Bootstrap

Chapter 6 deals with extreme value bootstrap. Efron's Bootstrap is known to be inefficient in specific scenarios, one of them being the extremes of independent and identically distributed random variables when the resample- and the sample sizes are equal. In such cases, the $m$ out of $n$ Bootstrap is used, with $m = o(n)$. The consistency of the $m$ out of $n$ Bootstrap with $m = o(n)$ has been widely investigated. However, the choice of m is critical in ensuring the optimal performance of the method. We studied the convergence rate of the $m$ out of $n$ Bootstrap for extremes of i.i.d. random variables and derived formulae for the optimal resample size $m$ in a univariate framework. An extension of the results to multivariate random variables is also given in the chapter. Results from a simulation study are presented as well. For the real data analysis, we have applied our methods to AQI data from several cities around the world and demonstrated the dire air quality condition in some of these cities. This work was done in collaboration with Prof. Soumendra Lahiri from Washington University in St. Louis.

# Chapter 2

# Quadratic Prediction of Time Series via Auto-Cumulants

## 2.1 Introduction

The prediction problem can often be parsed as an attempt to find an excellent "estimator"of a target random variable $Y$ given an available data random vector $\underline{X}$. Typically a joint distribution is posited, and the broader prediction problem involves determining the conditional distribution $Y|\underline{X}$. The mean (when it exists) of this conditional distribution is the conditional expectation $\mathbb{E}[Y|\underline{X}]$, and is known to minimize the mean square error (MSE) between $Y$ and all functions of $\underline{X}$. This general problem has received much attention since it can be used for many other prediction problems. When the joint distribution is Gaussian, the conditional expectation is a linear function of $\underline{X}$. This linear function is completely computable in terms of the first and second moments of the joint vector $(Y, \underline{X})$, as discussed in [24](Chapter 2).

Even when the joint vector is non-Gaussian, a practitioner still might use the linear solution – knowing that this solution has the minimal MSE among all linear estimators – because it is simple to compute. Nevertheless, there can be a substantial predictive loss when non-Gaussian features are present in the data, such as asymmetry and excess kurtosis [85], [23]. Nonlinearity in financial data has been documented in [64] and [2]; [104] discuss time irreversibility (i.e., nonlinearity) in macroeconomic data, whereas [59] discuss the implications of skewness in asset pricing. [98] discuss the importance of obtaining phase information for applications to image and speech reconstruction. [92] provides autocumulant calculations for popular econometric models, motivated by known nonlinearities in consumption and interest rate data. [130] provides an overview of the benefits of nonlinear time series analysis.

One can formulate nonlinear prediction by generalizing the linear minimizer of prediction MSE to a nonlinear solution, leading to the search for a "universal predictor"([73]). One approach to finding a universal predictor involves the Kolmogorov-Gabor (KG) polynomial, a finite-lag version of the universal predictor's truncated Volterra expansion ([138], [118], [22]), which is discussed in [134] and [127]). Recent work in econometrics utilizing the KG approach include [76] and [28]. There is also a substantial literature on bilinear processes ([107], [102], [105], [64], [83], [82], and [128]) and the uses of bi-spectral analysis ([51] and [108]). In this chapter, we formulate this universal predictor in terms of higher-order moments of a time series, which potentially could be estimated nonparametrically, avoiding the need to specify a model.

To fix ideas, let $\{X_t\}$ be a zero mean stationary time series with autocovariance function $\gamma(\cdot)$, and suppose that we observe a finite stretch $X_1, \ldots, X_T$ of the time series. The prediction target is a future value $Y = X_{T+L}$ of the $\{X_t\}$ process (for some given $L \geq 1$, the forecast lead) on the basis of the past $P$ observations $\underline{X} = [X_{T-P+1}, \ldots, X_T]'$ (where $'$ denotes the transpose). In the case that $P = T$ we use all the available sample for prediction, but it may be advantageous to take $P < T$ when a large value of $P$ necessitates the estimation of many parameters.

As a first step, one might consider predictors that are quadratic functions of $\underline{X}$, with the understanding that first and second moments will no longer be sufficient to describe the solution – as established in (2.3.3) below. Such a quadratic predictor could be computed with either parametric or nonparametric approaches: if a particular parametric model is supposed that specifies the needed auto-cumulants (but perhaps is agnostic about other auto-cumulants not needed to describe the solution), then one can simply plug into the formula once the model has been fitted. A nonparametric approach would forego modeling, and instead proceed with consistent estimation of the auto-cumulants. Adequate estimation of polyspectra has already been addressed in the statistics literature: see [21], [111], [81], [10], and [12]. In the signal processing literature there is also much attention given to the subject: [96], [91], and [95]. See also [63] which investigate the cross bi-spectrum in spatial data.

Given that much is already known about polyspectral estimation and the properties of specific kinds of nonlinear models, our focus in this chapter is instead on the properties of nonlinear predictors, elicited through the analysis of polyspectra and the development of recursive computational algorithms. With a deeper understanding of the properties of nonlinear predictors (of the multivariate polynomial type), we aim to facilitate the application of nonlinear prediction to time series data. We remark that the use of polynomials, and the quadratic basis in particular, is not canonical, but is merely a convenient, agnostic choice that is motivated by the Volterra expansion of the optimal predictor; this also indicates that taking cubic and quartic terms will further improve the MSE. One can utilize different bases, but in order to do calculations one must either know the type of nonlinear process or have resort to auto-cumulant computations. Our proposal is useful in contexts where the exact specification of the nonlinear process is not known, or there are computational difficulties associated with finding its optimal predictor.

We now briefly describe the specific findings and main contributions of the chapter. For the mean corrected random variates $Y = X_{T+L}$ and $\underline{X} = [X_{T-P+1}, \ldots, X_T]'$, the class of quadratic predictors we consider here has the generic form:

$$g(\underline{X}) = a + \underline{b}' \, \underline{X} + \underline{X}' \, B \, \underline{X}$$

for some constant $a$, coefficient vector $\underline{b}$ for the linear part, and (symmetric) coefficient matrix $B$ for the quadratic part. The zero mean condition on $Y$ and unbiasedness considerations suggest taking $a = -\mathbb{E}[\underline{X}' \, B \, \underline{X}] = -\text{tr.}(B \, \Sigma_{\underline{X},\underline{X}})$, leaving $\underline{b}$ and $B$ as the free parameters of the quadratic predictor; here $\Sigma_{\underline{X},\underline{X}} = \text{Var}[\underline{X}]$ is the covariance matrix of $\underline{X}$ with $jk$th entry $\gamma(j-k)$, and tr. is the trace operator. Using some suitable vectorization steps, we develop below a generalized version of the Yule-walker equations for the quadratic prediction problem, and derive an explicit expression for the optimal choices of $\underline{b}$ and $B$ under the squared error loss. We also derive necessary and sufficient conditions under which the quadratic prediction approach improves upon linear prediction, providing a complete characterization – see Theorem 2.4.1. We call a stochastic process *second order forecastable* if it satisfies these necessary and sufficient conditions. Thus, there is benefit in using the quadratic prediction approach *only* for such second order forecastable processes.

We also give examples of some popular nonlinear time series models, such as ARCH and GARCH models (which are shown to have some strange behavior) and nonlinear Hermite processes that are second order forecastable. On computational aspects of the proposed methodology, we give an outline of an algorithm for computing the higher order auto-cumulants and associated polyspectra that yield the coefficients in the optimal quadratic predictor. Numerical results reported in the chapter show nontrivial improvements over linear prediction, with relative gains in mean squared (prediction) error being as high as 25%; See Table 1 in Section 6 below. To summarize, the key contributions of the chapter are to develop a new quadratic prediction methodology, provide results for identification of second order forecastable processes (where the quadratic prediction method can offer improvements over classical linear prediction theory), and develop necessary computing tools to make the methodology applicable in practice.

The rest of the chapter is organized as follows. In Section 2.2, we describe the quadratic prediction methodology. In Section 2.3, we quantify potential gains from using the quadratic prediction approach over the traditional linear prediction methodology. In Section 2.4 we provide a complete characterization of the class of second order forecastable processes under some regularity conditions, and Section 2.5 gives some examples of such processes arising from nonlinear time series models. Computational aspects and results from a numerical study are reported in Section 2.6. In Section 2.7 we provide two real data examples involving the Unemployment Rate and the Wolfer Sunspots data. The quadratic predictors give reductions of 16.5% and 29.2% in the mean squared errors in the two cases respectively. The Appendix 2.8 contains proofs of the main results, as well as some technical details related to derivation of the auto-cumulants of Hermite processes.

## 2.2 The Quadratic Yule Walker Equations and the Best Quadratic Predictor

To state the formula for the quadratic predictor, we require the notion of auto-moment. With $\mathbb{Z}$ denoting the set of all integers and with $t_0 = 0$, for $r \geq 2$, let

$$\gamma_r(t_1, \ldots, t_{r-1}) = \mathbb{E} \prod_{j=0}^{r-1} X_{t_j}, \quad t_1, \ldots, t_{r-1} \in \mathbb{Z}$$

denote the $r$th order auto-moment function of the stationary process $\{X_t\}$. In particular, for $r = 2$, $\gamma_2(t) \equiv \gamma(t) = \mathbb{E} X_0 X_t$ is the autocovariance function of $\{X_t\}$.

Next, recall that $Y = X_{T+L}$ and $\underline{X}' = [X_{T-P+1}, \ldots, X_T]$ have mean zero. The entire random vector consisting of $\underline{X}$ with $Y$ as the final component is denoted $\underline{Z}$. If $\underline{Z}$ is Gaussian, $\mathbb{E}[Y|\underline{X}] = \underline{b}' \, \underline{X}$ with $\underline{b}' = \text{Cov}[Y, \underline{X}] \, \text{Var}[\underline{X}]^{-1}$. Moreover, even when $\underline{Z}$ is non-Gaussian this same solution minimizes the linear prediction problem

$$\mathbb{E}[(Y - \underline{b}' \, \underline{X})^2] = \text{Var}[Y] - 2 \, \underline{b}' \, \text{Cov}[\underline{X}, Y] + \underline{b}' \, \text{Var}[\underline{X}] \, \underline{b},$$

which is verified by computing the gradient and Hessian with respect to $\underline{b}$. Whereas the generic nonlinear prediction problem minimizes the MSE difference between $Y$ and all functions $g(\underline{X})$, the quadratic problem posits a predictor of the form

$$\begin{aligned} g(\underline{X}) &= \underline{b}' \, \underline{X} + \underline{X}' \, B \, \underline{X} - \mathbb{E}[\underline{X}' \, B \, \underline{X}] \\ &= \sum_{t=1}^{P} b_t \, X_{T+1-t} + \sum_{1 \leq s \leq u \leq P} B_{su} \left( X_{T+1-s} X_{T+1-u} - \gamma(u-s) \right), \end{aligned} \tag{2.2.1}$$

where $\underline{b} = (b_1, \ldots, b_P)' \in \mathbb{R}^P$ and $B$ is a $P \times P$ real matrix with entries $B_{su}$. The second term of (2.2.1) is a bilinear form, involving a two-dimensional array (i.e., the matrix $B$). The centering of this bilinear form is needed to ensure that the predictor $g(\underline{X})$ has mean zero, as otherwise a bias is introduced. It is important to specify that $B$ is (weakly) lower-triangular, as we seek to identify the coefficients of $B$ that minimize MSE, and these will not be identifiable unless we restrict the bilinear form. The entries of the linear form are not constrained. The optimal quadratic predictor is obtained by choosing the coefficients $\underline{b}$ and $B$ such that $\mathbb{E}[(Y - g(\underline{X}))^2]$ is minimized. This can be done by taking partial derivatives of the quadratic form with respect to the free coefficients $\{b_t : 1 \leq t \leq P\}$ and $\{B_{su} : 1 \leq s \leq u \leq P\}$. Setting these partial derivatives equal to zero, one obtains the equations

$$\begin{aligned}
&\sum_{t'=1}^{P} b_{t'}\, \gamma(t-t') + \sum_{1 \leq s' \leq u' \leq P} B_{s'u'}\, \gamma_3(t-s', t-u') = \gamma(L+t-1) \\
&\sum_{1 \leq s' \leq u' \leq P} B_{s'u'}\, \left(\gamma_4(u-s', u-u', u-s) - \gamma(u-s)\gamma(u'-s')\right) \\
&\qquad\qquad + \sum_{t'=1}^{P} b_{t'}\, \gamma_3(u-t', u-s) = \gamma_3(L+u-1, u-s). \qquad (2.2.2)
\end{aligned}$$

These are the *generalized Yule-Walker equations* for the quadratic prediction problem. Note that in addition to the autocovariance function, it involves the third and fourth order auto-moments of the $\{X_t\}$ process.

The equations in (2.2.2) can be expressed in a compact form and solved for the coefficients by recasting them using suitable matrix notation. To that end, note that the bilinear form can be expressed in the following two alternative ways :

$$\underline{X}'\, B\, \underline{X} = \text{tr.}\{\underline{X}\, \underline{X}'\, B\} = \text{vec}[B]'\, \text{vec}[\underline{X}\, \underline{X}'],$$

where recall that for an $m \times n$ matrix $A$, vec$[A]$ is the $mn \times 1$ vector obtained by stacking the columns of $A$. The lower-triangular structure of $B$ ensures that we can without loss of generality consider the (weak) vech (where only the elements on or below the diagonal are included in the column-wise vectorization of $B$) in lieu of vec in the above. Let $\underline{W} =$ vech$[\underline{X}\,\underline{X}']$, set $\beta' = [\underline{b}', \text{vech}[B]']$, and let $\Sigma_{A,B} = \text{Cov}[A, B]$ for any random vectors $A$ and $B$. Then, we have the following result on the best quadratic predictor (BQP) of $Y$:

**Proposition 2.2.1.** Suppose that $\Sigma_{\underline{X},\underline{X}}$ and $S \equiv \Sigma_{\underline{W},\underline{W}} - \Sigma_{\underline{W},\underline{X}}\,\Sigma^{-1}_{\underline{X},\underline{X}}\,\Sigma_{\underline{X},\underline{W}}$ are invertible. Then, the BQP of $Y$ is given by $[\underline{X}', \underline{W}']\widehat{\beta}$ where

$$\widehat{\beta} = \begin{bmatrix} \Sigma^{-1}_{\underline{X},\underline{X}}\,\Sigma_{\underline{X},Y} - \Sigma^{-1}_{\underline{X},\underline{X}}\,\Sigma_{\underline{X},\underline{W}}\,S^{-1}\left(\Sigma_{\underline{W},Y} - \Sigma_{\underline{W},\underline{X}}\,\Sigma^{-1}_{\underline{X},\underline{X}}\,\Sigma_{\underline{X},Y}\right) \\ S^{-1}\left(\Sigma_{\underline{W},Y} - \Sigma_{\underline{W},\underline{X}}\,\Sigma^{-1}_{\underline{X},\underline{X}}\,\Sigma_{\underline{X},Y}\right) \end{bmatrix}.$$

Note that the optimal weights $\widehat{\beta}$ depend on the autocovariance function as well as the third and fourth order auto-moments of the $\{X_t\}$ process, as expected from the generalized Yule-Walker equations (2.2.2). In the next section we explore the benefits of using the quadratic predictor over its linear counterpart, leading to the quadratic prediction principle that gives a criterion for establishing superiority of the BQP over the best linear predictor (BLP).

## 2.3 Improvement over the linear predictor

Clearly, the quadratic portion of the solution disappears entirely if and only if

$$\Sigma_{\underline{W},Y} - \Sigma_{\underline{W},\underline{X}}\,\Sigma^{-1}_{\underline{X},\underline{X}}\,\Sigma_{\underline{X},Y} = 0, \qquad (2.3.1)$$

in which case $\widehat{\underline{b}}$ (the first component of $\widehat{\beta}$) also reduces to the linear solution $\Sigma_{\underline{X},\underline{X}}^{-1} \, \Sigma_{\underline{X},Y}$. An important observation here is that condition (2.3.1) involves the third auto-cumulant, but not the fourth (and higher) order auto-cumulants. The quantity on the left of (2.3.1) can also be viewed as

$$\Sigma_{\underline{W},\widehat{E}^{(1)}}, \tag{2.3.2}$$

where, with $\widehat{Y}^{(1)}$ denoting the BLP of $Y = X_{t+L}$, the random variable $\widehat{E}^{(1)} = Y - \widehat{Y}^{(1)}$ gives the error in the linear predictor. Hence it follows that there is no benefit to quadratic prediction if and only if the error in the linear predictor is uncorrelated with $\underline{W} = \text{vech}[\underline{X} \, \underline{X}']$. This is certainly the case for Gaussian $\underline{Z}$, where all third auto-moments are zero; even though $S$ is invertible (it is now given by $\Sigma_{\underline{W},\underline{W}}$), the condition (2.3.1) is true. In general, the full expression for the BQP, $\widehat{Y}^{(2)}$, is

$$\widehat{Y}^{(2)} = \widehat{Y}^{(1)} + \Sigma_{\widehat{E}^{(1)},\underline{W}} \, S^{-1} \left[ \underline{W} - \Sigma_{\underline{W},\underline{X}} \, \Sigma_{\underline{X},\underline{X}}^{-1} \, \underline{X} \right]. \tag{2.3.3}$$

This expresses the quadratic estimator as the linear estimator plus a modification that is only present if (2.3.1) is violated. Also, this modification is based on the difference between $\underline{W}$ and its linear projection upon $\underline{X}$. It is easy to check that the minimum mean squared prediction error (attained by the BQP, $\widehat{Y}^{(2)}$) is

$$\Sigma_{Y,Y} - \Sigma_{Y,\underline{X}} \, \Sigma_{\underline{X},\underline{X}}^{-1} \, \Sigma_{\underline{X},Y} - \Sigma_{\widehat{E}^{(1)},\underline{W}} \, S^{-1} \, \Sigma_{\underline{W},\widehat{E}^{(1)}},$$

where the first two terms correspond to the prediction error for a linear problem. In other words, the efficiency loss of using a linear estimator when a quadratic is warranted is the non-negative quantity

$$\Sigma_{\widehat{E}^{(1)},\underline{W}} \, S^{-1} \, \Sigma_{\underline{W},\widehat{E}^{(1)}}, \tag{2.3.4}$$

which is non-negligible when (2.3.1) is violated. We summarize this discussion in the following result.

**Theorem 2.3.1.** Suppose that $\Sigma^{-1}_{\underline{X},\underline{X}}$ and $S^{-1}$ exist. Then,
(i) the optimal quadratic predictor of $Y$ under the squared error loss function is given by (2.3.3).
(ii) The minimal quadratic prediction error is given by

$$\Sigma_{Y,Y} - \Sigma_{Y,\underline{X}}\,\Sigma^{-1}_{\underline{X},\underline{X}}\,\Sigma_{\underline{X},Y} - \Sigma_{\widehat{E}^{(1)},\underline{W}}\,S^{-1}\,\Sigma_{\underline{W},\widehat{E}^{(1)}}.$$

(iii) The quadratic predictor improves upon the linear predictor if and only if (2.3.1) fails or equivalently,

$$\Sigma_{\widehat{E}^{(1)},\underline{W}}\,S^{-1}\,\Sigma_{\underline{W},\widehat{E}^{(1)}} \neq 0.$$

Theorem 2.3.1 gives the formulae for the best quadratic predictor and the minimal prediction error of a quadratic predictor. It also precisely describes situations where quadratic prediction may improve upon linear prediction. Thus, Theorem 2.3.1 leads us to the following general *quadratic prediction principle*:
*"There is no benefit to quadratic prediction if and only if the linear prediction error is orthogonal to quadratic functions of the data, i.e., when quantity (2.3.2) equals zero."*

## 2.4 Second Order Forecastable Processes

In this section, we present a characterization of time series models where the quadratic prediction improves on linear prediction. For definiteness, we shall restrict attention to the case where the goal is to predict $Y = X_{t+1}$ based on an infinite past $\{X_t, X_{t-1}, \ldots, \}$. We say a time series is a *second order forecastable processes* if and only if the MSE of its one-step ahead BQP is less than the MSE of its one-step ahead BLP; this generalizes the classical set of linear processes, for which the one-step ahead BLP is the conditional expectation. Below, we develop some key results on linear projections of quadratic terms, and provide a characterization of second order forecastable processes. A stationary process with moments of all orders can be described in terms of its polyspectra; this is the approach to a frequency domain analysis of time series advocated by [18]. We proceed by describing these results and relating them to the familiar case of a linear process. Any strictly stationary time series $\{X_t\}$ with moments of all orders has auto-cumulant functions $\kappa$ of order $r+1$ (for $r \geq 1$) defined via

$$\kappa_{r+1}(\underline{h}) = \mathrm{cum}[X_{t+h_1}, X_{t+h_2}, \ldots, X_{t+h_r}, X_t],$$

where $\underline{h} = [h_1, h_2, \ldots, h_r]'$ is a $r$-vector of lags. Note that the order of the auto-cumulant corresponds to the number of variables included $(r+1)$, not the number of lags $(r)$. Strict stationarity – or more generally, stationarity of order $(r+1)$ – guarantees that $\kappa_{r+1}$ is only a function of the lags, and hence $t$ is immaterial.

Recall the standing assumption that $\mathbb{E}[X_t] = 0$, and also recall the definition of the auto-moment functions $\gamma$ of order $(r+1)$:

$$\gamma_{r+1}(\underline{h}) = \mathbb{E}[X_{t+h_1} X_{t+h_2} \ldots X_{t+h_r} X_t].$$

For $r = 1, 2$, we have $\kappa_{r+1} = \gamma_{r+1}$, but for $r \geq 3$ the auto-cumulant and auto-moment functions are distinct.

For the discussion below, we fix $r \geq 1$ and assume that the order $(r+1)$ auto-cumulant function, denoted simply by $\kappa_{r+1}(\underline{h}) = \kappa(\underline{h})$ for ease of exposition, is absolutely summable over $\underline{h} \in \mathbb{Z}^r$. With such an assumption, the polyspectrum of order $(r+1)$ is well-defined. The corresponding polyspectral density of order $(r+1)$ is given by

$$f(\underline{\lambda}) = \sum_{\underline{h} \in \mathbb{Z}^r} \kappa(\underline{h}) \exp\{-i \, \underline{\lambda}' \underline{h}\},$$

where $i = \sqrt{-1}$ and where $\underline{\lambda} = [\lambda_1, \ldots, \lambda_r]'$ denotes a $r$-vector of frequencies. [21] provides an elegant discussion as to why it is preferable to consider the Fourier transform of auto-cumulants rather than that of auto-moments. When applying a linear filter $\Psi(B) = \sum_{j \in \mathbb{Z}} \psi_j B^j$ to such an $\{X_t\}$, yielding a new $\{Y_t\}$ defined by $Y_t = \Psi(B) X_t$, one can relate the polyspectra of the filter output to the polyspectra of the filter input. Let $f_y$ and $f_x$ denote polyspectra of order $(r+1)$ for the $\{Y_t\}$ and $\{X_t\}$ processes; then by Theorem 2.8.1 of [18],

$$f_y(\underline{\lambda}) = \prod_{j=1}^{r} \Psi(e^{-i\lambda_j}) \, \Psi(e^{i \sum_{j=1}^{r} \lambda_j}) \, f_x(\underline{\lambda}). \tag{2.4.1}$$

Recall that it follows from the Wold decomposition ([90]) that a purely non-deterministic stationary time series $\{X_t\}$ can be expressed as $X_t = \Psi(B) \, Z_t$, where $\Psi(z)$ is a power series such that $\Psi(0) = 1$, and $\{Z_t\}$ is a white noise process, with its $r$th cumulant denoted by $\mu_r$. When $\{Z_t\}$ is i.i.d., we say the process $\{X_t\}$ is *linear*; this appellation is connected to the fact that the minimal MSE one-step ahead forecast function is linear in the past data. In such a case, the polyspectrum of order $(r+1)$ is given by

$$f(\underline{\lambda}) = \mu_{r+1} \prod_{j=1}^{r} \Psi(e^{-i\lambda_j}) \, \Psi(e^{i \sum_{j=1}^{r} \lambda_j}), \tag{2.4.2}$$

which follows from (2.4.1) and the fact that all polyspectra for an i.i.d. sequence are equal to the constant cumulant $\mu_{r+1}$. If we relax the assumption that $\{Z_t\}$ is i.i.d., the above formula will no longer be valid; potentially the $\{Z_t\}$ has a non-constant polyspectra. Conversely, given a polyspectrum it is possible to factorize it under certain conditions; see

[126]. For the case $r = 1$, the well-known spectral factorization theorem ([90]) yields

$$f(\lambda) = \mu_2 \, \Psi_2(e^{-i\lambda}) \, \Psi_2(e^{i\lambda}), \tag{2.4.3}$$

where $\Psi_2(z)$ is a power series such that $\Psi_2(0) = 1$. This factorization is possible when the process is invertible, i.e., the spectral density is strictly positive.

Within the above context we now provide an equivalent characterization of second order forecastable processes. The endeavor to predict $X_{t+1}$ in terms of both linear and quadratic functions of past data can be re-expressed as a linear function of both $\{X_{t-\ell}\}_{\ell\geq 0}$ and $\{X_{t-j}X_{t-k}\}_{j,k\geq 0}$. Using Lemma 2 of [9] the forecast only depends on the linear portion if and only if

$$\mathrm{Cov}[X_{t+1}, X_{t-j}X_{t-k} - \widehat{X_{t-j}X_{t-k}}] = 0 \tag{2.4.4}$$

for all $j, k \geq 0$, where $\widehat{X_{t-j}X_{t-k}}$ denotes the linear prediction of $X_{t-j}X_{t-k}$ on the basis of $\{X_{t-\ell}\}_{\ell\geq 0}$. In other words, if the above covariance is non-zero for some $j$ and $k$, the process is second order forecastable – following the ideas discussed in Sections 2 and 3. To understand this condition better, we first derive $\widehat{X_{t-j}X_{t-k}}$; this is expressible as the mean $\gamma(k-j)$ plus some causal filter $\Pi^{(j,k)}(B)$ applied to $X_t$, i.e.,

$$\widehat{X_{t-j}X_{t-k}} = \gamma(k-j) + \Pi^{(j,k)}(B)X_t \equiv \gamma(k-j) + \sum_{h\geq 0} \pi_h^{(j,k)} X_{t-h}. \tag{2.4.5}$$

To state the formula for this filter, we need new notations. Let the order $(r+1)$ autocumulant generating function be denoted as

$$f_{r+1}(\underline{z}) = \sum_{\underline{h}\in\mathbb{Z}^r} \kappa(\underline{h}) \, z_1^{h_1} \cdots z_r^{h_r},$$

which reduces to the polyspectral density $f_{r+1}(\underline{\lambda})$ when $z_j = e^{-i\lambda_j}$ for $j = 1, \ldots, r$. Next, for any function $g(z)$ that is analytic on an open annulus containing the unit circle of $\mathbb{C}$, let

$$(g(z))_z = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(e^{-i\lambda})\, d\lambda.$$

Also, for any Laurent series $\Psi(z)$ (see [3]), let $[\Psi(z)]_r^s = \sum_{j=r}^s \psi_j z^j$ for integers $r \leq s$. The following result gives the linear projection of the quadratic term $X_{t-j} X_{t-k}$ for any $j, k \geq 0$.

**Proposition 2.4.1.** Let $\{X_t\}$ be strictly stationary with third moments, and absolutely summable auto-cumulants of order 2 and 3. Suppose the spectral density is positive, so that the factorization (2.4.3) exists. Then for any $j, k \geq 0$, the power series in $z$ defined by

$$\Pi^{(j,k)}(z) = \frac{1}{\mu_2} \left[ z^j \left( y^{k-j} f_3(z\overline{y}, y) \right)_y / \Psi_2(z^{-1}) \right]_0^\infty \Psi_2(z)^{-1}$$

yields the filter $\Pi^{(j,k)}(B)$ that generates the optimal linear estimate of $X_{t-j} X_{t-k}$ via (2.4.5).

Using Proposition 2.4.1, we can now prove our main result about the characterization of the property of being a second order forecastable process. The main assumption is that there are no zeroes in the spectral density. This is not a substantially new restriction, because if the spectral density is not positive then linear forecasting is also impossible.

**Theorem 2.4.1.** Let $\{X_t\}$ be strictly stationary with fourth moments, and absolutely summable auto-cumulants of order 2, 3, and 4. Suppose the spectral density is positive, so that the factorization (2.4.3) exists. Then $\{X_t\}$ is second order forecastable if and only if the expression

$$\left( \left( z^{j+1} y^{k+1} f_3(z, y) / \Psi_2(\overline{zy}) \right)_y \right)_z \qquad (2.4.6)$$

is nonzero for some $j, k \geq 0$.

**Remark 2.4.1.** The condition (2.4.4) is also equivalent to

$$\mathrm{Cov}[X_{t+1} - \widehat{X}_{t+1}, X_{t-j}X_{t-k}] = 0$$

for all $j, k \geq 0$, and therefore by Theorem 2.4.1 condition (2.4.6) says that the linear forecast error is orthogonal to all quadratic functions of the past, i.e., the quadratic prediction principle holds.

**Remark 2.4.2.** As a consequence of Theorem 2.4.1, we shall say that any process $\{X_t\}$ satisfying those hypotheses is by definition a second order forecastable process if and only if (2.4.6) is nonzero for some $j, k \geq 0$. It is immediate that Gaussian processes and causal linear processes are not second order forecastable: in the former case, $f_3(z, y) \equiv 0$, and in the latter case, using (2.4.2),

$$f_3(z, y)/\Psi_2(\overline{zy}) = \mu_3\, \Psi_2(z)\Psi_2(y),$$

on the unit circle, so that (2.4.6) equals zero for all $j, k \geq 0$.

To check whether a given nonlinear process is second order forecastable, one needs the spectral factorization $\Psi_2$ (see [89] for algorithms) and an expression for the bi-spectral density, so that (2.4.6) can be directly calculated. However, a process that is second order forecastable may also be third order forecastable – this just says that the best cubic predictor has lower MSE than the BQP. In fact, writing $\mathcal{F}_d$ for the class of $d$th order forecastable processes, we see that $\mathcal{F}_2 \subset \mathcal{F}_3 \subset \ldots$, since trivially any BQP can be written as a cubic predictor where the third order terms are zero. In order to partition the space of nonlinear processes, we take the intersection of each $\mathcal{F}_d$ with the complement of all higher order classes, i.e., $\mathcal{A}_d$ is the set of *order $d$ augurable processes*, defined as processes in $\mathcal{F}_d$ such that the best order $k$ predictor, for all $k > d$, gives no reduction to the MSE. This is a stronger condition, so we refer to an augury rather than a forecast; it follows that $\mathcal{A}_d$ consists of all processes for which the Volterra expansion of the one-step ahead conditional expectation truncates to order $d$. Although the augurable classes are more elegant, since they form a partition, it is more difficult to check membership.

## 2.5 Illustrations of Second Order Forecastable Processes

We present results for two classes of second order forecastable processes, each of which is simple to simulate and study.

### 2.5.1 Hermite Processes

A class of nonlinear processes for which the auto-moments can be calculated fairly directly is the Hermite class. Let $\{Z_t\}$ be a mean zero, stationary Gaussian with autocovariance $c(h)$ such that $c(0) = 1$. The Hermite polynomials are defined for $k \geq 1$ ($H_0 \equiv 1$) as

$$H_k(x) = \frac{(-1)^k}{\sqrt{k!}} \, e^{x^2/2} \, \partial_x^k e^{-x^2/2}.$$

For a sequence of coefficients $\{J_k\}_{k\geq 1}$ that are square summable, let $g(x) = \sum_{k=1}^{\infty} J_k \, H_k(x)$ and define a Hermite process via $X_t = g(Z_t)$. This is a zero mean nonlinear process; see [67] for more details. We describe a general method for computing the auto-moments in the Appendix. The derivation involves the Hermite generating function and some combinatorial concepts, and may be of independent interest; see the Appendix for details.

### Exponential Hermite Process

Here we set $g(x) = e^x - \mu$ for $\mu = e^{c(0)/2}$, so that $\{X_t\}$ is a lognormal process. The automoments are:

$$\begin{aligned}
\gamma(h_1) &= \mu^2 \left(\exp\{c(h_1)\} - 1\right) \\
\gamma_3(h_1, h_2) &= \mu^3 \, \big(\exp\{c(h_1) + c(h_2) + c(h_1 - h_2)\} \\
&\quad - \exp\{c(h_1)\} - \exp\{c(h_2)\} - \exp\{c(h_1 - h_2)\} + 2\big) \\
\gamma_4(h_1, h_2, h_3) &= \mu^4 \, \big(\exp\{c(h_1) + c(h_2) + c(h_3) + c(h_1 - h_2) + c(h_1 - h_3) \\
&\quad + c(h_2 - h_3)\} - \exp\{c(h_1 - h_2) + c(h_1 - h_3) + c(h_2 - h_3)\} \\
&\quad - \exp\{c(h_2) + c(h_3) + c(h_2 - h_3)\} \\
&\quad - \exp\{c(h_1) + c(h_3) + c(h_1 - h_3)\} \\
&\quad - \exp\{c(h_1) + c(h_2) + c(h_1 - h_2)\} \\
&\quad + \exp\{c(h_2 - h_3)\} + \exp\{c(h_1 - h_3)\} + \exp\{c(h_1 - h_2)\} \\
&\quad + \exp\{c(h_3)\} + \exp\{c(h_2)\} + \exp\{c(h_1)\} - 3\big) \,,
\end{aligned}$$

which are easily derived using the formula for the expectation of the lognormal distribution. For such a process, the minimal MSE predictor among all function classes is known to be the exponential of the sum of the linear estimator plus half its MSE; hence we know there is a benefit to nonlinear prediction, and the lognormal process will generally be a second order forecastable process.

Of course, in practice we may not know that our data follows such a lognormal process, or may have difficulty fitting the model; if we use nonparametric estimation of the automoments (see below) and utilize quadratic prediction, we can expect a benefit even when the exact model specification is unknown. For example, if $\{Z_t\}$ is an MA($q$) then the third auto-moment function is zero whenever $|h_1|, |h_2| > q$, or $|h_1|, |h_1 - h_2| > q$, or $|h_2|, |h_1 - h_2| > q$. In the case $q = 1$, we obtain

$$\begin{aligned}
&\gamma_3(0,0) = \mu^3\,(e^3 - 3e + 2) \\
&\gamma_3(\pm 1, 0) = \gamma_3(0, \pm 1) = \gamma_3(1,1) = \gamma_3(-1,-1) = \mu^3\,(e^{1+2c(1)} - 2e^{c(1)} - e + 2) \\
&\gamma_3(2,1) = \gamma_3(1,2) = \gamma_3(-2,-1) = \gamma_3(-1,-2) = \mu^3\,(e^{2c(1)} - 2e^{c(1)} + 1),
\end{aligned}$$

all other values being zero. Hence the bi-spectrum is

$$\begin{aligned}
f(z,y) = \mu^3 \Big( & (e^3 - 3e + 2) \\
& + (e^{2c(1)} - 2e^{c(1)} + 1)\,(z^2 y + z^{-2} y^{-1} + z y^2 + z^{-2} y^{-2}) \\
& + (e^{1+2c(1)} - 2e^{c(1)} - e + 2)\,(z + z^{-1} + y + y^{-1} + zy + z^{-1} y^{-1}) \Big).
\end{aligned}$$

However, the autocovariance function corresponds to an MA(1) process, and hence $\Psi_2(z) = 1 - \theta z$ for some $\theta$ determined from $\gamma(0)$ and $\gamma(1)$ via spectral factorization. It follows that (2.4.6) equals $\sum_{\ell=0}^{\infty} \theta^\ell \gamma_3(\ell - j - 1, \ell - k - 1)$, which is nonzero in general.

### Squared Hermite Process

Another case is given by assuming that only $J_1$ and $J_2$ are non-zero, so that the process is expressed as:

$$X_t = J_1\, H_1(Z_t) + J_2\, H_2(Z_t) = J_1\, Z_t + J_2\, Z_t^2 - J_2.$$

Using the general method in the Appendix, the auto-moments are given in terms of $J_1$ and $J_2$ as follows:

$$\begin{aligned}
\gamma(h_1) &= J_1^2\, c(h_1) + J_2^2\, c(h_1)^2 \\
\gamma_3(h_1,h_2) &= J_2^3\, 8^{1/2}\, c(h_1)\, c(h_2)\, c(h_1-h_2) \\
\gamma_4(h_1,h_2,h_3) &= J_2^4\, \left(c(h_3)^2\, c(h_1-h_2)^2 + c(h_2)^2\, c(h_1-h_3)^2 + c(h_1)^2\, c(h_3-h_2)^2\right) \\
&\quad + 4J_2^4\, (c(h_1)\, c(h_2)\, c(h_1-h_3)\, c(h_2-h_3) + c(h_2)\, c(h_3)\, c(h_1-h_2)\, c(h_1-h_3) \\
&\qquad +c(h_1)\, c(h_3)\, c(h_1-h_2)\, c(h_2-h_3)) \\
&\quad + J_1^4\, (c(h_3)\, c(h_1-h_2) + c(h_1)\, c(h_2-h_3) + c(h_2)\, c(h_1-h_3)) \\
&\quad + J_1^2\, J_2^2\, \left(c(h_3)\, c(h_1-h_2)^2 + c(h_1-h_2)\, c(h_3)^2 + c(h_1-h_3)\, c(h_2)^2\right. \\
&\qquad \left.+c(h_2-h_3)\, c(h_1)^2 + c(h_1)\, c(h_2-h_3)^2 + c(h_2)\, c(h_1-h_3)^2\right) \\
&\quad + 2\, J_1^2\, J_2^2\, (c(h_1)\, c(h_1-h_2)\, c(h_2-h_3) + c(h_1)\, c(h_2)\, c(h_1-h_3) \\
&\qquad +c(h_2)\, c(h_1-h_2)\, c(h_1-h_3) + c(h_1)\, c(h_1-h_3)\, c(h_2-h_3) \\
&\qquad +c(h_3)\, c(h_1-h_2)\, c(h_2-h_3) + c(h_2)\, c(h_1-h_3)\, c(h_2-h_3) \\
&\qquad +c(h_3)\, c(h_1-h_2)\, c(h_1-h_3) + c(h_1)\, c(h_3)\, c(h_2-h_3) \\
&\qquad +c(h_2)\, c(h_3)\, c(h_1-h_3) + c(h_2)\, c(h_3)\, c(h_1-h_2) \\
&\qquad +c(h_1)\, c(h_2)\, c(h_2-h_3) + c(h_1)\, c(h_3)\, c(h_1-h_2))\, .
\end{aligned}$$

So long as $J_2 \neq 0$, such squared Hermite processes can be second order forecastable. For example, if $\{Z_t\}$ is an MA$(q)$, then $\gamma_3(h_1,h_2) = 0$ if $|h_1| > q$ or $|h_2| > q$ or $|h_1-h_2| > q$. In the case that $q = 1$, we find

$$\begin{aligned}
\gamma_3(0,0) &= J_2^3\, 8^{1/2} \\
\gamma_3(\pm1,0) = \gamma_3(0,\pm1) = \gamma_3(1,1) = \gamma_3(-1,-1) &= J_2^3\, 8^{1/2}\, c(1)^2,
\end{aligned}$$

and is zero otherwise. Hence the bi-spectrum is

$$f(z,y) = J_2^3\, 8^{1/2} \left(1 + c(1)^2 \left(z + z^{-1} + y + y^{-1} + zy + [zy]^{-1}\right)\right),$$

and the autocovariance function corresponds to an MA(1) process. With $\Psi_2(z) = 1 - \theta z$ for some $\theta$ determined from $\gamma(0)$ and $\gamma(1)$, we find that (2.4.6) equals

$$\begin{aligned} J_2^3\, 8^{1/2} \Big(\theta^{j+1}\, 1_{\{k=j\}} + c(1)^2 \big[ (\theta^{j+2} + \theta^{j+1})\, 1_{\{k=j+1\}} \\ + (\theta^{j+2} + \theta^{j})\, 1_{\{k=j\}} + (\theta^{j+1} + \theta^{j})\, 1_{\{k=j-1\}} \big]\Big). \end{aligned}$$

Therefore, in many cases the squared Hermite processes are second order forecastable.

### 2.5.2 ARCH and GARCH Processes

The class of ARCH and GARCH processes is extremely popular in modeling the log returns of stocks and indices in the financial sector. The market efficiency axiom indicates that any forecasts of such a process should have MSE equal to the variance, i.e., there is no benefit to be gained by prediction. Technically this phenomenon is explained by the fact that the GARCH process is a white noise with non-trivial serial dependence; the squared process has non-trivial correlation, which allows the volatility to be forecasted with some success. Moreover, the conditional expectation formula for the one-step ahead forecast is linear for a GARCH process, so the best predictor (in the MSE sense) is linear and there can not be any further advantages in using the quadratic prediction. In particular, (2.4.6) must be zero for all $j, k \geq 0$; we verify this below. This example shows that the quadratic prediction need not always improve upon the linear prediction even for certain nonlinear processes. We also show, on the contrary, that backcasts of the GARCH process are nonlinear and hence, there is potential advantage if we use the quadratic approach for backcasting. We now provide details of the arguments needed to establish these claims.

Conventionally, GARCH processes are defined in terms of driving input $\{Z_t\}$ that are symmetric, the first cases being studied having involved Gaussian distributions ([16]). This was generalized to fat-tailed and asymmetric inputs – see [74], [132] and [70]. These adaptations were driven by empirical considerations; here, we can show directly how kurtosis and asymmetry in the inputs impact the auto-cumulants and polyspectra of the GARCH process, and use them to verify (2.4.6).

The GARCH(p,q) process is given by

$$\begin{aligned} X_t &= \sigma_t \, Z_t \\ \sigma_t^2 &= a_0 + \sum_{j=1}^{p} a_j \, X_{t-j}^2 + \sum_{j=1}^{q} b_j \, \sigma_{t-j}^2, \end{aligned}$$

where $Z_t \sim$ i.i.d.$(0,1)$. Let $\omega(x) = \sum_{j=1}^p a_j x^j$ and $\theta(x) = 1 - \sum_{j=1}^q b_j x^j$. Set $\pi(z) = 1 - \omega(z)/\theta(z)$ and $\phi(z) = \theta(z) - \omega(z)$, so that $\pi(z) = \phi(z)/\theta(z)$. This power series will be written with a minus convention, i.e., $\pi(z) = 1 - \sum_{\ell \geq 1} \pi_\ell z^\ell$. Because the GARCH is a white noise, $f(\lambda) \equiv f_2(\lambda) = \mathbb{E}[X^2]$ for all $\lambda$ , and

$$f_3(z, y) = \mathbb{E}[X^3] \left( \pi(zy)^{-1} + \pi(z^{-1})^{-1} + \pi(y^{-1})^{-1} - 2 \right).$$

The expression for the fourth auto-cumulant function is omitted because it is extremely complicated, although the special case of the autocovariance for $\{X_t^2\}$ has a nice formula due to [16]. Because $\Psi_2(z) \equiv 1$ (i.e., the GARCH process is a white noise), it follows that (2.4.6) equals zero for all $j, k \geq 0$. Thus, the forecast function is linear (and equals zero).

Although the GARCH process is not a linear process, contrary to our intuition, there is no advantage in using the quadratic prediction in this case.

If instead we examine one-step behind backcasts, the quadratic predictor has an advantage over the linear. To see this, observe that backcasting is equivalent to forecasting the time-reversed GARCH process, for which we obtain (see Ch. 11, [90]):

$$f_3(z,y) = \mathbb{E}[X^3]\,\Big(\pi\big([zy]^{-1}\big)^{-1} + \pi(z)^{-1} + \pi(y)^{-1} - 2\Big), \text{ and}$$
$$\Big(\big(z^{j+1}y^{k+1} f_3(z,y)/\Psi_2(\overline{zy})\big)_y\Big)_z = \mathbb{E}[X^3]\,1_{\{j=k\}}\,\widetilde{\pi}_{j+1},$$

where $\sum_{h\geq 0}\widetilde{\pi}_h x^h \;=\; \pi(x)^{-1}$. As a result, the BQP for the time reserved process will be better than its linear counterpart.

## 2.6 Computational Matters and Numerical Examples

We have implemented the methodology of this chapter and applied it to various nonlinear processes, including the numerical examples reported below. In this section we describe how the computations are done, and summarize the results. Recall that $\underline{W} = \text{vech}[\underline{X}\,\underline{X}']$. Construction of the matrix $\Sigma_{\underline{W},\underline{W}}$ proceeds by first building a larger 4-array out of the covariance of vec$[\underline{X}\,\underline{X}']$ with itself, allowing for redundancies. The indices $i$,$j$,$k$,$\ell$ for the four dimensions of the array each range from 1 to $P$. In R, applying the *matrix* operator to a 4-array constructs a block matrix, whereby $j$ and $\ell$ are row and column block indices, and $i$ and $k$ are row and column indices within each block. For example, $3,1,4,1$ represents the $3,4$ entry in the upper left block of the matrix. In this ordering, the indices $i$ and $j$ correspond to various entries in the vector vec$[\underline{X}\,\underline{X}']$, conceived of as the transpose of

$$[\underline{X}'\,X_{T-P+1}, \underline{X}'\,X_{T-P+2}, \ldots, \underline{X}'\,X_T],$$

or the collection of $\underline{X}' X_{T-P+j}$ with $i$ giving the index within each $\underline{X}$. It follows that the $i, j, k, \ell$ entry of the array equals

$$\mathrm{Cov}[X_{T-P+i} X_{T-P+j}, X_{T-P+k} X_{T-P+\ell}] = \gamma_4(i-\ell, j-\ell, k-\ell) - \gamma(i-j)\, \gamma(k-\ell).$$

(Note that auto-moment functions have many symmetries in their arguments, so there are many ways of writing the same quantity.) Once the entries of the 4-array have been filled in (inefficiently, by utilizing 4 nested loops over $T$ elements), then certain row and column entries corresponding to the lower triangular entries of $\underline{X}\,\underline{X}'$ are omitted from the matrization of the array. In a similar fashion, we construct $\Sigma_{\underline{X},\underline{W}}$.

Hence the formulas for the BQP and its prediction error can be applied once the auto-moments are known. In practice, these can be obtained by fitting nonlinear models and plugging in the parameter estimates; alternatively, in cases where it is not practical to fit a nonlinear model, we can use nonparametric estimators. As described in [21], simple estimators of auto-moments and auto-cumulants can be constructed as follows: for a sample of size $T$ and given a lag vector $\underline{h} = [h_1, h_2, \ldots, h_r]'$, define the index set

$$\mathcal{T}_{\underline{h}} = \cap_{\ell=0}^{r} \{1 - h_\ell, \ldots, T - h_\ell\}$$

with $h_0 = 0$. Note that if any $h_\ell$ is greater than $T$ or less than 1, then $\mathcal{T}_{\underline{h}} = \varnothing$. Then define the sample auto-moment of lag $\underline{h}$ as

$$\widehat{\gamma}_{r+1}(\underline{h}) = T^{-1} \sum_{t \in \mathcal{T}_{\underline{h}}} (X_{t+h_0} - \overline{X})(X_{t+h_1} - \overline{X}) \cdots (X_{t+h_r} - \overline{X}),$$

where $\overline{X} = T^{-1} \sum_{t=1}^{T} X_t$, and the sum is extended to be zero in cases where $\mathcal{T}_{\underline{h}} = \varnothing$. For any fixed $\underline{h}$, these estimators are asymptotically unbiased, with variance $O(T^{-1})$ assuming that all auto-cumulant functions are absolutely summable – see (2.6.1) of [18]. However, values of $\underline{h}$ such that $\mathcal{T}_{\underline{h}}$ is small imply that there will be some bias in finite samples; note that $|h_j| < P$ for $1 \leq j \leq r$, so by restricting $P$ to be much smaller than $T$ we can ensure that bias is minimized. However, we reckon that there is an efficiency loss to taking $P$ small. In the simulations below, we set $P$ to be a moderately large value to balance the trade off, following the choices considered in existing literature for other nonlinear prediction methods (cf. [46]).

For the simulation study, we consider the five different models as listed below (with the Innovation distributions are given in parentheses):

- **Model IA:** $X_t = A\epsilon_t + B\epsilon_{t-1}^2 - B$ (Unif(-1,1))
- **Model IB:** $X_t = A\epsilon_t + B\epsilon_{t-1}^2 - B$ (Exp(1)-1)
- **Model IIA:** $X_t = J_1\,H_1(Z_t) + J_2\,H_2(Z_t) = J_1\,Z_t + J_2\,Z_t^2 - J_2.$ (Gaussian(0,1))
- **Model IIB:** $X_t = J_1\,H_1(Z_t) + J_2\,H_2(Z_t) = J_1\,Z_t + J_2\,Z_t^2 - J_2.$ (Exp(1)-1)
- **Model IIIA:** $X_t = \sum_{j\geq 0}\beta^j\,\prod_{n=0}^{j-1} e_{t-nk-\ell}\,e_{t-jk}$ (Exp(1)-1)
- **Model IIIB:** $X_t = \sum_{j\geq 0}\beta^j\,\prod_{n=0}^{j-1} e_{t-nk-\ell}\,e_{t-jk}$ (Gaussian(0,1))

Three subcases of each of these models are taken using different choices of parameters.

Table 2.1: MSE Comparison for $T = 100$ and $P = 20$. The values in the parentheses represent relative percentage improvements in MSE when Quadratic Prediction is used compared to the respective competing methods.

| Model | Parameters | Quad Pred | Fan &Yao | Linear |
|---|---|---|---|---|
| IA | $A = -0.235,\ B = 0.376$ | 0.15 | 0.22 (31.81) | 0.32 (53.12) |
| | $A = -0.350,\ B = 0.100$ | 0.14 | 0.18 (22.22) | 0.24 (41.67) |
| | $A = 0.350,\ B = -0.100$ | 0.18 | 0.21 (14.29) | 0.28 (35.71) |
| IB | $A = -0.235,\ B = 0.376$ | 0.59 | 1.28 (53.91) | 1.12 (47.32) |
| | $A = -0.350,\ B = 0.100$ | 0.20 | 0.32 (37.50) | 0.42 (52.38) |
| | $A = 0.350,\ B = -0.100$ | 0.27 | 0.16 (−68.75) | 0.27 (1.09) |
| IIA | $\rho = 0.8,\ J_1 = 0.1,\ J_2 = 2$ | 0.77 | 1.22 (36.89) | 1.32 (41.67) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 0.5$ | 0.43 | 0.85 (49.41) | 0.73 (41.09) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 10$ | 1.14 | 15.80 (92.78) | 1.58 (27.84) |
| IIB | $\rho = 0.8,\ J_1 = 0.1,\ J_2 = 2$ | 2.50 | 8.65 (71.09) | 7.19 (65.23) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 0.5$ | 1.04 | 3.97 (73.81) | 1.38 (24.63) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 10$ | 3.42 | 14.86 (76.98) | 7.98 (57.14) |
| IIIA | $k = 1,\ l = 2,\ \beta = 0.3$ | 1.05 | 3.25 (67.69) | 3.89 (73.01) |
| | $k = 2,\ l = 5,\ \beta = 0.3$ | 1.15 | 1.39 (17.27) | 3.14 (63.37) |
| | $k = 5,\ l = 2,\ \beta = 0.3$ | 0.83 | 1.56 (46.79) | 2.47 (66.39) |
| IIIB | $k = 1,\ l = 2,\ \beta = 0.3$ | 0.92 | 1.26 (26.98) | 2.02 (54.45) |
| | $k = 2,\ l = 5,\ \beta = 0.3$ | 0.52 | 1.06 (50.94) | 2.57 (79.76) |
| | $k = 5,\ l = 2,\ \beta = 0.3$ | 0.83 | 1.56 (46.79) | 2.47 (66.39) |

Table 2.1 gives MSE for 5 different models for linear and Quadratic prediction, and also MSE obtained by using a nonparametric approach proposed in Chapter 10 of [46]. The sample size is taken to be $T = 100$ and the past window length is taken to be $P = 20$. More

tables are given in the Supplement for different choices of sample size (T) and past window length (P). From the tables, we find significant improvement in almost in all of the cases over the competing approaches. For the nonparametric method of [46], the improvement obtained by the quadratic approach can be as high as 92% although there are cases where the nonparametric approach is superior (cf. Model IB, subcase 3). On the other hand, the quadratic prediction always provided improvements over the linear prediction, with the amount of improvement ranging from modest (e.g., 1.09% for Model IB, case 3) to substantial (79.76% for Model IIIB, case 2). Hence, while quadratic prediction may not always provide substantially better results than linear prediction (say, when the process is Gaussian), it is always at least as good as and performs better than its linear counterpart under the presence of nonlinearity. Of course, when the sample size is not reasonably large, the improvement in MSE (2.3.4) of using the BQP over the BQL could be offset by additional error due to parameter estimation uncertainty, in view of the fact that quadratic filters require the estimation of more filter coefficients.

## 2.7 Data Analysis

We consider two applications. The first example treats the case of forecasting the Wolfer sunspots, and the second example examines nonlinear forecasts of unemployment data.

### 2.7.1 Wolfer Sunspots

The time series of Wolfer sunspots has been heavily studied. We consider a monthly vintage starting in 1749. Examination of the series shows large cyclical movements due to the known solar behavior, and the oscillations have an asymmetric shape. As the period is roughly 11 years, or 132 time units, longer-term forecasts should use auto-cumulants containing this many lags. We instead examine the one-step ahead forecast, which should not be greatly impacted by the solar oscillation. With $P = 30$ as the size of the predictor set used in the quadratic prediction problem, and with the whole sample used to estimate the auto-moments nonparametrically, we determine the one-step ahead prediction MSE resulting from both the linear and quadratic estimators: there is a **29.2 % reduction** in MSE.

### 2.7.2 Unemployment Rate

We also examined unemployment rate data from the Bureau of Labor Statistics. This series is the monthly Seasonally Adjusted Unemployment Rate (16 years and over, series id LNS14000000), covering the period January 1948 through July 2019, of the Labor Force Statistics from the Current Population Survey. This was downloaded from the Bureau of Labor Statistics on 4:30 PM, August 8, 2019 (Data Source). The series is of considerable interest to economists and policy-makers, and is fairly smooth with occasional bursts of activity. A crude autoregressive fit indicates an AR(13) may be adequate from the standpoint of linear time series modeling, and hence we will use $P = 13$ as the size of the predictor set (though with the entire sample used to nonparametrically estimate the auto-moments). As a result, the one-step ahead prediction MSE is **16.5 % reduced** for the quadratic estimator as compared to the linear.

## 2.8 Proofs

### 2.8.1 Proof of Proposition 2.2.1

Write $\mathcal{X}' = [\underline{X}', \text{vech}[\underline{X}\,\underline{X}']']$ for the complete data vector, and set $\beta' = [\underline{b}', \text{vech}[B]']$. Then, we can express a generic predictor $g(\underline{X}) = \underline{b}'\underline{X} + \underline{X}'B\underline{X} - \mathbb{E}\underline{X}'B\underline{X}$ as $g(\underline{X}) = \beta'\{\mathcal{X} - \mathbb{E}[\mathcal{X}]\}$. Next recall that $\Sigma_{A,B} = \text{Cov}[A,B]$ for random vectors $A$ and $B$. Hence, the quadratic MSE can be expressed as

$$\mathcal{Q}(\beta) \equiv \mathbb{E}[(Y - g(\underline{X}))^2] = \text{Var}[Y] - 2\,[\Sigma_{Y,\underline{X}},\, \Sigma_{Y,\underline{W}}]\,\beta + \beta'\,M\,\beta,$$

where

$$M = \begin{bmatrix} \Sigma_{\underline{X},\underline{X}} & \Sigma_{\underline{X},\underline{W}} \\ \Sigma_{\underline{W},\underline{X}} & \Sigma_{\underline{W},\underline{W}} \end{bmatrix}.$$

Now setting $\frac{\partial \mathcal{Q}(\beta)}{\partial \beta} = 0$, we get the matrix equivalent of the generalized Yule-Walker equations in (2.2.2):

$$M\beta = [\Sigma_{Y,\underline{X}},\, \Sigma_{Y,\underline{W}}].$$

Note that when both $\Sigma_{\underline{X},\underline{X}}$ and the Schur complement $S$ are invertible, the matrix $M$ is invertible. The proposition now follows by expressing $M^{-1}$ in terms of $\Sigma^{-1}_{\underline{X},\underline{X}}$ and $S^{-1}$ and simplifying the resulting expression. □

### 2.8.2 Proof of Proposition 2.4.1

Without loss of generality suppose that $\mathbb{E}[X_t] = 0$. Take $j, k \geq 0$ as fixed integers throughout the proof. First we note that the linear estimator $\widehat{X_{t-j}X_{t-k}}$ has to take the form (2.4.5) by basic linear projection theory ([24]), given that $\mathbb{E}[X_{t-j}X_{t-k}] = \gamma(k-j)$. The error process for the optimal linear estimator needs to be uncorrelated with $X_{t-\ell}$ for all $\ell \geq 0$, which yields the equations

$$\begin{aligned}
\gamma_3(\ell - j, \ell - k) = \sum_{h \geq 0} \pi_h^{(j,k)} \gamma(\ell - h), \quad \text{which holds if and only if} \\
\left( \left( z^{j-\ell} y^{k-\ell} f_3(z,y) \right)_y \right)_z = \left( \Pi^{(j,k)}(z)\, z^{-\ell} f(z) \right)_z, \quad \text{which holds if and only if} \\
\left( z^{j-\ell} \left( y^{k-j} f_3(z\overline{y}, y) \right)_y \right)_z = \mu_2 \left( \Pi^{(j,k)}(z)\, z^{-\ell} \Psi_2(z) \Psi_2(\overline{z}) \right)_z .
\end{aligned}$$

The last line is obtained by using the change of variable $z \mapsto z\overline{y}$ (this amounts to shifting the frequency $\lambda$ in $z = e^{-i\lambda}$, so there is no impact from the chain rule), and applying the spectral factorization (2.4.3). Consider summing this last equation against $\beta_\ell$ for $\ell \geq 0$, where we define these coefficients for any desired $h \geq 0$ via $\beta_\ell = \widetilde{\psi}^{(2)}_{\ell - h}$ for $\ell \geq h$, and zero otherwise. Here $\widetilde{\psi}^{(2)}_k$ is the $k$th coefficient of $\Psi_2(z)^{-1}$. This definition means that $\sum_{\ell \geq 0} \beta_\ell z^\ell = \Psi_2(z)^{-1} z^h$, and we can provide this construction for any $h \geq 0$. The application of these coefficients yields a new system of equations, which hold for all $h \geq 0$:

$$\left( z^{j-h} \left( y^{k-j} f_3(z\overline{y}, y) \right)_y / \Psi_2(\overline{z}) \right)_z = \mu_2 \left( \Pi^{(j,k)}(z)\, z^{-h} \Psi_2(z) \right)_z .$$

Note that $\Pi^{(j,k)}(z)\, \Psi_2(z)$ is a power series (i.e., it corresponds to some causal filter), and hence the right hand side of the above equation is just $\mu_2$ times the $h^{th}$ coefficient of this power series. It follows that there can be no anti-causal portion of the left-hand side of the equation, i.e.,

$$\begin{aligned}
\left[ z^j \left( y^{k-j} f_3(zy^{-1}, y) \right)_y / \Psi_2(z^{-1}) \right]_{-\infty}^{-1} = 0, \quad \text{and} \\
\left[ z^j \left( y^{k-j} f_3(zy^{-1}, y) \right)_y / \Psi_2(z^{-1}) \right]_{0}^{\infty} = \mu_2 \, \Pi^{(j,k)}(z) \Psi_2(z).
\end{aligned}$$

Note that $\mu_2 \neq 0$, so dividing by $\mu_2 \Psi_2(z)$ yields the stated formula. □

### 2.8.3 Proof of Theorem 2.4.1

The theorem claims that (2.4.4) holds (for all $j, k \geq 0$) if and only if (2.4.6) equals zero (for all $j, k \geq 0$). Fixing arbitrary $j, k \geq 0$, (2.4.4) holds if and only if

$$\gamma_3(-j-1, -k-1) = \sum_{h \geq 0} \pi_h^{(j,k)} \, \gamma(-1-h).$$

Utilizing the result of Proposition 2.4.1, (2.4.4) holds if and only if

$$\begin{aligned}
\left(\left(z^{j+1} y^{k+1} \, f_3(z,y)\right)_y\right)_z &= \left(\Pi^{(j,k)}(z) \, z f(z)\right) \\
&= \left(z\Psi_2(\overline{z})\left[z^j \left(y^{k-j} \, f_3(z\overline{y}, y)\right)_y / \Psi_2(\overline{z})\right]_0^\infty\right) \\
&= \left(z\Psi_2(\overline{z}) z^j \left(y^{k-j} \, f_3(z\overline{y}, y)\right)_y / \Psi_2(\overline{z})\right) \\
&\quad - \left(z\Psi_2(\overline{z})\left[z^j \left(y^{k-j} \, f_3(z\overline{y}, y)\right)_y / \Psi_2(\overline{z})\right]_{-\infty}^{-1}\right).
\end{aligned}$$

To obtain the last equality, we have used the fact that a Laurent series $\Theta(z)$ can be written as $[\Theta(z)]_0^\infty = \Theta(z) - [\Theta(z)]_{-\infty}^{-1}$. Next, note that

$$\left(z\Psi_2(\overline{z}) z^j \left(y^{k-j} \, f_3(z\overline{y}, y)\right)_y / \Psi_2(\overline{z})\right)_z = \left(\left(\left(z\overline{y}^{j+1} y^{k+1} \, f_3(z\overline{y}, y)\right)\right)_y\right)_z.$$

Therefore, (2.4.4) holds if and only if

$$\begin{aligned}
0 &= \left(z\Psi_2(\overline{z})\left[z^j \left(y^{k-j} \, f_3(z\overline{y}, y)\right)_y / \Psi_2(\overline{z})\right]_{-\infty}^{-1}\right) \\
&= \left(\Psi_2(\overline{z})\left[z^{j+1} \left(y^{k-j} \, f_3(z\overline{y}, y)\right)_y / \Psi_2(\overline{z})\right]_{-\infty}^{0}\right),
\end{aligned}$$

which uses the fact that for any Laurent series $\Theta(z)$, $z[\Theta(z)]_{-\infty}^{-1} = [z\Theta(z)]_{-\infty}^{0}$. The final expression is the integral of the product of two power series, and hence the product of their index-zero coefficients must be zero; because $\Psi_2(0) = 1$, (2.4.4) holds if and only if

$$0 = \left( z^{j+1} \left( y^{k-j} \, f_3(z\overline{y}, y) \right)_y / \Psi_2(\overline{z}) \right)_z .$$

Now applying the transformation $z \mapsto zy$, we obtain (2.4.6) equals zero for all $j, k \geq 0$. □

## 2.9 Computing Auto-cumulants of Hermite Processes

Define the generating function $h(x,t) = \exp\{xt - t^2/2\}$, which is related to the Hermite polynomials via

$$h(x,t) = \sum_{k=0}^{\infty} \frac{t^k}{\sqrt{k!}} H_k(x).$$

It is known that the autocovariance function is given by

$$\gamma(h) = \sum_{\ell \geq 0} J_\ell^2 \, c(h)^\ell,$$

and we generalize this below. It can be shown that the formula for the order $(r+1)$ auto-moment is

$$\gamma_{r+1}(\underline{h}) = \sum_{\underline{\ell} \geq 0} J_{\ell_0} \cdot J_{\ell_1} \cdots J_{\ell_r} \prod_{j=0}^{r} (\ell_j!)^{-1/2} \\ \times \frac{\partial^{\ell_j}}{\partial s_j^{\ell_j}} \mathbb{E}\left[ \exp\left\{ \sum_{i=0}^{r} s_i Z_{t+h_i} - \sum_{i=0}^{r} s_i^2/2 \right\} \right] |_{s_j=0},$$

where $h_0 = 0$, and $\underline{\ell} = [\ell_0, \ell_1, \ldots, \ell_r]'$. The expectation can be expanded as follows: let $K = \{(m,n) : 0 \leq m, n \leq r, m \neq n\}$. Then it follows from the formula for the mean of the lognormal distribution that

$$\mathbb{E}\left[ \exp\left\{ \sum_{i=0}^{r} s_i Z_{t+h_i} - \sum_{i=0}^{r} s_i^2/2 \right\} \right] = \exp\left\{ \sum_{(m,n) \in K} s_m s_n \, c(h_m - h_n) \right\},$$

and by differentiation the auto-moments can be determined. We now consider the cases of the third and fourth auto-moments. For $r = 2$ we compute

$$\begin{aligned}&\frac{\partial^{\ell_0}}{\partial s_0^{\ell_0}}\frac{\partial^{\ell_1}}{\partial s_1^{\ell_1}}\frac{\partial^{\ell_2}}{\partial s_2^{\ell_2}}\exp\{s_0 s_1\, c(h_1) + s_0 s_2\, c(h_2) + s_1 s_2\, c(h_1 - h_2)\}\Big|_{s_0=s_1=s_2=0}\\ =&\frac{\partial^{\ell_0}}{\partial s_0^{\ell_0}}\frac{\partial^{\ell_1}}{\partial s_1^{\ell_1}}\frac{\partial^{\ell_2}}{\partial s_2^{\ell_2}}\sum_{n_0,n_1,n_2\geq 0}\Big[\frac{s_0^{n_0+n_1}\,s_1^{n_0+n_2}\,s_2^{n_1+n_2}}{n_0!n_1!n_2!}\\ &\qquad\qquad\times c(h_1)^{n_0}\,c(h_2)^{n_1}\,c(h_1-h_2)^{n_2}\Big]\Big|_{s_0=s_1=s_2=0},\end{aligned}$$

which is nonzero only if $\ell_0 = n_0 + n_1$, $\ell_1 = n_0 + n_2$, and $\ell_2 = n_1 + n_2$. The integer solution to this system is unique, and is given by

$$n_0 = (\ell_0 + \ell_1 - \ell_2)/2, \qquad n_1 = (\ell_0 - \ell_1 + \ell_2)/2, \qquad n_2 = (-\ell_0 + \ell_1 + \ell_2)/2, \tag{2.9.1}$$

so long as $\ell_0$, $\ell_1$, and $\ell_2$ are even integers. Therefore

$$\gamma_3(h_1,h_2) = \sum_{\ell_0,\ell_1,\ell_2\in 2\mathbb{N}^+} J_{\ell_0}\cdot J_{\ell_1}\cdot J_{\ell_2}\,\frac{\sqrt{\ell_0!\ell_1!\ell_2!}}{n_0!n_1!n_2!}\,c(h_1)^{n_0}\,c(h_2)^{n_1}\,c(h_1-h_2)^{n_2},$$

where $n_0$, $n_1$, and $n_2$ satisfy (2.9.1) and $\mathbb{N}^+ = \mathbb{N}\cup\{0\}$. For $r = 3$ we have

$$\begin{aligned}&\frac{\partial^{\ell_0}}{\partial s_0^{\ell_0}}\frac{\partial^{\ell_1}}{\partial s_1^{\ell_1}}\frac{\partial^{\ell_2}}{\partial s_2^{\ell_2}}\frac{\partial^{\ell_3}}{\partial s_3^{\ell_3}}\exp\{s_0 s_1\, c(h_1) + s_0 s_2\, c(h_2) + s_0 s_3\, c(h_3) +\\ &\qquad s_1 s_2\, c(h_1-h_2) + s_1 s_3\, c(h_1-h_3) + s_2 s_3\, c(h_2-h_3)\}\,|_{s_0=s_1=s_2=s_3=0}\\ =&\frac{\partial^{\ell_0}}{\partial s_0^{\ell_0}}\frac{\partial^{\ell_1}}{\partial s_1^{\ell_1}}\frac{\partial^{\ell_2}}{\partial s_2^{\ell_2}}\frac{\partial^{\ell_3}}{\partial s_3^{\ell_3}}\sum_{n_0,n_1,n_2,n_3,n_4,n_5\geq 0}\frac{s_0^{n_0+n_1+n_2}\,s_1^{n_0+n_3+n_4}\,s_2^{n_1+n_3+n_5}\,s_3^{n_2+n_4+n_5}}{n_0!n_1!n_2!n_3!n_4!n_5!}\\ &\qquad c(h_1)^{n_0}\,c(h_2)^{n_1}\,c(h_3)^{n_2}\,c(h_1-h_2)^{n_3}\,c(h_1-h_3)^{n_4}\,c(h_2-h_3)^{n_5}|_{s_0=s_1=s_2=s_3=0},\end{aligned}$$

which is nonzero only if $\ell_0 = n_0+n_1+n_2$, $\ell_1 = n_0+n_3+n_4$, $\ell_2 = n_1+n_3+n_5$, and $\ell_3 = n_2+n_4+n_5$. Given $\ell_0$, $\ell_1$, $\ell_2$, and $\ell_3$, finding solutions for $n_0,\ldots,n_5$ is an under-determined system. If we allow $n_4$ and $n_5$ to be free variables (non-negative integers), then

$$\begin{aligned}
n_0 &= (\ell_0 + \ell_1 - \ell_2 - \ell_3)/2 + n_5 \\
n_1 &= (\ell_0 - \ell_1 + \ell_2 - \ell_3)/2 + n_4 \\
n_2 &= \ell_3 - n_4 - n_5 \\
n_3 &= (-\ell_0 + \ell_1 + \ell_2 + \ell_3)/2 - n_4 - n_5,
\end{aligned}$$

and subject to these constraints

$$\begin{aligned}
&\gamma_4(h_1, h_2, h_3) \\
&= \sum_{\ell_0,\ell_1,\ell_2,\ell_3 \in 2\mathbb{N}^+} J_{\ell_0} \cdot J_{\ell_1} \cdot J_{\ell_2} \cdot J_{\ell_3} \prod_{j=0}^{3} (\ell_j!)^{-1/2} \sum_{n_4,n_5 \geq 0} \frac{\ell_0!\ell_1!\ell_2!\ell_3!}{n_0!n_1!n_2!n_3!n_4!n_5!} \\
&\qquad c(h_1)^{n_0}\, c(h_2)^{n_1}\, c(h_3)^{n_2}\, c(h_1 - h_2)^{n_3}\, c(h_1 - h_3)^{n_4}\, c(h_2 - h_3)^{n_5}.
\end{aligned}$$

Table 2.2: Configurations of $\ell$ and $n$ indices. Left column gives values for $\ell_0$, $\ell_1$, $\ell_2$, and $\ell_3$, and right column gives corresponding possible values for $n_0$, $n_1$, $n_2$, $n_3$, $n_4$, and $n_5$.

| $\ell$ | $n$ |
|---|---|
| 1 1 1 1 | [0 0 1 1 0 0], [1 0 0 0 0 1], [0 1 0 0 1 0] |
| 1 1 2 2 | [0 0 1 1 0 1], [0 1 0 0 1 1], [1 0 0 0 0 2] |
| 1 2 1 2 | [0 0 1 1 1 0], [1 0 0 0 1 1], [0 1 0 0 2 0] |
| 1 2 2 1 | [0 0 1 2 0 0], [1 0 0 1 0 1], [0 1 0 1 1 0] |
| 2 1 1 2 | [0 0 2 1 0 0], [1 0 1 0 0 1], [0 1 1 0 1 0] |
| 2 1 2 1 | [0 1 1 1 0 0], [1 1 0 0 0 1], [0 2 0 0 1 0] |
| 2 2 1 1 | [1 0 1 1 0 0], [2 0 0 0 0 1], [1 1 0 0 1 0] |
| 2 2 2 2 | [0 0 2 2 0 0], [1 0 1 1 0 1], [2 0 0 0 0 2],<br>[1 1 0 0 1 1], [0 1 1 1 1 0], [0 2 0 0 2 0] |

We can now apply these formulas to the case of a quadratic Hermite process. For $r = 2$, the only nonzero terms occur when $\ell_0 = \ell_1 = \ell_2 = 2$, which implies $n_0 = n_1 = n_2 = 1$, and we obtain the stated formula. For $r = 3$, there are sixteen configurations for the sum over the $\ell$ indices, as each can take the value one or two. It turns out that eight of these configurations can yield solutions in terms of $n_0, \ldots, n_5$, which in turn are each constrained to be zero or one. These configurations are described in Table 2.2. By carefully organizing the results, we obtain the stated expression for the fourth-order auto-moment.

## 2.10 Supplemental Tables

Table 2.3: MSE Comparison for $T = 100$ and $P = 20$. The values in the parenthesis represent the relative percentage improvement in MSE when Quadratic Prediction is used compared to the corresponding column.

| Model | Parameters | Quad Pred | Fan &Yao | Linear |
|---|---|---|---|---|
| IA | $A = -0.235,\ B = 0.376$ | 0.15 | 0.22 (31.81) | 0.32 (53.12) |
| | $A = -0.350,\ B = 0.100$ | 0.14 | 0.18 (22.22) | 0.24 (41.67) |
| | $A = 0.350,\ B = -0.100$ | 0.18 | 0.21 (14.29) | 0.28 (35.71) |
| IB | $A = -0.235,\ B = 0.376$ | 0.59 | 1.28 (53.91) | 1.12 (47.32) |
| | $A = -0.350,\ B = 0.100$ | 0.20 | 0.32 (37.50) | 0.42 (52.38) |
| | $A = 0.350,\ B = -0.100$ | 0.27 | 0.16 (−68.75) | 0.27 (1.09) |
| IIA | $\rho = 0.8,\ J_1 = 0.1,\ J_2 = 2$ | 0.77 | 1.22 (36.89) | 1.32 (41.67) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 0.5$ | 0.43 | 0.85 (49.41) | 0.73 (41.09) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 10$ | 1.14 | 15.80 (92.78) | 1.58 (27.84) |
| IIB | $\rho = 0.8,\ J_1 = 0.1,\ J_2 = 2$ | 2.50 | 8.65 (71.09) | 7.19 (65.23) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 0.5$ | 1.04 | 3.97 (73.81) | 1.38 (24.63) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 10$ | 3.42 | 14.86 (76.98) | 7.98 (57.14) |
| IIIA | $k = 1,\ l = 2,\ \beta = 0.3$ | 1.05 | 3.25 (67.69) | 3.89 (73.01) |
| | $k = 2,\ l = 5,\ \beta = 0.3$ | 1.15 | 1.39 (17.27) | 3.14 (63.37) |
| | $k = 5,\ l = 2,\ \beta = 0.3$ | 0.83 | 1.56 (46.79) | 2.47 (66.39) |
| IIIB | $k = 1,\ l = 2,\ \beta = 0.3$ | 0.92 | 1.26 (26.98) | 2.02 (54.45) |
| | $k = 2,\ l = 5,\ \beta = 0.3$ | 0.52 | 1.06 (50.94) | 2.57 (79.76) |
| | $k = 5,\ l = 2,\ \beta = 0.3$ | 0.83 | 1.56 (46.79) | 2.47 (66.39) |

Table 2.4: MSE Comparison for $T$ = 100 and $P$ = 10. The values in the parenthesis represent the relative percentage improvement in MSE when Quadratic Prediction is used compared to the corresponding column.

| Model | Parameters | Quad Pred | Fan &Yao | Linear |
|---|---|---|---|---|
| IA | $A = -0.235,\ B = 0.376$ | 0.08 | 0.11 (27.27) | 0.18 (55.55) |
|  | $A = -0.350,\ B = 0.100$ | 0.12 | 0.13 (7.69) | 0.23 (47.82) |
|  | $A = 0.350,\ B = -0.100$ | 0.12 | 0.18 (33.33) | 0.25 (52.01) |
| IB | $A = -0.235,\ B = 0.376$ | 0.32 | 2.55 (87.45) | 4.17 (92.32) |
|  | $A = -0.350,\ B = 0.100$ | 0.08 | 0.15 (46.67) | 0.31 (74.19) |
|  | $A = 0.350,\ B = -0.100$ | 0.11 | 0.19 (42.11) | 0.44 (75.01) |
| IIA | $\rho = 0.8,\ J_1 = 0.1,\ J_2 = 2$ | 2.23 | 11.38 (80.41) | 7.14 (68.76) |
|  | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 0.5$ | 0.30 | 0.61 (50.81) | 0.83 (63.85) |
|  | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 10$ | 7.91 | 38.84 (79.63) | 18.39 (56.98) |
| IIB | $\rho = 0.8,\ J_1 = 0.1,\ J_2 = 2$ | 2.50 | 8.65 (71.09) | 7.19 (65.23) |
|  | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 0.5$ | 1.04 | 3.97 (73.80) | 1.38 (24.63) |
|  | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 10$ | 4.90 | 56.26 (91.29) | 21.87 (77.59) |
| IIIA | $k = 1,\ l = 2,\ \beta = 0.3$ | 0.93 | 1.30 (28.46) | 1.83 (49.18) |
|  | $k = 2,\ l = 5,\ \beta = 0.3$ | 0.97 | 0.75 (−29.33) | 2.41 (59.75) |
|  | $k = 5,\ l = 2,\ \beta = 0.3$ | 0.73 | 1.71 (57.31) | 2.65 (72.45) |
| IIIB | $k = 1,\ l = 2,\ \beta = 0.3$ | 1.26 | 1.41 (10.64) | 2.71 (53.51) |
|  | $k = 2,\ l = 5,\ \beta = 0.3$ | 0.52 | 1.06 (50.94) | 2.57 (79.76) |
|  | $k = 5,\ l = 2,\ \beta = 0.3$ | 0.90 | 1.48 (39.18) | 2.24 (59.82) |

Table 2.5: MSE Comparison for $T = 500$ and $P = 10$. The values in the parenthesis represent the relative percentage improvement in MSE when Quadratic Prediction is used compared to the corresponding column.

| Model | Parameters | Quad Pred | Fan &Yao | Linear |
|---|---|---|---|---|
| IA | $A = -0.235,\ B = 0.376$ | 0.03 | 0.13 (76.92) | 0.14 (78.57) |
| | $A = -0.350,\ B = 0.100$ | 0.19 | 0.23 (17.39) | 0.33 (42.42) |
| | $A = 0.350,\ B = -0.100$ | 0.11 | 0.28 (60.71) | 0.35 (68.57) |
| IB | $A = -0.235,\ B = 0.376$ | 0.12 | 1.55 (92.25) | 4.23 (97.16) |
| | $A = -0.350,\ B = 0.100$ | 0.03 | 0.25 (46.67) | 0.41 (74.19) |
| | $A = 0.350,\ B = -0.100$ | 0.11 | 0.19 (88.00) | 0.44 (92.68) |
| IIA | $\rho = 0.8,\ J_1 = 0.1,\ J_2 = 2$ | 1.23 | 8.17 (84.94) | 6.31 (80.51) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 0.5$ | 0.23 | 0.41 (43.91) | 0.53 (56.61) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 10$ | 4.33 | 28.32 (84.71) | 16.91 (74.39) |
| IIB | $\rho = 0.8,\ J_1 = 0.1,\ J_2 = 2$ | 3.45 | 9.16 (62.33) | 8.49 (59.36) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 0.5$ | 2.04 | 4.67 (56.31) | 7.23 (71.78) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 10$ | 3.90 | 36.17 (89.21) | 11.89 (67.19) |
| IIIA | $k = 1,\ l = 2,\ \beta = 0.3$ | 1.42 | 2.31 (38.52) | 2.43 (41.56) |
| | $k = 2,\ l = 5,\ \beta = 0.3$ | 0.67 | 0.55 (−21.81) | 3.31 (79.75) |
| | $k = 5,\ l = 2,\ \beta = 0.3$ | 0.92 | 1.31 (29.01) | 3.45 (73.45) |
| IIIB | $k = 1,\ l = 2,\ \beta = 0.3$ | 1.34 | 1.93 (30.56) | 3.71 (63.88) |
| | $k = 2,\ l = 5,\ \beta = 0.3$ | 1.23 | 1.97 (37.56) | 2.31 (46.75) |
| | $k = 5,\ l = 2,\ \beta = 0.3$ | 1.31 | 2.03 (35.47) | 2.91 (54.98) |

Table 2.6: MSE Comparison for $T$ = 500 and $P$ = 20. The values in the parenthesis represent the relative percentage improvement in MSE when Quadratic Prediction is used compared to the corresponding column.

| Model | Parameters | Quad Pred | Fan &Yao | Linear |
|---|---|---|---|---|
| IA | $A = -0.235,\ B = 0.376$ | 0.16 | 0.20 (20.03) | 0.37 (56.76) |
| | $A = -0.350,\ B = 0.100$ | 0.12 | 0.16 (25.02) | 0.28 (57.14) |
| | $A = 0.350,\ B = -0.100$ | 0.15 | 0.12 (−25.01) | 0.28 (46.43) |
| IB | $A = -0.235,\ B = 0.376$ | 0.64 | 1.62 (60.49) | 2.31 (72.29) |
| | $A = -0.350,\ B = 0.100$ | 0.18 | 0.27 (33.33) | 0.58 (68.96) |
| | $A = 0.350,\ B = -0.100$ | 0.16 | 0.24 (33.33) | 0.49 (67.35) |
| IIA | $\rho = 0.8,\ J_1 = 0.1,\ J_2 = 2$ | 5.16 | 8.11 (36.37) | 9.09 (43.23) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 0.5$ | 0.42 | 0.53 (20.75) | 0.56 (25.02) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 10$ | 8.69 | 37.38 (76.76) | 14.79 (41.24) |
| IIB | $\rho = 0.8,\ J_1 = 0.1,\ J_2 = 2$ | 4.53 | 11.16 (59.41) | 13.43 (66.27) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 0.5$ | 0.37 | 0.54 (31.48) | 0.62 (40.32) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 10$ | 12.84 | 88.46 (85.49) | 34.93 (63.24) |
| IIIA | $k = 1,\ l = 2,\ \beta = 0.3$ | 0.31 | 0.91 (65.94) | 1.38 (77.54) |
| | $k = 2,\ l = 5,\ \beta = 0.3$ | 0.59 | 1.12 (47.32) | 1.70 (65.29) |
| | $k = 5,\ l = 2,\ \beta = 0.3$ | 0.67 | 1.45 (53.79) | 1.01 (33.67) |
| IIIB | $k = 1,\ l = 2,\ \beta = 0.3$ | 0.19 | 0.53 (64.15) | 1.41 (86.52) |
| | $k = 2,\ l = 5,\ \beta = 0.3$ | 0.94 | 2.06 (54.37) | 4.67 (79.87) |
| | $k = 5,\ l = 2,\ \beta = 0.3$ | 0.96 | 2.35 (59.15) | 3.74 (74.33) |

Table 2.7: MSE Comparison for $T = 50$ and $P = 10$. The values in the parenthesis represent the relative percentage improvement in MSE when Quadratic Prediction is used compared to the corresponding column.

| Model | Parameters | Quad Pred | Fan &Yao | Linear |
|---|---|---|---|---|
| IA | $A = -0.235,\ B = 0.376$ | 0.20 | 0.23 (13.04) | 0.39 (48.72) |
| | $A = -0.350,\ B = 0.100$ | 0.23 | 0.22 (−4.54) | 0.27 (14.82) |
| | $A = 0.350,\ B = -0.100$ | 0.29 | 0.21 (−38.09) | 0.30 (3.33) |
| IB | $A = -0.235,\ B = 0.376$ | 0.18 | 6.94 (97.40) | 5.08 (96.45) |
| | $A = -0.350,\ B = 0.100$ | 0.24 | 0.27 (11.11) | 0.35 (31.43) |
| | $A = 0.350,\ B = -0.100$ | 0.20 | 0.25 (20.03) | 0.56 (64.28) |
| IIA | $\rho = 0.8,\ J_1 = 0.1,\ J_2 = 2$ | 3.71 | 7.68 (51.69) | 7.22 (48.61) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 0.5$ | 0.46 | 0.61 (24.59) | 0.62 (25.81) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 10$ | 10.23 | 47.11 (78.29) | 26.36 (61.19) |
| IIB | $\rho = 0.8,\ J_1 = 0.1,\ J_2 = 2$ | 4.32 | 25.85 (83.29) | 12.66 (65.88) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 0.5$ | 0.58 | 2.53 (77.07) | 1.16 (50.05) |
| | $\rho = 0.8,\ J_1 = 0.5,\ J_2 = 10$ | 9.27 | 64.19 (85.58) | 38.52 (75.93) |
| IIIA | $k = 1,\ l = 2,\ \beta = 0.3$ | 1.59 | 1.58 (−0.63) | 1.94 (18.04) |
| | $k = 2,\ l = 5,\ \beta = 0.3$ | 1.19 | 1.93 (38.34) | 1.45 (17.93) |
| | $k = 5,\ l = 2,\ \beta = 0.3$ | 0.94 | 1.31 (28.24) | 1.51 (37.75) |
| IIIB | $k = 1,\ l = 2,\ \beta = 0.3$ | 1.39 | 1.63 (14.73) | 2.06 (35.25) |
| | $k = 2,\ l = 5,\ \beta = 0.3$ | 1.64 | 1.66 (1.21) | 2.09 (21.53) |
| | $k = 5,\ l = 2,\ \beta = 0.3$ | 1.57 | 2.19 (28.32) | 1.58 (0.63) |

# Chapter 3

# Polyspectral Mean Estimation of General Nonlinear Processes

## 3.1 Introduction

Spectral Analysis is an important tool in time series analysis. The spectral density provides important insights into the underlying periodicities of a time series; see [24], [90]. However, when the stochastic process is nonlinear, higher-order auto-cumulants [18] can furnish valuable insights that are not discernible from the spectral density, as argued in [20]. In particular, three- and four-way interactions of a nonlinear process can be measured through the third and fourth order polyspectra; see [21], [20], [23], [51], [85], [91], and [12]. Linear functionals of higher-order spectra (here referred to as polyspectral means) are often encountered in the analysis of nonlinear time series and are therefore a serious object of inference. Whereas there is an existent literature on estimation of polyspectra ([20], [137], [103], [121]), inference for polyspectral means remains a gap. This chapter aims to fill this gap by developing asymptotic distributional results on polyspectral means, and by developing statistical inference methodology for nonlinear times series analysis.

Let $\{X_t\}$ be a $(k+1)^{th}$ order stationary time series, that is, $E[|X_t|^{k+1}] < \infty$ and $EX_tX_{t+h_1}\dots X_{t+h_r} = EX_1X_{1+h_1}\dots X_{1+h_r}$ for all integers $t, h_1, \dots h_r$ and for all $1 \leq r \leq k$, for given $k \geq 1$. The most commonly used case is $k = 1$, yielding the class of second order stationary (SOS) processes. Thus, for a SOS process $\{X_t\}$, $EX_t = EX_1$ is a constant and its cross-covariances depend only on the time-difference $\text{Cov}(X_t, X_{t+h}) = \gamma(h)$ for all $t, h \in \mathbb{Z}$, where $\mathbb{Z} = \{0, \pm1, \pm2, \dots\}$ denotes the set of all integers. When the autocovariance function $\gamma(h) \equiv \gamma_h$ is absolutely summable, $\{X_t\}$ has a spectral density, given by $f(\lambda) = \sum_{h\in\mathbb{Z}} \gamma(h)e^{-\iota h\lambda}$, $\lambda \in [-\pi, \pi]$, where $\iota = \sqrt{-1}$. Important features of a stochastic process can be extracted with a spectral mean, which is defined as $\int_{[-\pi,\pi]} f(\lambda)g(\lambda)d\lambda$, where $g(\cdot)$ is a weight function. Spectral means can provide us important insight about the time series, based on different g functions. For example, a spectral mean with $g(\lambda) = cos(h\lambda)$ corresponds to the lag $h$ autocovariance, whereas $g(\lambda) = \mathbb{1}(-a, a)$ gives the spectral content in the interval $(-a, a)$, where $\mathbb{1}(\cdot)$ denotes the indicator function.

Analogous definitions can be made for higher order spectra. The $(k+1)^{th}$ order autocumulant is defined as $\gamma(h_1, \dots, h_k) = Cum(X_t, X_{t+h_1}, \dots, X_{t+h_k})$ for all $t$, where $Cum(X, Y, \dots, Z)$ denotes the cumulant of jointly distributed random variables $\{X, Y, \dots, Z\}$ (cf. [18]) and where $h_1, \dots, h_k$ are integer lags. If the auto-cumulant function is absolutely summable, then we can define the polyspectra of order $k$ as the Fourier transform of the $(k+1)^{th}$ order auto-cumulants:

$$f_k(\underline{\lambda}_k) = \sum_{\underline{h}\in\mathbb{Z}^k} \gamma(\underline{h})e^{-\iota \underline{h}'\underline{\lambda}},$$

where $\underline{h} = [h_1, \ldots, h_k]'$, $\underline{\lambda} = [\lambda_1, \ldots, \lambda_k]'$ and $A'$ denotes the transpose of a matrix $A$. A polyspectral mean with weight function $g : [-\pi, \pi]^k \to \mathbb{R}$ is defined as

$$M_g(f_k) = \int_{[-\pi,\pi]^k} f_k(\underline{\lambda}) g(\underline{\lambda}) d\underline{\lambda}. \tag{3.1.1}$$

Thus, spectral means correspond to the case $k = 1$, in which case we drop the subscript $k$ (and write $f_1 = f$). Polyspectral means give us important information about a time series that can not be obtained from the spectral distribution, e.g., when the time series is nonlinear or the innovation process is non-Gaussian. Additionally, we can extract several features from a time series by obtaining the polyspectral mean from different weight functions. These features can play an important role in identifying (dis-)similarities among different time series (cf. [31], [99], [44], [139]). For an illustrative example, we consider the Gross Domestic Product (GDP) data from 140 countries over 40 years. Our goal is to obtain clustering of the countries based on patterns of their GDP growth rates. The left panel of Figure 3.6 below gives the raw time series and the right panel gives their differenced and scaled versions. As a part of exploratory data analysis, we carried out test checks for Gaussianity and for linearity. Based on the tests, a significant proportion of the time series are non-Gaussian, and some of them are nonlinear. Hence, using spectral means alone to capture relevant time series features may be inadequate. Here we use (estimated) higher order polyspectral means with different weight functions to elicit salient features of the GDP data, in order to capture possible nonlinear dynamics of the economies. See Section 3.6 for more details.

Estimation of spectral and polyspectral means can be carried out using the periodogram and its higher order versions, which are defined in terms of the discrete Fourier transform of a sample $\{X_1, \ldots, X_T\}$ (described in Section 3.2). However, distributional properties of the polyspectral mean estimators are not very well-studied. In an important work, [38] proved the asymptotic normality of spectral means, i.e., the case $k = 1$; however, in the general case ($k > 1$), there does not seem to be any work on the asymptotic distribution of polyspectral mean estimators. One of the major contributions of this chapter is to establish asymptotic normality results for the polyspectral means of general order. We develop some nontrivial combinatorial arguments involving higher order auto-cumulants to show that, under mild conditions, the estimators of the $k$th order polyspectral mean parameter $M_g(f_k)$ in (3.1.1) are asymptotically normal. We also obtain an explicit expression for the asymptotic variance, which is shown to depend on certain polyspectral means of order $2k+1$. This result agrees with the known results on spectral means when specialized to the case $k = 1$, where the limiting variance is known to involve the trispectrum $f_{2k+1} = f_3$. In particular, the results of this chapter provide a unified way to construct studentized versions of spectral mean estimators and higher order polyspectral mean estimators, which can be used to carry out large sample statistical inference on polyspectral mean parameters of any *arbitrary* order $k \geq 1$.

The second major contribution of the chapter is the development of a new test procedure to assess the linearity assumption on a time series. Earlier work on the problem is typically based on the squared modulus of the estimated bispectrum ([106], [29], [11], [8]). In contrast, here we make a key observation that under the linearity hypothesis (i.e., the null hypothesis) the ratio of the bispectrum to a suitable power transformation of the spectral density must be a constant at *all* frequency pairs in $[-\pi, \pi]^2$. This observation allows us to construct a set of higher order polyspectral means that must be zero when the process is linear. On the other hand, for a nonlinear process, not all of these spectral means can be equal to zero. Here we exploit this fact and develop a new test statistic that can be used to test the significance of these polyspectral means. We also derive the asymptotic distribution of the test statistic under the null hypothesis, and provide the critical values needed for calibrating the proposed test.

The rest of the chapter is organized as follows. Section 3.2 covers some background material, and provides some examples of polyspectra and polyspectral means. Section 3.3 gives the regularity conditions and the asymptotic properties of the polyspectral means of a general order. The linearity test is described in Section 3.4. Simulations are presented in Section 3.5, and two data applications, one involving the Sunspot data and the other involving the GDP growth rate (with a discussion of clustering time series via bispectral means), are presented in Section 3.6. The proofs of the theoretical results are given in Section 3.7.

## 3.2 Background and examples

First we will define estimators of the polyspectral mean $M_g\left(f_k\right)$ of (3.1.1) corresponding to a given weight function $g$. Following [20], we define an estimator of $M_g\left(f_k\right)$ by replacing $f_k$ by the $k^{th}$ order periodogram.

Specifically, given a mean zero sample $\{X_1,\ldots,X_T\}$ from the time series $\{X_t\}$, the $k^{th}$ order periodogram is defined as

$$\hat{f}_k(\underline{\lambda}) = T^{-1} d(\lambda_1)\cdots d(\lambda_k) d(-[\underline{\lambda}])$$

where, with $\iota = \sqrt{-1}$), $d(\lambda) = \sum_{t=1}^{T} X_t e^{-\iota\lambda t}$ is the Discrete Fourier Transform (DFT) of $\{X_1,\ldots,X_T\}$ at $\lambda \in [-\pi,\pi]$, and where $[\underline{\lambda}]$ is a shorthand for $\sum_k \lambda_k$. We also define the Fourier frequencies as $\lambda_j = \frac{2\pi j}{T}$, where $j$ runs from $-\lfloor \frac{T}{2} \rfloor$ to $\lfloor \frac{T}{2} \rfloor$. Then letting $\sum_{\underline{\lambda}}$ denote a shorthand for a summation over $-\lfloor \frac{T}{2} \rfloor \leq j_1, \ldots, j_k \leq \lfloor \frac{T}{2} \rfloor$, we can define the estimator of the polyspectral mean $M_g(f_k)$ as

$$\begin{aligned}\widehat{M_g(f)} &\equiv (2\pi)^k T^{-k} \sum_{\underline{\lambda}} \hat{f}_k(\underline{\lambda}) g(\underline{\lambda}) \Phi(\underline{\lambda}) \\ &= (2\pi)^k T^{-k-1} \sum_{\underline{\lambda}} d(\lambda_1) \cdots d(\lambda_k) d(-[\underline{\lambda}]) g(\underline{\lambda}) \Phi(\underline{\lambda}),\end{aligned}$$

where $\Phi(\underline{\lambda})$ is an indicator function that is non-zero only when the $\lambda$'s do not lie in any sub-manifold, i.e., $\sum_j \lambda_{i_j} \not\equiv 0 \pmod{2\pi}$ for any subset $\lambda_{i_1}, \ldots, \lambda_{i_m}$ of $\underline{\lambda}$. We further assume that the weighing function $g$ is continuous in $\underline{\lambda}$ and it satisfies the symmetry condition

$$g(\underline{\lambda}) = \overline{g(-\underline{\lambda})} \tag{3.2.1}$$

for all $\underline{\lambda}$. *Moreover, assume that* $\int_{\underline{\lambda}} |g(\underline{\lambda})| < \infty$.

Note that for the polyspectral mean we can ignore the sub-manifolds, since they form a measure zero set; hence we can only use estimates when $\underline{\lambda}$ are not contained in such sub-manifolds. This is a significant issue, since in [21] the estimator of the polyspectral density is given by a kernel-weighted average of the $k^{th}$ order periodogram, where the average is taken by avoiding the sub-manifolds. Such an estimator introduces a bandwidth term, which in general makes the convergence rate much slower. Since we are working with polyspectral means, it is not necessary to smooth the $k^{th}$ order periodogram, and hence we can ignore the bandwidth problem and focus on regions that avoid the sub-manifolds.

In the following, we provide some examples of polyspectral means, which demonstrate how diverse features of a nonlinear process can be extracted with different weighting functions. The sample is considered to be mean-centered.

**Example** Let $g(\lambda_1, \lambda_2) = \frac{e^{\iota\lambda_1 h_1 + \iota\lambda_2 h_2}}{4\pi^2}$. Then it follows that:

$$M_g(f_2) = \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} f_2(\underline{\lambda}) g(\underline{\lambda}) d\underline{\lambda}$$
$$= \gamma(h_1, h_2),$$

where $\gamma(\cdot, \cdot)$ is the third order autocumulant function.

$$\widehat{M_g(f_2)} = T^{-3} \sum_{\lambda_1, \lambda_2 \neq 0} d(\lambda_1) d(\lambda_2) d(-\lambda_1 - \lambda_2) e^{\iota\lambda_1 h_1 + \iota\lambda_2 h_2}$$
$$\sim T^{-1} \sum_{t_3} X_{t_3} X_{t_3 + h_1} X_{t_3 + h_2} = \hat{\gamma}(h_1, h_2)$$

Hence the estimate of the polyspectral mean gives the sample autocumulant of third order.

**Example** Let $g(\lambda_1, \lambda_2) = \frac{cos(h_1\lambda_1) cos(h_2\lambda_2)}{4\pi^2}$ for $h_1 > h_2 > 0$. Then it follows that

$$\begin{aligned} M_g(f_2) &= \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} f_2(\underline{\lambda}) g(\underline{\lambda}) d\underline{\lambda} \\ &= (4\pi^2)^{-1} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \sum_{k_1, k_2 \in \mathbb{Z}} \gamma(k_1, k_2) e^{\iota\lambda_1 k_1 + \iota\lambda_2 k_2} cos(\lambda_1 h_1) cos(\lambda_2 h_2) d\lambda_1 d\lambda_2 \\ &= \frac{1}{4} \{\gamma(h_1, h_2) + \gamma(h_1, -h_2) + \gamma(-h_1, h_2) + \gamma(-h_1, -h_2)\} \end{aligned}$$

$$\begin{aligned}
\widehat{M_g(f_2)} &= (2\pi)^2 T^{-2} \sum_{\lambda_1,\lambda_2 \neq 0} T^{-1} d(\lambda_1) d(\lambda_2) d(-\lambda_1 - \lambda_2) g(\lambda_1, \lambda_2) \\
&= T^{-1} T^{-2} \sum_{\lambda_1,\lambda_2} \sum_{t_1,t_2,t_3} X_{t_1} X_{t_2} X_{t_3} e^{-\iota\lambda_1 t_1 - \iota\lambda_2 t_2 + \iota(\lambda_1+\lambda_2)t_3} cos(\lambda_1 h_1) cos(\lambda_2 h_2) d\lambda_1 d\lambda_2 \\
&- 2\pi T^{-1} T^{-2} \sum_{\lambda_1} \sum_{t_1.t_2,t_3} X_{t_1} X_{t_2} X_{t_3} e^{-\iota\lambda_1(t_1-t_3)} cos(\lambda_1 h_1) d\lambda_1 \\
&- 2\pi T^{-1} T^{-2} \sum_{\lambda_2} \sum_{t_1.t_2,t_3} X_{t_1} X_{t_2} X_{t_3} e^{-\iota\lambda_2(t_2-t_3)} cos(\lambda_2 h_2) d\lambda_2 \\
&= \frac{1}{4}\{\hat{\gamma}(h_1,h_2) + \hat{\gamma}(-h_1,h_2) + \hat{\gamma}(h_1,-h_2) + \hat{\gamma}(-h_1,-h_2)\}
\end{aligned}$$

where $\hat{\gamma}(\cdot,\cdot)$ denotes the sample autocumulant of third order. The last line is true since we considered the sample to be location centered. This result is similar to the spectral mean case, where $g(\lambda) = cos(h\lambda)/(2\pi)$ gives the lag h autocovariance.

**Example** Consider the trispectral mean with the weight function

$$g(\lambda_1, \lambda_2, \lambda_3) = e^{\iota(\lambda_1 h_1 + \lambda_2 h_2 + \lambda_3 h_3)}/(8\pi^3)$$

In this case:

$$\begin{aligned}
M_g(f_3) &= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f_3(\underline{\lambda}) g(\underline{\lambda}) d\underline{\lambda} \\
&= (8\pi^3)^{-1} \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \sum_{k_1,k_2,k_3\in\mathbb{Z}} \{\gamma(k_1,k_2,k_3) - \gamma(k_1)\gamma(k_3-k_2) - \gamma(k_2)\gamma(k_3-k_1) - \gamma(k_3)\gamma(k_2-k_1)\} \\
&\qquad e^{\iota\lambda_1(h_1-k_1)+\iota\lambda_2(h_2-k_2)+\lambda_3(h_3-k_3)} d\lambda_1 d\lambda_2 d\lambda_3 \\
&= \gamma(h_1,h_2,h_3) - \gamma(h_1)\gamma(h_3-h_2) - \gamma(h_2)\gamma(h_1-h_3) - \gamma(h_3)\gamma(h_2-h_1),
\end{aligned}$$

where $\gamma(\cdot,\cdot)$ is the third order autocumulant function.

Suppose $S = \{\lambda_1, \lambda_2, \lambda_3 \neq 0 : \lambda_1 + \lambda_2 \neq 0, \lambda_2 + \lambda_3 \neq 0, \lambda_1 + \lambda_3 \neq 0\}$. Then,

$$\begin{aligned}\widehat{M_g(f_3)} &= (2\pi)^3 T^{-3} \sum_{\underline{\lambda} \in S} T^{-1} d(\lambda_1) d(\lambda_2) d(\lambda_3) d(-\lambda_1 - \lambda_2 - \lambda_3) g(\lambda_1, \lambda_2, \lambda_3) \\ &= T^{-4} \sum_{\underline{\lambda} \in S} \sum_{t_1, t_2, t_3, t_4} X_{t_1} X_{t_2} X_{t_3} X_{t_4} e^{-\iota\lambda_1(t_1 - t_4 - h_1) - \iota\lambda_2(t_2 - t_4 - h_2) - \iota\lambda_3(t_3 - t_4 - h_3)}\end{aligned}$$

If $\lambda_i = 0$ for some $i$, as seen earlier we will get a $\bar{X}$ term which will be zero under the zero mean assumption. Consider the case for $\lambda_1 + \lambda_2 = 0$:

$$\begin{aligned}&T^{-4} \sum_{\lambda_1, \lambda_2} \sum_{t_1, t_2, t_3, t_4} X_{t_1} X_{t_2} X_{t_3} X_{t_4} e^{-\iota\lambda_1(t_1 - t_2 - h_1 + h_2)} e^{-\iota\lambda_2(t_3 - t_4 - h_3)} \\ &= \hat{\gamma}(h_1 - h_2) \hat{\gamma}(h_3)\end{aligned}$$

Hence, the estimate takes the form:

$$\widehat{M_g(f_3)} = \hat{\gamma}(h_1, h_2, h_3) - \hat{\gamma}(h_1)\hat{\gamma}(h_2 - h_3) - \hat{\gamma}(h_2)\hat{\gamma}(h_3 - h_1) - \hat{\gamma}(h_3)\hat{\gamma}(h_2 - h_1)$$

**Example** $g(\lambda_1, \lambda_2) = \mathbb{I}(\lambda_1 \in (-h_1, h_1), \lambda_2 \in (-h_2, h_2))$

$$\begin{aligned}\widehat{M_g(f)} &= T^{-2} \sum_{\lambda_1, \lambda_2} T^{-1} d(\lambda_1) d(\lambda_2) d(-\lambda_1 - \lambda_2) g(\lambda_1, \lambda_2) \\ &= (2\pi)^{-2} \int_{-h_1}^{h_1} \int_{-h_2}^{h_2} I_2(\lambda_1, \lambda_2) d\lambda_1 d\lambda_2\end{aligned}$$

which will give an estimate of the content of the bispectra in the rectangular region $(-h_1, h_1) \cup (-h_2, h_2)$. Here $I_2(\lambda_1, \lambda)$ is the $2^{nd}$ order periodogram.

**Example** $g(\lambda_1, \lambda_2) = \mathbb{I}(\lambda_1 \in (-h_1, h_1), \lambda_2 \in (-h_2, h_2))$

Here $M_g(f_2)$ gives the bispectral content in the rectangular region $(-h_1, h_1) \cup (-h_2, h_2)$.

$$\begin{aligned}\widehat{M_g(f)} &= T^{-2} \sum_{\lambda_1, \lambda_2} T^{-1} d(\lambda_1) d(\lambda_2) d(-\lambda_1 - \lambda_2) g(\lambda_1, \lambda_2) \\ &= (2\pi)^{-2} \int_{-h_1}^{h_1} \int_{-h_2}^{h_2} I_2(\lambda_1, \lambda_2) d\lambda_1 d\lambda_2\end{aligned}$$

which will give an estimate of the content of the bispectra in the rectangular region $(-h_1, h_1) \cup (-h_2, h_2)$. Here $I_2(\lambda_1, \lambda)$ is the $2^{nd}$ order periodogram.

**Example** $g(\lambda_1, \lambda_2) = (\pi - |\lambda_1|)(\pi - |\lambda_2|)$

$$\begin{aligned}\widehat{M_g(f)} &= (2\pi)^2 T^{-2} \sum_{\lambda_1, \lambda_2 \neq 0} T^{-1} d(\lambda_1) d(\lambda_2) d(-\lambda_1 - \lambda_2) g(\lambda_1, \lambda_2) \\ &\sim T^{-1} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \sum_{t_1, t_2, t_3} X_{t_1} X_{t_2} X_{t_3} e^{-\iota\lambda_1 t_1 - \iota\lambda_2 t_2 + \iota(\lambda_1 + \lambda_2) t_3} (\pi - |\lambda_1|)(\pi - |\lambda_2|) \\ &= T^{-1} \sum_{t_1, t_2, t_3} X_{t_1} X_{t_2} X_{t_3} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} (\pi - |\lambda_1|) e^{-\iota\lambda_1 (t_3 - t_1)} (\pi - |\lambda_2|) e^{-\iota\lambda_2 (t_3 - t_2)}\end{aligned}$$

Now,

$$\begin{aligned}&\int_{-\pi}^{\pi} (\pi - |\lambda|) e^{-\iota\lambda(t_3 - t_1)} d\lambda \\ &= \pi^2 \int_{-1}^{1} (1 - |z|) e^{-\iota\pi z (t_3 - t_1)} dz \\ &= \pi^2 \int_{-1}^{1} \Pi(z) * \Pi(z) e^{-\iota\pi z (t_3 - t_1)} dz \\ &= \pi^2 \frac{sin\left(\frac{\pi(t_3 - t_1)}{2}\right)}{\frac{\pi(t_3 - t_1)}{2}}\end{aligned}$$

where $\Pi(t) = \mathbb{I}(|t| \le 0.5)$, and $*$ denote the convolution function. Hence, the final expression would be:

$$\widehat{M_g(f)} = \pi^2 T^{-1} \sum_{t_1,t_2,t_3} X_{t_1} X_{t_2} X_{t_3} \frac{sin\left(\frac{\pi(t_3-t_1)}{2}\right)}{\frac{\pi(t_3-t_1)}{2}} \frac{sin\left(\frac{\pi(t_3-t_2)}{2}\right)}{\frac{\pi(t_3-t_2)}{2}}$$

The terms are non-zero only when $t_3 = t_1 + (2N-1)$ and $t_3 = t_2 + 2N - 1$, if we consider the $t_i$'s to be ordered.

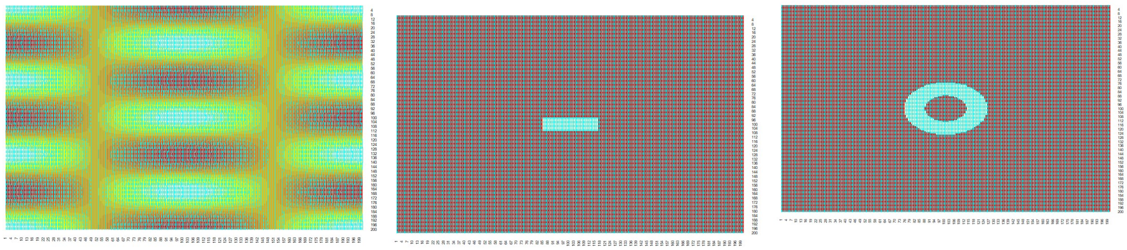


Figure 3.1: Heatmap for different g functions ((a) $g(\underline{\lambda}) = cos(2\lambda)cos(3\lambda)$, (b) $g(\underline{\lambda}) = \mathbb{I}(\lambda_1 \le 0.2)\mathbb{I}(\lambda_2 \le 0.2)$, (c) $g(\underline{\lambda}) = \mathbb{I}(0.1 \le \lambda_1^2 + \lambda^2 \le 0.2)$

## 3.3 Asymptotic Results for Polyspectral Means

The spectral means is classical in the time series literature. For example, in [17] it was shown that the sample spectral mean converges to a Gaussian distribution with a trispectrum appearing in the asymptotic variance. A particular case is the sample autocovariance function, which is asymptotically Gaussian with variance involving a trispectral mean. In this section we prove that the $k^{th}$ polyspectral mean estimate proposed in Section 3.2 is asymptotically normal, showing that cumulants of order higher than 2 tend to zero asymptotically. We will also compute the asymptotic variance term, which involves polyspectra of order $2k+1$.

We will need the following summability assumption on $k^{th}$ order auto-cumulants, as stated in [21].

**Assumption A:**

$$\sum_{v_1,\ldots,v_{k-1}=-\infty}^{\infty} |v_j Cum(v_1,\ldots,v_{k-1},v_k)| < \infty, \qquad \text{for } j = 1,\ldots,k-1. \tag{3.3.1}$$

Assumption A is directly related to the smoothness of the $k^{th}$ order polyspectra.

**Proposition 3.3.1.** Suppose that, for some $k \geq 1$, $\{X_t\}$ is a $(k+1)$th order stationary time series with finite $(2k+1)$th order moment that satisfies Assumption A. Let $g$ be a weight function satisfying the symmetry condition (3.2.1) and has finite integral in $(-\pi,\pi)$. Then as $T \to \infty$,

$$E\widehat{M_g(f)} = M_g(f) + O\left(T^{-1}\right). \tag{3.3.2}$$

Proposition 1 shows that the expectation goes to the theoretical polyspectral mean asymptotically, at a rate of $T^{-1}$. The proof is quite simple, and is given in Section 3.7.1. This result shows that the estimate defined for polyspectral mean is asymptotically unbiased and hence a valid estimate for polyspectral mean. Next, we want to compute the variance, or second cumulant of the polyspectral mean. Since polyspectral mean is a potentially complex value, we need to take cumulant for complex valued processes here. We can consider the cumulant in different directions, viz. $Cum(X,X), Cum(X,\bar{X})$ or $Cum(\bar{X},\bar{X})$. All the cases will have similar proofs, and we will only look at the circular cumulant [34] $Cum(X,\bar{X})$ in this chapter. Here we will briefly go over the outline of the proof (Details given in 3.7.2). We can write:

$$Cum\left(M_g\left(\hat{f}\right), \overline{M_g\left(\hat{f}\right)}\right) = E\left\{M_g\left(\hat{f}\right)\overline{M_g\left(\hat{f}\right)}\right\} - M_g\left(f\right)\overline{M_g\left(f\right)} \tag{3.3.3}$$

Now,

$$\begin{aligned}
&E\left\{M_g\left(\hat{f}\right)\overline{M_g\left(\hat{f}\right)}\right\} \\
&= E\left\{T^{-2k}T^{-2}\sum_{\underline{\lambda},\underline{\omega}}\left(2\pi\right)^{-2k} d\left(\lambda_1\right)\dots d\left(\lambda_k\right) d\left(-[\underline{\lambda}]\right) d\left(-\omega_1\right)\dots d\left(-\omega_k\right) d\left([\underline{\omega}]\right) g\left(\underline{\lambda}\right)\overline{g(\underline{\omega})}\Phi(\underline{\lambda})\Phi(\underline{\omega})\right\} \\
&= T^{-2k-2}\left(2\pi\right)^{-2k}\sum_{\underline{\lambda},\underline{\omega}} g\left(\underline{\lambda}\right)\overline{g(\underline{\omega})}E\left\{\prod_{\lambda\in\underline{\lambda}'} d\left(\lambda\right)\prod_{\omega\in\underline{\omega}'} d(-\omega)\Phi(\underline{\lambda})\Phi(\underline{\omega})\right\}
\end{aligned}$$

where $\underline{\lambda}' = \left(\lambda_1, \dots, \lambda_k, -\sum_{i=1}^{k}\lambda_i\right)$ (and with a similar definition for $\underline{\omega}'$).

Considering only the part inside expectation of the equation, we can see:

$$E\left\{d\left(\lambda_1\right)\dots d\left(\lambda_k\right) d\left(-[\underline{\lambda}]\right) d\left(-\omega_1\right)\dots d\left(-\omega_k\right) d\left([\underline{\omega}]\right)\right\} \tag{3.3.4}$$

$$= \sum_{\sigma\in I_{2k+2}}\prod_{b\in\sigma} Cum(b) \tag{3.3.5}$$

where $\sigma$ is a partition of $\{1, \dots, 2k+2\}$ and $Cum(b)$ is the joint cumulant of the set of indices in b which is an element of $\sigma$. In other words, the elements $Cum(b)$ are the joint cumulants of the Discrete Fourier Transforms at the $\lambda_j$'s whose subscripts belong to the set b of partition $\sigma$. From Lemma 1 of [20], under Assumption A (3.3.1), this term is non-zero iff the sum of the frequencies within the set is 0. The $\Phi(\cdot)$ ensures that the mass is concentrated only on the principal manifold $\sum_{j=1}^{k}\omega_j \equiv 0 \pmod{2\pi}$ of k-dimensional wave and $\sum_{j=1}^{l}\omega_j \not\equiv 0 \pmod{2\pi}$ for all $l \neq k$. In other words, the Fourier frequencies do not lie in any sub-manifold, i.e. no subset of $\underline{\lambda}$ or $\underline{\omega}$ has their sum equal to 0. Hence, the terms will be non-zero in two possible ways:

- There are two partitions of $\{\underline{\lambda}, -\underline{\omega}\}$, which is $\{\underline{\lambda}\}$ and $\{-\underline{\omega}\}$.

- The partitions contain a mixture of $\underline{\lambda}$ and $\underline{\omega}$.

The first case will give the second term of 3.3.3, so the only remaining terms will be those arising from the second case. Now suppose we have m such mixture partitions of $\{\lambda_1, \dots, \lambda_k, -[\underline{\lambda}]\}$ and $\{-\omega_1, \dots, -\omega_k, [\underline{\omega}]\}$, which we shall call constraints. Then all such combinations can be compactly written as:

$$\begin{aligned} \{A\underline{\lambda}' - B\underline{\omega}' = 0_{m\times 1} \Big| \; & A \text{ and } B \text{ are } m \times k+1 \text{ binary matrices such that:} \\ & \text{every row must be non-empty and} \\ & \text{every column must have exactly one } 1\} \end{aligned} \tag{3.3.6}$$

where $\underline{\lambda}' = \{\lambda_1, \dots, \lambda_k, -[\underline{\lambda}]\}$. Same for $\underline{\omega}'$. This is because all the partitions must be disjoint and the union of all partitions must contain all the individual $\lambda$'s. For example, for bispectra $k = 2$ and $m = 2$, we will have the following set of possible choices of $A$ and $B$:

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

Hence taking $A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and $B = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, we will have the constraint $\lambda_1 + \lambda_2 = \omega_1 + \omega_2$, i.e. one partition contains $\{\lambda_1, \lambda_2, \omega_1, \omega_2\}$ while the other partition contains $\{-\lambda_1 - \lambda_2, -\omega_1 - \omega_2\}$. Note that we also have two additional constraints in the form $\sum_j \lambda'_j = \sum_j \omega'_j = 0$. Suppose: $L_m = \{l_1, \dots, l_m\}$ be such that $\sum_{j=1}^{m} l_j = k+1$, $l_j > 0$. For example, for $m = 2$ and

$k = 2,\ L_m = \{(1,2),(2,1)\}$. Additionally, suppose:

$$\zeta_{L_m} = \left\{ A_{m\times(k+1)} \middle| A_{i\cdot} \text{ has } l_i \text{ 1's and every column has exactly one 1} \right\} \tag{3.3.7}$$

Let us define $f_\alpha(\underline{\lambda}_\alpha)$ to be the polyspectra of order $\alpha$ ($\underline{\lambda}_\alpha$ is a vector of length $\alpha$), $r_{Aj}^{(s)}$ and $r_{Bj}^{(s)}$ denote the sum of the $j^{th}$ row of $A$ and $B$ respectively, $r_j^{(s)} = r_{Aj}^{(s)} + r_{Bj}^{(s)} - 1$, and $\lambda_{r_{A_j}}$ denotes the subset of $\underline{\lambda}'$ which contains the elements corresponding to the non-zero positions of $r^{th}$ row of $A$. Similarly, define $\omega_{r_{B_j}}$. Let $L_m$ and $L'_m$ be the set of all possible rows of $A$ and $B$. Using the matrices from (3.3.6), and some combinatorial arguments, it can be shown that:

$$V = \sum_{m=1}^{k+1} \sum_{\underline{l}_m, \underline{l}'_m \in L_m} \sum_{A\in\zeta_{l_m}, B\in\zeta_{l'_m}} \underbrace{\int_{-\pi}^{\pi}\cdots\int_{-\pi}^{\pi}}_{A\underline{\lambda}'-B\underline{\omega}'=0} g(\underline{\lambda})\overline{g(\underline{\omega})} \prod_{j=1}^{m} \tilde{f}_{r_j^{(s)}}\left(\lambda_{r_{A_j}}, \omega_{r_{B_j}}\right) d\underline{\lambda} d\underline{\omega}$$

where $V$ is the asymptotic variance of the polyspectral mean estimate, $\tilde{f}_k(\underline{\lambda}') = f_k(\underline{\lambda})$ is the $k^{th}$ order polyspectra, and $r_j^{(s)} = r_{A_j}^{(s)} + r_{B_j}^{(s)} - 1$ denote the sum of the $j^{th}$ row of A and B subtracted by 1. It is to be noted that the case when all the $\underline{\lambda}'$ and $\underline{\omega}'$ lie in the same partition is covered in the case $m = 1$, and is simply the polyspectral mean of order $2k+1$ with weight function $g(\underline{\lambda})g(\underline{\omega})$, as mentioned in the introduction. Therefore, we have arrived at the next proposition, on the expression of the asymptotic variance (Detailed proof given in 3.7.2).

**Proposition 3.3.2.** Suppose $X_1, \ldots, X_T$ satisfies the zero-mean stationarity conditions, and the weight function g satisfies the symmetry condition. Furthermore, suppose Assumption

1 is satisfied. Then, the second cumulant is of the form $T^{-1}V + O(T^{-2})$, where:

$$V = \sum_{m=1}^{k+1} \sum_{\underline{l}_m, \underline{l}'_m \in L_m} \sum_{A \in \zeta_{l_m}, B \in \zeta_{l'_m}} \underbrace{\int_{-\pi}^{\pi} \dots \int_{-\pi}^{\pi}}_{A\underline{\lambda}' - B\underline{\omega}' = 0} g(\underline{\lambda})\overline{g(\underline{\omega})} \prod_{j=1}^{m} \tilde{f}_{r_j^{(s)}} \left( \lambda_{r_{A_j}}, \omega_{r_{B_j}} \right) d\underline{\lambda} d\underline{\omega}$$

Now that we have derived the mean and variance of the polyspectral mean estimate, all that remains to prove asymptotic normality is to show that the higher order autocumulants goes to 0 at a rate faster that $T^{-1}$. To do that we would need Theorem 2.3.3 of [19], which states:

$$Cum\left( \prod_{j=1}^{J_1} X_{1j}, \dots, \prod_{j=1}^{J_I} X_{Ij} \right) = \sum_{\nu} Cum(X_{ij}; ij \in \nu_1) \dots Cum(X_{ij}; ij \in \mu_p)$$

where the summation is over all indecomposible partition $\nu = \nu_1 \cup \dots \cup \nu_p$ of the table.

$$\begin{pmatrix} (1,1) & \cdots & (1,J_1) \\ . & & . \\ . & & . \\ . & & . \\ (I,1) & \cdots & (I,J_I) \end{pmatrix}$$

Using this result, it can be shown that in our case, cumulants of order r goes to 0 at a rate $T^{-r+1}$ (Detailed Proof given in 3.7.3). Hence we can finally write our theorem:

**Theorem 3.3.1.** Suppose $X_1, \dots, X_T$ is a sample from a stationary time series, and $M_g(f)$ and $\widehat{M_g(f)}$ be as defined earlier. Let g be a Hermitian weight function such that $\int_{\underline{\lambda}} |g(\underline{\lambda})| d\underline{\lambda} < \infty$. Let $A$ and $B$ are $m \times (k+1)$ binary matrices as defined in (3.3.6).

Then, we can write:

$$\sqrt{T}\left(\widehat{M_g(f)} - M_g(f)\right) \Rightarrow \mathcal{N}(0, V)$$

where

$$V = \sum_{m=1}^{k+1} \sum_{\underline{l}_m, \underline{l}'_m \in L_m} \sum_{A \in \zeta_{l_m}, B \in \zeta_{l'_m}} \underbrace{\int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi}}_{A\underline{\lambda}' - B\underline{\omega}' = 0} g(\underline{\lambda})\overline{g(\underline{\omega})} \prod_{j=1}^{m} \tilde{f}_{r_j^{(s)}}\left(\lambda_{r_{A_j}}, \omega_{r_{B_j}}\right) d\underline{\lambda} d\underline{\omega} \tag{3.3.8}$$

The case $m = 1$ in the sum is given by:

$$\int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} g(\underline{\lambda})\overline{g(\underline{\omega})} f_{2k+1}(\lambda_1, \ldots, \lambda_k, -[\underline{\lambda}], \omega_1, \ldots, \omega_k) d\underline{\lambda} d\underline{\omega}$$

.

Therefore, we have obtained the asymptotic distribution of estimate polyspectral mean for general weight function. A simple extension the the above proof shows that the correlation between polyspectral mean of different orders goes to 0 at a rate of $T^{-1}$.

**Corollary 3.3.1.** Suppose we have a stationary time series satisfying all the conditions in the previous theorem. Suppose we consider the polyspectral mean of order $k$ and $k+1$ as $\widehat{M_{g_1}(f_k)}$ and $\widehat{M_{g_2}(f_{k+1})}$ respectively. Then it can be shown that

$$Cov(\widehat{M_{g_1}(f_k)}, \widehat{M_{g_2}(f_{k+1})})$$

goes to 0 at a rate of $T^{-1}$. More precisely:

$$Cov(\widehat{M_{g_1}(f_k)}, \overline{\widehat{M_{g_2}(f_{k+1})}}) = T^{-1}\left\{ \sum_{m=1}^{k+1} \sum_{\underline{l}_m, \underline{l}'_m \in L_m} \sum_{A \in \zeta_{l_m}, B \in \zeta_{l'_m}} \underbrace{\int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi}}_{A\underline{\lambda}' - B\underline{\omega}' = 0} \prod_{j=1}^{m} g_1\left(\underline{\lambda}_k\right) \overline{g_2(\underline{\omega}_{k+1})} \prod_{j=1}^{m} f_{r_j^{(s)}}\left(\lambda_{r_{A_j}}, \omega_{r_{B_j}}\right) \right\}$$

The proof follows from the fact that:

$$
\begin{aligned}
&Cov(\widehat{M_{g_1}(f_k)}, \overline{\widehat{M_{g_1}(f_k)}}) \\
&= E\left\{T^{-k}\sum_{\underline{\lambda}_k}\hat{f}_k(\underline{\lambda}_k)g_1(\underline{\lambda}_k)T^{-k-1}\sum_{\underline{\lambda}_{k+1}}\hat{f}_{k+1}(\underline{\lambda}_{k+1})\overline{g_2(\underline{\lambda}_{k+1})}\right\} - M_{g_1}(f_k)M_{g_2}(f_{k+1}) \\
&= T^{-2k-3}(2\pi)^{2k+1}\sum_{\underline{\lambda}_k,\underline{\omega}_{k+1}} g_1(\underline{\lambda}_k)\overline{g_2(\underline{\lambda}_{k+1})}\sum_{\sigma\in I_{2k+3}}\prod_{b\in\sigma} Cum(b) - M_{g_1}(f_k)M_{g_2}(f_{k+1})
\end{aligned}
$$

In this case, the proof would be similar, except the A matrix will be of dimension $m \times (k+1)$ and B matrix will be of dimension $m \times (k+2)$. The proof can be easily extended to covariance between polyspectral means of any two orders. The next corollary gives the joint distribution of polyspectral mean estimates of different weight functions. It can be shown that the polyspectral mean estimates of different weight functions asymptotically follow a multivariate normal distribution with a non-diagonal covariance matrix, given by 3.3.9. The proof is disussed in the first part of Section 3.7.4.

**Corollary 3.3.2.** Under the assumptions of Theorem 3.3.1 and the given definitions, suppose we consider the polyspectral means of order k for two different weight functions $g_1$ and $g_2$. Then it can be shown that the covariance can be written as:

$$
Cov\left(\widehat{M_{g_1}(f_k)}, \widehat{M_{g_2}(f_k)}\right) = \sum_{m=1}^{k+1}\sum_{\underline{l}_m,\underline{l}'_m\in L_m}\sum_{A\in\zeta_{l_m},B\in\zeta_{l'_m}}\underbrace{\int_{-\pi}^{\pi}\cdots\int_{-\pi}^{\pi}}_{A\underline{\lambda}'-B\underline{\omega}'=0} g_1(\underline{\lambda})\overline{g_2(\underline{\omega})}\prod_{j=1}^{m}\tilde{f}_{r_j^{(s)}}\left(\lambda_{r_{A_j}},\omega_{r_{B_j}}\right)d\underline{\lambda}d\underline{\omega} \tag{3.3.9}
$$

Also, if we take multiple weight functions, then the d-dimensional vector $(\widehat{M_{g_1}(f_k)},\ldots,\widehat{M_{g_d}(f_k)})$ will converge to a multivariate normal distribution with mean 0 and covariance matrix given by (3.3.9).

Corollary 3.3.2 will be useful in devising the linearity test of time series that we will propose in the next section.

## 3.4 Testing of Linear Process Hypothesis Using Bispectrum

As discussed earlier, it is often of interest to determine if a process is linear. [88] showed that it is possible to use quadratic prediction instead of the widely used linear predictor to get significant improvement in prediction error. In this section we will provide a novel test of linearity using bispectrum. As we know, if $\{X_t\}$ is a linear process of the form $X_t = \psi(B)\epsilon_t$, for some known $\psi$, then the bispectrum will be of the form $f_2(\lambda, \omega) = \mu_3 \psi(e^{-\iota\lambda})\psi(e^{-\iota\omega})\psi(e^{\iota(\lambda+\omega)})$. Therefore, we construct a statistic:

$$\mathcal{T}(\lambda, \omega) = \frac{f(\lambda, \omega)}{\psi(e^{-\iota\lambda})\psi(e^{-\iota\omega})\psi(e^{\iota(\lambda+\omega)})} = \frac{f(\lambda, \omega)}{\Psi(\lambda, \omega)}$$

then $\mathcal{T}(\lambda, \omega)$ will be constant under the null hypothesis that $X_t = \psi(B)\epsilon_t$. Hence, for any j and k, the integral

$$\int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \mathcal{T}(\lambda, \omega) e^{\iota\lambda j} e^{\iota\lambda k} d\lambda d\omega$$

should be 0 whenever one of j and k is non zero. Hence, we can construct the following test statistic:

$$\mathcal{T}_{BLT} = \sum_{(j,k)\neq(0,0), 0\leq j,k\leq M} T \left| \langle\langle \hat{\mathcal{T}}(\lambda, \omega)\rangle_j\rangle_k \right|^2$$

where $\langle f \rangle_k = \int_{-\pi}^{\pi} f(\lambda) e^{\iota\lambda k} d\lambda$. The choice of M is considered to be arbitrary. Now, if we consider $g_{j,k}(x_1, x_2) = \frac{e^{\iota j x_1 + \iota k x_2}}{\Psi(x_1, x_2)}$, then the test statistic, after scaling by $V_{j,k}$ (the asymptotic variance corresponding to the weight function $g_{j,k}$), becomes:

$$\mathcal{T}_{BLT} = \sum_{(j,k)\neq(0,0),0\leq j,k\leq M} \frac{T|M_{g_{j,k}}(\hat{f})|^2}{V_{j,k}}$$

Note that under $H_0$, $V_{j,k}$ is known, (cf. (3.3.8)). Hence, $\mathcal{T}_{BLT}$ is a legitimate test statistic. As discussed in Corollary 3.3.2, the asymptotic distribution of a vector of polyspectral means with different choice of $g$ function is a multivariate normal with covariance term given by (3.3.9). Hence, suppose we consider the functions $g_{j,k}(x_1, x_2) = \frac{e^{\iota j x_1 + \iota k x_2}}{\Psi(x_1,x_2)}$, where $j$ and $k$ runs from 0 to $M$, such that both are not zero. Let us construct the a vector of tuples (i,j):

$$\vartheta = \{(0,1),(0,2)\ldots(0,M)(1,0)\ldots(1,M)\ldots(M,0)\ldots(M,M)\}$$

. Then the vector of polyspectral means $\left\{\widehat{M_{g_v}(f)}\middle| v \in \vartheta\right\}$ converges in distribution to a multivariate normal variable with known covariance matrix, say $\mathbb{CV}_{BLT}^{(M)}$. There are many applications where researchers have used linear processes to model time series data, and this test is useful to identify deviations from the linear models they proposed. For example in Section 3.4 we considered the Sunspot Data. [1] fitted an AR(1) model to the dataset, and this test can be used to verify to check whether the model is reasonable for this dataset.

The test considers $\psi(\cdot)$ to be known, and as such the limiting distribution under $H_0$ is completely known. Then we can state the following theorem.

**Theorem 3.4.1.** Suppose $\{X_t\}$ be a stationary time series satisfying Assumption A. Suppose $g_{j,k}$, $\vartheta$ and $\mathbb{CV}_{BLT}^{(M)}$ be as defined earlier. To test $H_0 : X_t = \psi(B)\epsilon_t$ for some known $\psi$, we can construct the test statistic:

$$\mathcal{T}_{BLT} = \sum_{(j,k)\neq(0,0),0\leq j,k\leq M} T\left|\langle\langle\hat{\mathcal{T}}(\lambda,\omega)\rangle_j\rangle_k\right|^2$$

where $\hat{\mathcal{T}}(\lambda,\omega) = \frac{\hat{f}(\lambda,\omega)}{\psi(e^{-\iota\lambda})\psi(e^{-\iota\omega})\psi(e^{\iota(\lambda+\omega)})}$, and $\langle f\rangle_k = \int_{-\pi}^{\pi} f(\lambda)e^{\iota\lambda k}d\lambda$. Then the asymptotic distribution of $\mathcal{T}_{BLT}$ is $\sum_{j=1}^{M^2-1}\lambda_j\zeta_j$, where $\zeta_j$ are iid $\chi_1^2$ random variables, and $\lambda_1,\dots,\lambda_{M^2-1}$ are the eigen values of $\mathbb{CV}_{BLT}^{(M)}$. The exact expression of $\mathbb{CV}_{BLT}^{(M)}$ is given in (3.7.9).

## 3.5 Simulation

We will consider the following models for simulation:

- **AR(2):** $X_t = X_{t-1} - 0.9X_{t-2} + \epsilon_t$, where $\epsilon_t \sim Exp(1) - 1$ (r =1.1, n=100) (Similar to Example 4.7 of [119]).

  In this case, $\phi(z) = 1 - z + 0.9z^2$, $\theta(z) = 1$, where $\phi(z)X_t = \theta(z)\epsilon_t$. Also,

  $$|\phi(e^{-2\pi\iota\omega})|^2 = 2.81 - 3.8cos(2\pi\omega) + 1.8cos(4\pi\omega)$$

  Hence, the spectral density is of the form:

  $$f(\omega) = \frac{1}{\phi(e^{-2\pi\iota\lambda})\phi(e^{2\pi\iota\lambda})} = \frac{1}{2.81 - 3.8cos(2\pi\omega) + 1.8cos(4\pi\omega)}$$

  Similarly the bispectra and trispectra are of the form:

  $$f(\lambda,\omega) = \frac{2}{\phi(e^{-2\pi\iota\lambda})\phi(e^{-2\pi\iota\omega})\phi(e^{2\pi\iota(\lambda+\omega)})}$$
  $$f(\lambda_1,\lambda_2,\lambda_3) = \frac{9}{\phi(e^{-2\pi\iota\lambda_1})\phi(e^{-2\pi\iota\lambda_2})\phi(e^{-2\pi\iota\lambda_3})\phi(e^{2\pi\iota(\lambda_1+\lambda_2+\lambda_3)})}$$

- **ARMA(2,1):** $X_t - X_{t-1} + 0.9X_{t-2} = \epsilon_t + 0.8\epsilon_{t-1}$, where $\epsilon_t \sim Exp(1) - 1$ i.e. $\phi(z) = 1 - z + 0.9z^2$ and $\theta(z) = 1 + 0.8z$, where $\phi(B)X_t = \theta(B)\epsilon_t$. Hence, $\Psi(z) = \frac{\theta(z)}{\phi(z)} = \frac{1+0.8z}{1-z+0.9z^2}$. The spectra, bispectra and trispectra are:

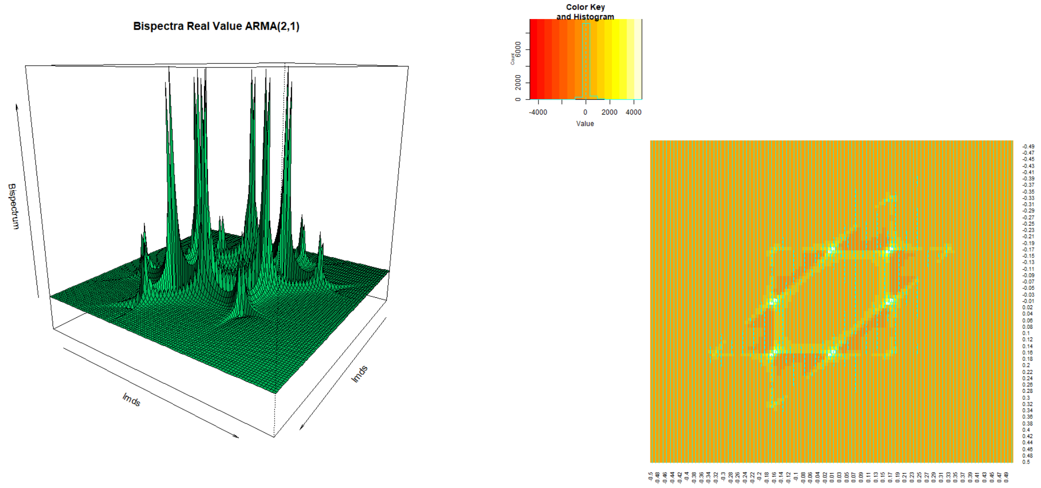


Figure 3.2: Bispectra of ARMA(2,1) Process

$$f(\omega) = |\Psi(e^{-2\pi\iota\lambda})|^2$$

$$f(\lambda,\omega) = 2\Psi(e^{-2\pi\iota\lambda})\Psi(e^{-2\pi\iota\omega})\Psi(e^{2\pi\iota(\lambda+\omega)})$$

$$f(\lambda_1,\lambda_2,\lambda_3) = 9\Psi(e^{-2\pi\iota\lambda_1})\Psi(e^{-2\pi\iota\lambda_2})\Psi(e^{-2\pi\iota\lambda_3})\Psi(e^{2\pi\iota(\lambda_1+\lambda_2+\lambda_3)})$$

The bispectra of the ARMA(2,1) process is depicted in Figure 3.2.

- **Squared Hermite:** $X_t = J_1\,H_1(Z_t) + J_2\,H_2(Z_t) = J_1\,Z_t + J_2\,Z_t^2 - J_2$ where $J_1 = 2$, $J_2 = 5$ and $Z_t$ is a MA(1) process such that $Z_t = \epsilon_t + 0.4\epsilon_{t-1}$, $\epsilon_t \sim Exp(1) - 1$.

The spectra, bispectra, trispectra are as follows:

$$
\begin{aligned}
f(\lambda) &= \sigma_{\epsilon}^2(J_1^2 + J_2^2)(1 + \theta^2 + 2\theta cos(2\pi\lambda)) \\
&= 29(1.16 + 0.8cos(2\pi\lambda)) \\
f(\lambda_1,\lambda_2) &= J_2^3\sqrt{8}(1+\theta^2)\left\{(1+\theta^2)^2 + 2\theta^2\left\{cos(2\pi\lambda_1) + cos(2\pi\lambda_2) + cos(2\pi(\lambda_1+\lambda_2))\right\}\right\} \\
&= 410.12\left\{1.3456 + 0.32\left\{cos(2\pi\lambda_1) + cos(2\pi\lambda_2) + cos(2\pi(\lambda_1+\lambda_2))\right\}\right\} \\
f(\lambda_1,\lambda_2,\lambda_3) &= 15J_2^4(1+\theta^2)^4 + 3J_1^4(1+\theta^2)^2 + 30J_1^2J_2^2(1+\theta^2)^3 + \\
&\quad \left\{15J_2^4(1+\theta^2)^2\theta^2 + 3J_1^4(1+\theta^2)\theta + 15J_1^2J_2^2(1+\theta^2)\theta(1+\theta+\theta^2)\right\}2cos(2\pi\lambda_1) + \\
&\quad \left\{15J_2^4(1+\theta^2)^2\theta^2 + 3J_1^4(1+\theta^2)\theta + 15J_1^2J_2^2(1+\theta^2)\theta(1+\theta+\theta^2)\right\}2cos(2\pi\lambda_2) + \\
&\quad \left\{15J_2^4(1+\theta^2)^2\theta^2 + 3J_1^4(1+\theta^2)\theta + 15J_1^2J_2^2(1+\theta^2)\theta(1+\theta+\theta^2)\right\}2cos(2\pi\lambda_3) + \\
&\quad \Bigg\{J_2^4\left\{(1+\theta^2)^4 + 6\theta^4 + (1+\theta^2)^2\theta^2\right\} + J_1^4\left\{(1+\theta^2)^2 + 2\theta^2\right\} + \\
&\quad J_1^2J_2^2\left\{2(1+\theta^2)^3 + 12\theta^3 + 8(1+\theta^2)^2\theta + 8\theta^2(1+\theta)\right\}\Bigg\}\Bigg\{2cos(2\pi(\lambda_1+\lambda_2)) + \\
&\quad 2cos(2\pi(\lambda_2+\lambda_3)) + 2cos(2\pi(\lambda_3+\lambda_1))\Bigg\} + \Bigg\{15J_2^4(1+\theta^2)^2\theta^2 + 3J_1^4(1+\theta^2)\theta + \\
&\quad 6J_1^2J_2^2\left\{3\theta(1+\theta^2)^2 + 2\theta^2(1+\theta^2)\right\}\Bigg\}2cos(2\pi(\lambda_1+\lambda_2+\lambda_3))
\end{aligned}
$$

Additionally, for each of these models we will consider the following weight functions:

- $g_1(\lambda_1,\lambda_2) = cos(3\lambda_1)cos(\lambda_2)$
- $g_2(\lambda_1,\lambda_2) = \mathbb{I}_{[-0.2,0.2]}(\lambda_1)\mathbb{I}_{[-0.5,0.5]}(\lambda_2)$
- $g(\lambda_1,\lambda_2) = 1 - \sqrt{\frac{\lambda_1^2+\lambda_2^2}{2}}$

The results of the simulation are given in Table 3.1. As we can see for all these cases the computed asymptotic variance in 3.3.8 is close to the simulated ones in all the cases.

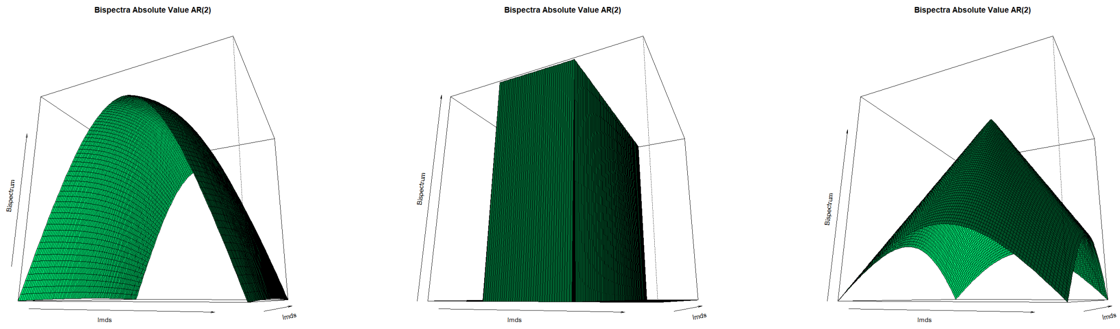


Figure 3.3: 3d Plot of the Weight Functions

Table 3.1: Simulation of different Models and different weight functions (Sample Size: 100)

| Model | g($\underline{\lambda}$) | Scaled MSE | MAPE |
|---|---|---|---|
| Model 1: AR(2) with Exp(1) - 1 | $g_1(\lambda_1, \lambda_2)$ | 0.12 | 0.27 |
| | $g_2(\lambda_1, \lambda_2)$ | 0.19 | 0.37 |
| | $g_3(\lambda_1, \lambda_2)$ | 0.26 | 0.39 |
| Model 2: AR(2) with $\chi_4^2 - 4$ | $g_1(\lambda_1, \lambda_2)$ | 0.07 | $< 0.05$ |
| | $g_2(\lambda_1, \lambda_2)$ | 0.26 | 0.41 |
| | $g_3(\lambda_1, \lambda_2)$ | 1.06 | 0.82 |
| Model 3: ARMA(2,1) with $Exp(1) - 1$ | $g_1(\lambda_1, \lambda_2)$ | 0.39 | 0.48 |
| | $g_2(\lambda_1, \lambda_2)$ | 0.40 | 0.42 |
| | $g_3(\lambda_1, \lambda_2)$ | 0.14 | 0.32 |
| Model 4: ARMA(2,1) with $\chi_4^2 - 4$ | $g_1(\lambda_1, \lambda_2)$ | 0.15 | 0.34 |
| | $g_2(\lambda_1, \lambda_2)$ | 0.28 | 0.46 |
| | $g_3(\lambda_1, \lambda_2)$ | 1.07 | 0.82 |
| Model 5: Sq. Hermite Process | $g_1(\lambda_1, \lambda_2)$ | 0.27 | 0.38 |
| | $g_2(\lambda_1, \lambda_2)$ | 0.14 | 0.27 |
| | $g_3(\lambda_1, \lambda_2)$ | 0.95 | 0.91 |

We also conducted a simulation study for the power of the linearity test we proposed. In this case we took the process $X_t$ be be generated from:

$$X_t = \epsilon_t + 0.4\epsilon_{t-1} + \theta\epsilon_{t-1}^2 - \theta$$

We then changed the value of $\theta$ and computed the power curve for the different choices. Ideally, the $\theta$ will control the non-linearity of the process, and as such increasing it should increase the power of the test, which can be seen in Figure 3.4. We have taken the sample size to be 100, and the choices of $\theta$ ranging from 0 to 10.

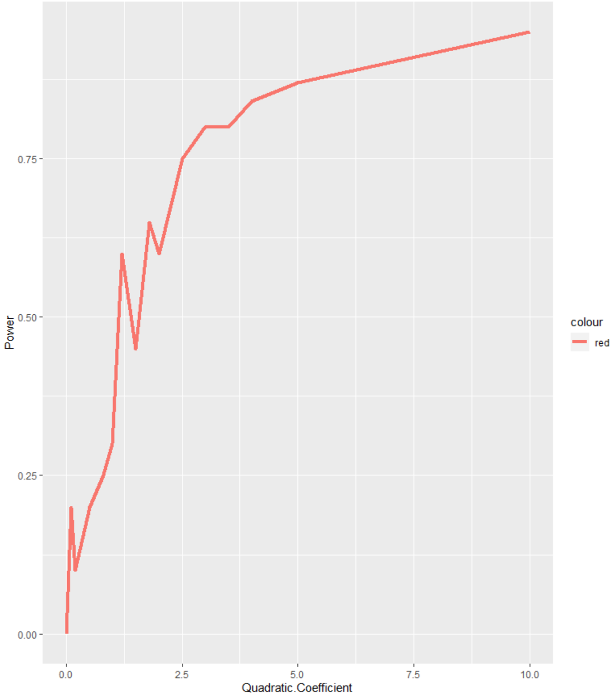


Figure 3.4: Power Curve for different choices of $\theta$

## 3.6 Real Data Analysis

### 3.6.1 Sunspot Data Linearity Test

Sunspots are temporary phenomena on the Sun's photosphere that appear as spots relatively darker than surrounding areas. Those spots are regions of reduced surface temperature arising due to concentrations of magnetic field flux that inhibit convection. It is already known that the solar activity follows a periodic pattern repeating every 11 years. Figure 3.5 shows the time series of the sunspot data collected in a monthly interval.

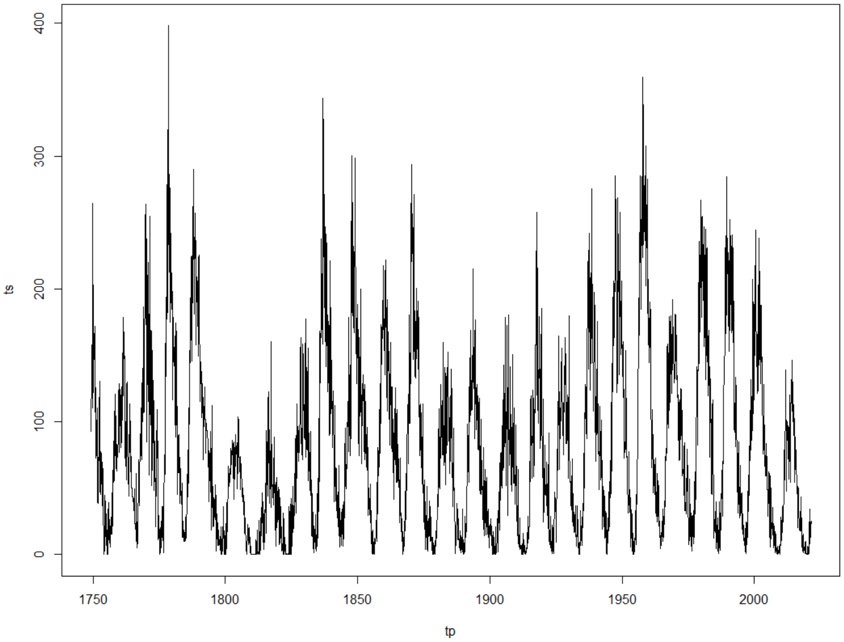


Figure 3.5: Sunspot Data Collected Monthly (Source: [120])

[88] depicted that it is possible to gain significant improvement in prediction in sunspot data by using quadratic prediction instead of linear prediction. This shows that there is high possibility of presence of nonlinearity in the sunspot data. Here we will use the test statistic proposed in Section **??** to test whether the sunspot data follows a particular linear process, namely $X_t = 0.976X_{t-1} + \epsilon_t$ which was used in [1]. In this case:

$$\sum_{(j,k)\neq(0,0),0\leq j,k\leq M} \frac{TM_{g_{j,k}}(\hat{f})^2}{V_{j,k}} = 3740.057$$

with the p-value $< 0.005$.

Hence, our proposed test rejects the null hypothesis that the sunspot data follows the linear process proposed by [1].

### 3.6.2 GDP Trend Clustering

An application of the polyspectral mean can be for time series clustering. As we have discussed earlier, we can extract different types of information from a time series by using different weight functions for the polyspectral mean. In this section, we have analyzed the Gross Domestic Product (GDP) of 136 countries for past 40 years, and have attempted to classify them based on the bispectral mean taken with different weight functions. The time series of the GDP is first differenced to ensure stationarity, and then scaled in order to extract just the trend information (shown in Figure 3.6). Hence, the clustering is based solely on how the GDP varied over time, and not the actual value of GDP. The weight functions considered are as follows:

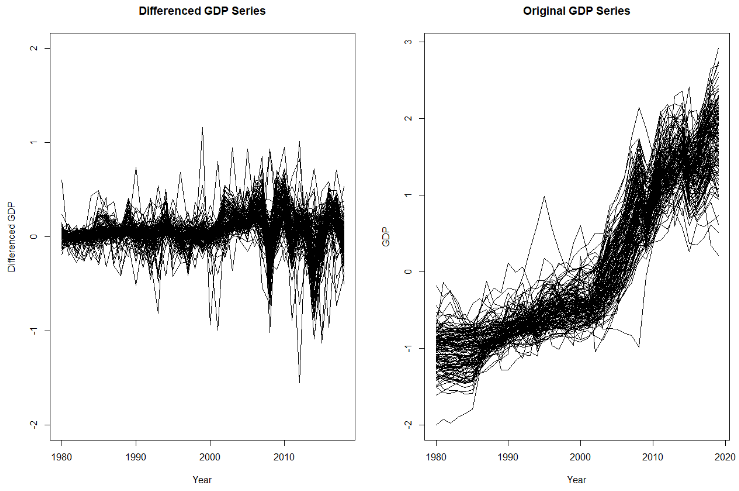


Figure 3.6: Original and Differenced GDP trend for the countries over 40 years (1980-2020)

- $g(\lambda_1, \lambda_2) = \mathbb{I}(a < \lambda_1^2 + \lambda_2^2 < b)$, where a and b are taken such that the interval $(0,1)$ are split into 10 segments. Hence, we have 10 such bispectral means for each annulus.
- $g(\lambda_1, \lambda_2) = (1 - |\lambda_1|)(1 - |\lambda_2|)$
- $g(\lambda_1, \lambda_2) = cos(3\lambda_1)cos(\lambda_2)$

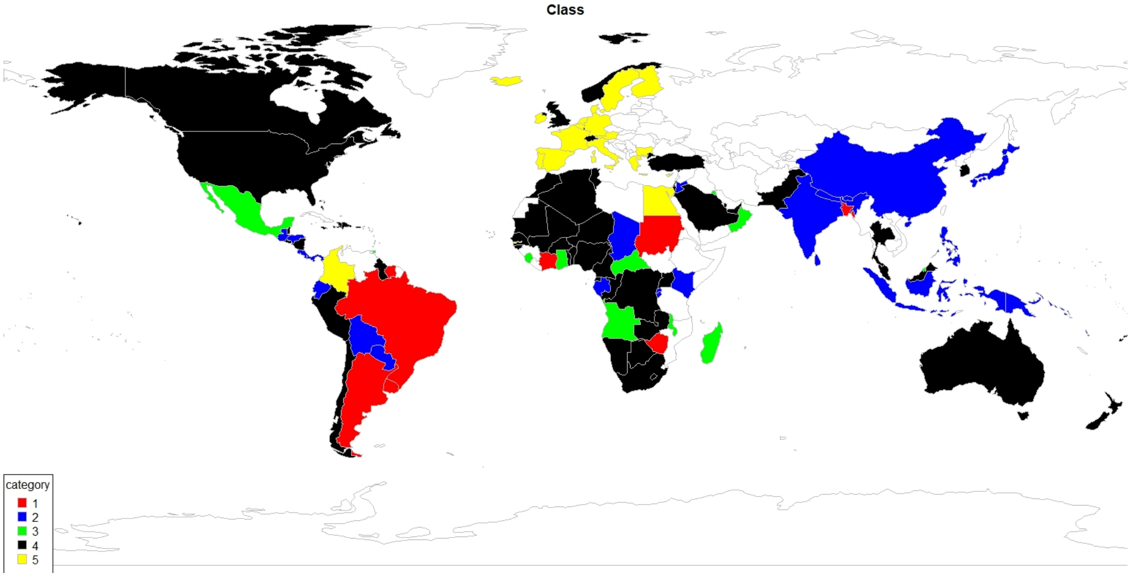


Figure 3.7: World Map GDP Clustering

The computed bispectral means are then used to classify the countries into 5 classes, as shown in Figure 3.7. As we can see, the developed EU countries all fall under the same category.The South Asian and South-East countries form another category, while all types of categories are found in Africa and South America. Some developed countries like USA and least developed countries like Benin seems to fall under same category, since both their time series seem to have similar ascending structure.

## 3.7 Proofs

### 3.7.1 Proof of Proposition 3.3.1

*Proof.*

$$E\left(\widehat{M_g(f_k))}\right) = (2\pi)^k T^{-k-1} \sum_{\underline{\lambda}} E\left(d\left(\lambda_1\right) \dots d\left(\lambda_k,\right) d\left(-[\underline{\lambda}]\right)\right) g\left(\underline{\lambda}\right)$$

Now, using results from [20]:

$$E\left(d\left(\lambda_1\right) \dots d\left(\lambda_k\right) d\left(-[\underline{\lambda}]\right)\right) = \sum_{\sigma \in \tau_n} \prod_{b \in \sigma} Cum(b)$$

the term $Cum(b)$ will be non-zero only when the sum inside is 0, which can only occur when the partition contains all the $\lambda$'s, since the Fourier Frequencies are assumed not to lie in any sub-manifold. Thus:

$$\begin{aligned} E\left(d\left(\lambda_1\right) \dots d\left(\lambda_k\right) d\left(-[\underline{\lambda}]\right)\right) &= \sum_{\sigma \in \tau_n} \prod_{b \in \sigma} Cum(b) \\ &= Cum\left(d\left(\lambda_1\right), \dots, d\left(\lambda_k\right), d\left(-[\underline{\lambda}]\right)\right) \\ &= Tf\left(\underline{\lambda}\right) + O(1) \end{aligned}$$

The last equality is a direct outcome of Lemma 1 of [20].

Hence, using sum to integral and properties of the weight function:

$$
\begin{aligned}
E\widehat{M_g(f)} &= (2\pi)^k T^{-k-1} \sum_{\underline{\lambda}} \{Tf(\underline{\lambda}) + O(1)\}\, g(\underline{\lambda}) \\
&= M_g\left(f\right) + O\left(T^{-1}\right)
\end{aligned}
$$

□

### 3.7.2 Proof of Proposition 2

*Proof.* For the second cumulant, we can consider $Cum(X, X), Cum(X, \bar{X})$ or $Cum(\bar{X}, \bar{X})$. All the cases will have similar proofs, and we will only look at the circular cumulant ([34]) $Cum(X, \bar{X})$ in this paper. We can write:

$$
Cum\left(\widehat{M_g\left(f\right)}, \overline{\widehat{M_g\left(f\right)}}\right) = E\left\{\widehat{M_g\left(f\right)}\,\overline{\widehat{M_g\left(f\right)}}\right\} - M_g\left(f\right)\overline{M_g\left(f\right)} \tag{3.7.1}
$$

Now,

$$
\begin{aligned}
& E\left\{\widehat{M_g\left(f\right)}\,\overline{\widehat{M_g\left(f\right)}}\right\} \\
&= E\left\{T^{-2k}T^{-2}\sum_{\underline{\lambda},\underline{\omega}}\left(2\pi\right)^{2k} d\left(\lambda_1\right)\ldots d\left(\lambda_k\right) d\left(-[\underline{\lambda}]\right) d\left(-\omega_1\right)\ldots d\left(-\omega_k\right) d\left([\underline{\omega}]\right) g\left(\underline{\lambda}\right)\overline{g\left(\underline{\omega}\right)}\right\} \\
&= T^{-2k-2}\left(2\pi\right)^{2k}\sum_{\underline{\lambda},\underline{\omega}} g\left(\underline{\lambda}\right)\overline{g\left(\underline{\omega}\right)} E\left\{\prod_{\lambda\in\underline{\lambda}'} d\left(\lambda\right)\prod_{\omega\in\underline{\omega}'} d(-\omega)\right\}
\end{aligned}
$$

where $\underline{\lambda}' = \left(\lambda_1, \ldots, \lambda_k, -\sum_{i=1}^{k}\lambda_i\right)$. Same for $\underline{\omega}'$.

Looking only at the expectation:

$$E\left\{d\left(\lambda_1\right)\ldots d\left(\lambda_k\right)d\left(-[\underline{\lambda}]\right)d\left(-\omega_1\right)\ldots d\left(-\omega_k\right)d\left([\underline{\omega}]\right)\right\} \tag{3.7.2}$$

$$= \sum_{\sigma\in I_{2k+2}}\prod_{b\in\sigma} Cum(b) \tag{3.7.3}$$

where $\sigma$ is a permutation of $\{1,\ldots,2k+2\}$ and $Cum(b)$ is the joint cumulant of the set of indices in b which is an element of $\sigma$. From Lemma 1 of [20], under the assumption that $\sum|v_j c'_{a_1,\ldots,a_k}(v_1,\ldots,v_{k-1})| < \infty$, this term is non-zero iff the sum of the frequencies is 0. Hence, the terms will be non-zero in three possible ways:

- There are two partitions of $\{\underline{\lambda}', -\underline{\omega}'\}$, (where $\underline{\lambda}'$ is $\{\lambda_1,\ldots,\lambda_k,-[\underline{\lambda}]\}$) which is $\{\underline{\lambda}'\}$ and $\{-\underline{\omega}'\}$. Let us call this partition $\sigma_0$. Since, sum of elements of both partitions is zero, both of them will contribute non-zero values to the product, rendering the product to be non-zero. It can be written as:

$$\begin{aligned}
&\prod_{b\in\sigma_0} Cum(b)\\
&= Cum\left(d(\lambda_1),\ldots,d(\lambda_k),d\left(-[\underline{\lambda}]\right)\right)Cum\left(d(-\omega_1),\ldots,d(-\omega_k),d\left([\underline{\omega}]\right)\right)\\
&= \left(Tf(\underline{\lambda})+O(1)\right)\left(Tf(-\underline{\omega})+O(1)\right)\\
&= T^2 f(\underline{\lambda})f(-\underline{\omega})+O(T)
\end{aligned}$$

Hence, entering it in the sum:

$$
\begin{aligned}
&T^{-2k-2}(2\pi)^{2k}\sum_{\underline{\lambda},\underline{\omega}} g(\underline{\lambda})\overline{g(\underline{\omega})}\left\{T^{2}f(\underline{\lambda})f(-\underline{\omega})+O(T)\right\}\\
&=T^{-2k}\,(2\pi)^{2k}\sum_{\underline{\lambda},\underline{\omega}} g(\underline{\lambda})\overline{g(\underline{\omega})}f(\underline{\lambda})f(-\underline{\omega})+(2\pi)^{2k}T^{-2k}\sum_{\underline{\lambda},\underline{\omega}} g(\underline{\lambda})\overline{g(\underline{\omega})}O\left(T^{-1}\right)
\end{aligned}
$$

Thus, the final expression will be:

$$
M_g(f)\overline{M_g(f)}+O\left(T^{-1}\right) \tag{3.7.4}
$$

which would cancel the second term of (3.3.3), leaving a $O(T^{-1})$ term.

- A second case would be where all the $\underline{\lambda}'$ and $\underline{\omega}'$ fall in the same partition.

$$
\begin{aligned}
&\prod_{b\in\sigma_0} Cum(b)\\
&=Cum\left(d(\lambda_1),\ldots,d(\lambda_k),d\left(-[\underline{\lambda}]\right),d(-\omega_1),\ldots,d(-\omega_k),d\left([\underline{\omega}]\right)\right)\\
&=\left(Tf_{2k+1}(\underline{\lambda},-[\underline{\lambda}],-\underline{\omega})+O(1)\right)
\end{aligned}
$$

Hence, entering it in the sum:

$$
\begin{aligned}
&T^{-2k-2}(2\pi)^{2k}\sum_{\underline{\lambda},\underline{\omega}} g(\underline{\lambda})\overline{g(\underline{\omega})}\left\{Tf(\underline{\lambda},-[\underline{\lambda}],-\underline{\omega})+O(1)\right\}\\
&=T^{-2k-1}(2\pi)^{2k}\sum_{\underline{\lambda},\underline{\omega}} g(\underline{\lambda})\overline{g(\underline{\omega})}f_{2k+1}(\underline{\lambda},-[\underline{\lambda}],-\underline{\omega})+(2\pi)^{2k}T^{-2k}\sum_{\underline{\lambda},\underline{\omega}} g(\underline{\lambda},\underline{\omega})O\left(T^{-2}\right)\\
&\sim T^{-1}\int_{-\pi}^{\pi}\ldots\int_{-\pi}^{\pi} g(\underline{\lambda})\overline{g(\underline{\omega})}f_{2k+1}(\underline{\lambda},-[\underline{\lambda}],-\underline{\omega})d\underline{\lambda}d\underline{\omega}+O(T^{-2})
\end{aligned}
$$

Hence, the final expression will be a polyspectral mean of order $2k + 1$, with a weight function depending on the original weight function. This case is in particular a special case of the third case, which we will discuss next.

- The other cases where the term would be non-zero are when the sums of all the elements within a block of a partition are 0, i.e. if the blocks of a partition $\sigma$ are $b_1, b_2, \ldots, b_p$, and $b_i$ block contains elements $\lambda_{b_{i1}}, \ldots, \lambda_{b_{il}}, -\omega_{b_{i1}}, \ldots, -\omega_{b_{im}}$, then the sums of all those elements must be 0 for all the blocks $b_i$. This is because if any of the blocks has elements which does not add up to 0, the corresponding $Cum(b)$ will be zero, and hence $\prod_{b\in\sigma} Cum(b)$ will become zero for that partition. Now, we have ensured that no subset of the Fourier frequencies lies in a sub-manifold, i.e. for no subset of $\lambda_1, \ldots, \lambda_k$, the sum would be 0. Hence, for one partition to have a non-zero value, all of it's blocks must contain at least one element from each of $\lambda$ and $\omega$.The only other case is covered in the first case, where the partition contains all elements of $\lambda$ and $\omega$. Now that we know that every block of the partition contains a mixture of $\underline{\lambda}'$ and $\underline{\omega}'$, let us call each mixture a constraint, since each mixture should have a sum 0, and hence would produce a constraint and thereby reducing the degrees of freedom by 1. Suppose we have m such constraints, i.e. m linear combinations of $\lambda_k$ and $\omega_k$ are 0. This also means we have m blocks. Then we will have $T^m$ from the integrand, since every block would contribute a T from the cumulant, and a O(1) from the residual. The sum will hence run for only the above cases, i.e. only when the linear combinations of the elements of partitions are equal to 0. In other words, only when there exists $m \times (k + 1)$ matrices $A$ and $B$ such that:

$$A\underline{\lambda}' - B\underline{\omega}' = 0_{m\times 1}$$

where $\underline{\lambda}' = \{\lambda_1, \ldots, \lambda_k, -[\underline{\lambda}]\}$. Same for $\underline{\omega}'$.

Also, every row of $A$ (and $B$) must be disjoint (this means that the rows of A (and B) cannot have 1 in the same position) and contain 1 or 0, and not all 0. Also, every column must have exactly one 1. This is because:

- A Fourier Frequency $\lambda$ can be only in one partition, and hence the rows of A must be disjoint
- All columns must have at least one 1, since a Fourier frequency $\lambda$ must be in at least one partition.
- Also, a column cannot have more than one 1, since that would violate the disjoint row property. Hence, every column must have exactly 1 "1".

An example would be:

$$A_{m,k+1} = \begin{pmatrix} 1 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \end{pmatrix}_{m \times k}$$

where m is the number of constraints. Suppose $f_\alpha(\underline{\lambda}_\alpha)$ be the polyspectra of order $\alpha$ ($\underline{\lambda}_\alpha$ is a vector of length $\alpha$). Also, assume $r_{Aj}^{(s)}$ and $r_{Bj}^{(s)}$ denotes the sum of the $j^{th}$ row of $A$ and $B$ respectively. Let $r_j^{(s)} = r_{Aj}^{(s)} + r_{Bj}^{(s)} - 1$. Let $\lambda_{r_{A_j}}$ denotes the subset of $\underline{\lambda}'$ which contains the elements corresponding to the non-zero positions of $r^{th}$ row of $A$. Similarly, define $\omega_{r_{B_j}}$. Then for each term we have ($\tilde{f}(\underline{\lambda}') = f(\underline{\lambda})$):

$$T^m \prod_{j=1}^{m} \tilde{f}_{r_j^{(s)}} \left( \lambda_{r_{A_j}}, \omega_{r_{B_j}} \right) + O\left(T^{m-1}\right)$$

Now, for each of these quantities, we have $m-1$ constraints, $m$ from the matrices, and 1 is subtracted because the sum is 0. Then, we have for each of the term:

$$\begin{aligned}&T^{-2k-2}\sum_{\underline{\lambda},\underline{\omega}} g\left(\underline{\lambda}\right)\overline{g\left(\underline{\omega}\right)}\left\{T^{m}\prod_{j=1}^{m}\tilde{f}_{r_j^{(s)}}\left(\lambda_{r_{A_j}},\omega_{r_{B_j}}\right)+O\left(T^{m-1}\right)\right\}\\&=T^{-1}T^{-2k+m-1}\sum_{\underline{\lambda},\underline{\omega}} g\left(\underline{\lambda}\right)\overline{g\left(\underline{\omega}\right)}\prod_{j=1}^{m}\tilde{f}_{r_j^{(s)}}\left(\lambda_{r_{A_j}},\omega_{r_{B_j}}\right)+O\left(T^{-2}\right)\end{aligned}$$

This goes to 0 at a rate of $T^{-1}$. Suppose: $L_m=\{l_1,\ldots,l_m\}$ be such that $\sum_{j=1}^{m} l_j=k+1$, $l_j>0$. For example, for $k=2$ and $m=2$, $L_2=\{(1,2),(2,1)\}$. Suppose:

$$\zeta_{L_m}=\left\{A_{m\times(k+1)}\middle|A_{i\cdot}\text{ has }l_i\text{ 1's and every column has exactly one 1}\right\}\tag{3.7.5}$$

Then for every choice of m, we will have a corresponding $L_m$, and for each elements $l_m\in L_m$, we will have corresponding sets $\zeta_{l_m}$. Then we will have the sum running over all possible choices of matrices A and B within the sets $\zeta_{l_m}$ for all choices of $l_m\in L_m$ such that $A\underline{\lambda}'-B\underline{\omega}'=0$. Then each of the summands will be of the form:

$$\begin{aligned}&T^{-2k-2}T^{m}\sum_{\underline{\lambda},\underline{\omega}|A\underline{\lambda}'-B\underline{\omega}'=0} g\left(\underline{\lambda}\right)\overline{g\left(\underline{\omega}\right)}\prod_{j=1}^{m}\tilde{f}_{r_j^{(s)}}\left(\lambda_{r_{A_j}},\omega_{r_{B_j}}\right)\\&=T^{-1}T^{-2k+m-1}\sum_{\underline{\lambda},\underline{\omega}|A\underline{\lambda}'-B\underline{\omega}'=0} g\left(\underline{\lambda}\right)\overline{g\left(\underline{\omega}\right)}\prod_{j=1}^{m}\tilde{f}_{r_j^{(s)}}\left(\lambda_{r_{A_j}},\omega_{r_{B_j}}\right)\end{aligned}\tag{3.7.6}$$

Then, the second cumulant would be of the form $T^{-1}V + O\left(T^{-2}\right)$, where:

$$
\begin{aligned}
V &= \sum_{m=1}^{k+1} \sum_{\underline{l_m} \in L_m} T^{-2k+m-1} \sum_{\underline{\lambda},\underline{\omega} \mid A\underline{\lambda}'-B\underline{\omega}'=0} g\left(\underline{\lambda}\right) \overline{g\left(\underline{\omega}\right)} \prod_{j=1}^{m} \tilde{f}_{r_j^{(s)}} \left(\lambda_{r_{A_j}}, \omega_{r_{B_j}}\right) \\
\sim V &= \sum_{m=1}^{k+1} \sum_{\underline{l}_m, \underline{l}'_m \in L_m} \sum_{A \in \zeta_{l_m}, B \in \zeta_{l'_m}} \underbrace{\int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi}}_{A\underline{\lambda}'-B\underline{\omega}'=0} g(\underline{\lambda}) \overline{g(\underline{\omega})} \prod_{j=1}^{m} \tilde{f}_{r_j^{(s)}} \left(\lambda_{r_{A_j}}, \omega_{r_{B_j}}\right) d\underline{\lambda} d\underline{\omega}
\end{aligned}
$$

The case $m = 1$ is the second case mentioned earlier, which would give a polyspectral mean of order $2k + 1$.

□

### 3.7.3 Proof of Theorem 3.3.1

*Proof.* Proposition 1 and 2 gives the mean and variance of the limiting distribution. The only thing remaining to prove asymptotic normality is to show that the higher order cumulants of the scaled transformation goes to zero as n goes to $\infty$. Suppose ($\underline{\lambda}' = (\lambda_1, \ldots, \lambda_k, -[\underline{\lambda}])$), then we can write ($\kappa_r$ be the $r^{th}$ order joint cumulant):

$$
\begin{aligned}
\kappa_r(\widehat{M_g(f)}) &= C_r \kappa(\underbrace{\widehat{M_g(f)}, \ldots, \widehat{M_g(f)}}_{\text{r times}}) \\
&= C_r \kappa\left(T^{-k-1} \sum_{\underline{\lambda}_1} g(\underline{\lambda}_1) \prod_j d(\lambda'_{1j}), \ldots, T^{-k-1} \sum_{\underline{\lambda}_r} g(\underline{\lambda}_r) \prod_j d(\lambda'_{rj})\right) \\
&= C_r T^{-rk-r} \sum_{\underline{\lambda_1}} \cdots \sum_{\underline{\lambda}_r} g(\underline{\lambda}_1) \ldots g(\underline{\lambda}_r) \kappa\left(\prod_j d(\lambda'_{1j}), \ldots, \prod_j d(\lambda'_{rj})\right)
\end{aligned}
$$

From Theorem 2.3.3 of [19], we know:

$$\kappa\left(\prod_{j=1}^{J_1} X_{1j}, \ldots, \prod_{j=1}^{J_I} X_{Ij}\right) = \sum_{\nu} \kappa(X_{ij}; ij \in \nu_1) \ldots \kappa(X_{ij}; ij \in \mu_p)$$

where the summation is over all indecomposible partition $\nu = \nu_1 \cup \ldots \cup \nu_p$ of the table.

$$\begin{pmatrix} (1,1) & \cdots & (1,J_1) \\ . & & . \\ . & & . \\ . & & . \\ (I,1) & \cdots & (I,J_I) \end{pmatrix}$$

In our case:

$$\kappa\left(\prod_{j=1}^{k+1} d(\lambda'_{1j}), \ldots, \prod_{j=1}^{k+1} d(\lambda'_{rj})\right) = \sum_{\nu} \kappa\left(d(\lambda'_{ij}) : ij \in \nu_1\right) \ldots \kappa\left(d(\lambda'_{ij}) : ij \in \nu_p\right)$$

and the table is:

$$\begin{pmatrix} (1,1) & \cdots & (1,k+1) \\ . & & . \\ . & & . \\ . & & . \\ (r,1) & \cdots & (r,k+1) \end{pmatrix}$$

From the expression of cumulants, we know that each of those $\kappa$'s are non zero only when their sum is equal to zero, and can give at most T for each of the partitions, i.e. the highest order from each of these elements is $T^p$. Hence, to get the leading term, we would need to take maximum number of such partitions such that the sum of $\lambda$'s in the partitions is zero. By our definition of $\widehat{M_g(f)}$, no subset of the $\underline{\lambda}_i$'s fall in a sub-manifold. Hence, each of these partitions must contain either all distinct $\underline{\lambda}'s$, or a mixture of $\lambda$'s. However, all distinct will not be an indecomposable partition. Hence, we are left with only mixture of $\lambda$'s. Also indecomposable would mean that we don't have any single $\underline{\lambda}$, i.e. all the partitions must be a mixture of all the $\lambda$'s. This is same as the matrix notation we described earlier. Here we just have $r$ matrices $A_1, \dots A_r$ such that $A_1\underline{\lambda}_1 + \dots A_r\underline{\lambda}_r = 0_{p\times 1}$ Hence, p partitions would give p constraints. thereby giving an order of $T^p$, and with these comes $p-1$ constraints. Therefore, the sum would need $T^{-rk+p-1}$. The remainder term would be $T^{-r+1}$. Thus, the cumulants of order higher than $r$ will go to 0 at a rate faster than $T^{-1}$, thereby proving the asymptotic normality of the scaled transformation.

□

### 3.7.4 Proof of Theorem 2

*Proof.* We defined the following test statistic:

$$\mathcal{T}_{BLT} = \sum_{(j,k)\neq(0,0), 0\leq j,k\leq M} T \left| \langle\langle \hat{\mathcal{T}}(\lambda,\omega)\rangle_j\rangle_k \right|^2$$

Now, suppose $g_{j,k} = (x_1, x_2) = \frac{e^{\iota j x_1 + \iota k x_2}}{\Psi(x_1,x_2)}$, then the test statistic after scaling by $V_{j,k}$ (the asymptotic variance obtained in Theorem 3.3.1) becomes:

$$\mathcal{T}_{BLT} = \sum_{(j,k)\neq(0,0),0\leq j,k\leq M} \frac{T|M_{g_{j,k}}(\hat{f})|^2}{V_{j,k}}. \tag{3.7.7}$$

First, we need to show that under this $g_{j,k}$, $M_g(f)$ is 0, which is trivial since under the null hypothesis $f(\underline{\lambda})$ is a constant multiple of the denominator of $g_{j,k}(\underline{x})$. Hence, when $(j,k) \neq (0,0)$, the integral will always be 0. Hence, in this case, the asymptotic distribution of $\sqrt{\frac{T\widehat{M_g(f)}}{V}}$ is $\mathcal{N}(0,1)$ for any given function g„ where $V$ is the corresponding asymptotic variance.

Let us first consider the joint distribution of $\left(\widehat{M_{g_{j_1,k_1}}(f)}, \widehat{M_{g_{j_2,k_2}}(f)}\right)$. Suppose we want to find the limiting distribution of $\upsilon = a_1\widehat{M_{g_{j_1,k_1}}(f)} + a_2\widehat{M_{g_{j_2,k_2}}(f)}$. We can write:

$$\begin{aligned}
\upsilon &= a_1(2\pi)^{-k}T^{-k}\sum_{\underline{\lambda}^{(a_1)}} T^{-1}d\left(\lambda_1^{(a_1)}\right)\dots d\left(\lambda_k^{(a_1)}\right)d\left(-\sum_{j=1}^{k}\lambda_j^{(a_1)}\right)g_{j_1,k_1}\left(\underline{\lambda}^{(a_1)}\right)\\
&\quad + a_2(2\pi)^{-k}T^{-k}\sum_{\underline{\lambda}^{(a_2)}} T^{-1}d\left(\lambda_1^{(a_2)}\right)\dots d\left(\lambda_k^{(a_2)}\right)d\left(-\sum_{j=1}^{k}\lambda_j^{(a_2)}\right)g_{(j_2,k_2)}\left(\underline{\lambda}^{(a_2)}\right)\\
&= (2\pi)^{-k}T^{-k}\sum_{\underline{\lambda}} T^{-1}d\left(\lambda_1\right)\dots d\left(\lambda_k\right)d\left(-\sum_{j=1}^{k}\lambda^{j}\right)h(\underline{\lambda})
\end{aligned}$$

where $h(\underline{\lambda}) = a_1 g_{(j_1,k_1)}(\underline{\lambda}) + a_2 g_{(j_1,k_1)}(\underline{\lambda})$, and since we have already established the asymptotic normality of polyspectral mean for any function, it follows that any linear combination is asymptotically normal. Therefore, the joint distribution is asymptotically normal.

Hence, finally, we now need to find the asymptotic covariance between $\widehat{M_{g_{j_1,k_1}}(f)}$ and $\widehat{M_{g_{j_2,k_2}}(f)}$.

$$E\left\{\widehat{M_{g_{j_1,k_1}}(f)}\overline{\widehat{M_{g_{j_2,k_2}}(f)}}\right\} - M_{g_{j_1,k_1}}(f)M_{g_{j_2,k_2}}(f)$$

As we saw earlier, one term that would come out is when the partition is $\underline{\lambda}$ and $\underline{\omega}$, there complete disjoint partitions. We only need to show that other partitions will not exist in this case. The covariance will be similar to form of V established in Theorem 3.3.1 as mentioned in Corollary 3.3.2:

$$Cov\left(\widehat{M_{g_1}(f_k)}, \widehat{M_{g_2}(f_k)}\right) = \sum_{m=1}^{k+1} \sum_{\underline{l}_m, \underline{l}'_m \in L_m} \sum_{A \in \zeta_{l_m}, B \in \zeta_{l'_m}} \underbrace{\int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi}}_{A\underline{\lambda}' - B\underline{\omega}' = 0} g_1(\underline{\lambda})\overline{g_2(\underline{\omega})} \prod_{j=1}^{m} \tilde{f}_{r_j^{(s)}}\left(\lambda_{r_{A_j}}, \omega_{r_{B_j}}\right) d\underline{\lambda} d\underline{\omega} \tag{3.7.8}$$

Hence, now we have established the joint distribution of $(M_{g_{j_1,k_1}}, M_{g_{j_2,k_2}})$ is a multivariate normal with covariance matrix:

$$\begin{pmatrix} V_{j_1,k_1} & CV_{(j_1,k_1),(j_2,k_2)} \\ CV_{(j_1,k_1),(j_2,k_2)} & V_{j_2,k_2} \end{pmatrix}$$

for any arbitrary $j_1, j_2, k_1, k_2$. Therefore $\upsilon$ is a multivariate normal distribution with a covariance matrix, say $V_\Upsilon$.

Suppose we have a $M^2 - 1$ vector where each element correspond to a set $(i,j)|i,j \leq M, (i,j) \neq (0,0)$, defined as earlier. Then, the asymptotic distribution of statistic $\mathcal{T}_{BLT}$

under null distribution can be written as $X^T X$, where $X \sim \mathcal{N}_{M^2-1}(0, \mathbb{CV}_{BLT}^{(M)})$, where:

$$\mathbb{CV}_{BLT}^{(M)}(a,b) = \begin{cases} 1 & \text{if } a = b \\ \frac{Cov\left(M_{g_{j_a}}(f_k), M_{g_{j_b}}(f_k)\right)}{\sqrt{V_{j_a,k_a} V_{j_b,k_b}}} & \text{if } a \neq b \end{cases} \tag{3.7.9}$$

Suppose $\mathbb{CV}_{BLT}^{(M)} = P^T \Lambda P$, where $\Lambda = (\lambda_1, \ldots, \lambda_{M^2-1})$ are the eigen values of $\mathbb{CV}_{BLT}^{(M)}$. Let $U = PY$, so that $U \sim \mathcal{N}_{M^2-1}(0, I)$.Then, $X^T X = U^T \Lambda U = \sum_{j=1}^{M^2-1} \lambda_j U_j^2$. Hence, the asymptotic distribution under null can be given by a weighted sum of $\chi_1^2$ random variables where the weights are given by the eigen values of the covariance matrix $\mathbb{CV}_{BLT}^{(M)}$. □

# Chapter 4

# THANOS: A Predictive Model of Electoral Campaigns using Twitter Data and Opinion Polls

## 4.1 Introduction

The rise of social media has been one of the most critical events in this century. Social media has given the common people a platform to share their views with local, state, national, and international community members. Ideas, political speeches, and social commentary are now readily available in every corner of the world. Not long after its inception, social media proved helpful in marketing, politics, and social relations. Political campaigns in democratic countries worldwide have become heavily dependent on social media to promote citizen engagement. For example, the U.S. presidential campaigns of Barrack

Obama extensively used different social media platforms, like Twitter, Facebook, and MySpace. The Trump campaign in 2016 and 2020 deployed sophisticated social media efforts tailored to targeted audiences. Thus, it has become possible to analyze social media posts to forecast political and societal events, including electoral outcomes, and build effective influence campaigns. While a skeptic may note here that many voters do not maintain active accounts on social media platforms and are unaware of this online discourse, we would note that the media often includes citations to social media in print and online reporting. Many "conventional" media sources have adopted graphics that enable them to duplicate Twitter and Facebook content in their articles, extending the reach of these campaigns.

Several studies have been conducted regarding this new realm of social interaction. There exists a train of thought that social media data do no better than chance in predicting electoral campaigns ([53]). Social media discourse and debate may not be representative of the broader electorate and/or eligible voters that will turn out on election day. However, the fact that social media has immense potential to influence political campaigns cannot be refuted. The 2016 U.S. election served as a wake-up call for the American political elite and intelligence community. In a U.S. national election, the intervention by a malign foreign actor, Russia, was unprecedented in scope and impact [60]. A flurry of studies and analyses have since focused on the Russian influence campaign in the U.S. ([35], [60], [116], [140], [112]) and campaigns to alter outcomes in other countries—Eastern Europe ([61]), Sweden ([75]), and Romania & Hungary ([124]). Other studies have more generally explored the impact of social media on the 2016 election ([5]), specifically examined the mechanics of the Russian influence campaign ([94]), or tried to explain the causal mechanisms behind the electoral outcome ([93], [57]). Despite such widespread research on the influence of social media, less attention has been devoted to the underlying general mechanisms for generating influence on these platforms and the degree to which this influence translates into real-world voting behavior. One such study was done on the electoral campaigns in Germany ([136]) for the federal election of the national parliament, which took place on September $27^{th}$, 2009. In another study, [4] performed a sentiment analysis on Facebook during the 2016 U.S. Presidential Election and examined the dynamics between candidate posts and comments received on Facebook.

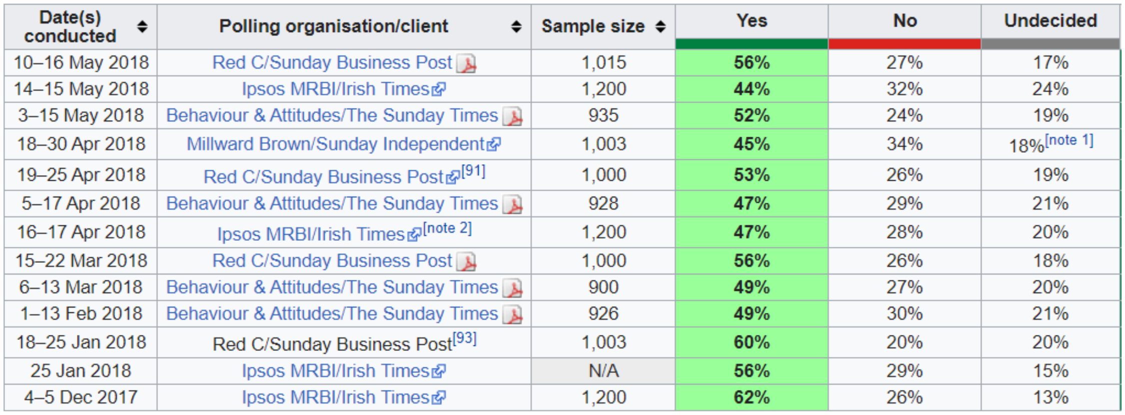
However, works on using social media to successfully forecast elections is sparse ([27], [133], [65], [26]). Our work considers the application of statistical analysis using Twitter data to predict the outcomes of recent elections in Ireland and the United States. For the first analysis, this chapter will aim to forecast the outcome of the voting for the $36^{th}$ Amendment of the Irish Constitution referendum (May 25, 2018) using a model built from Twitter data and several opinion polls conducted during that period. The second analysis involves forecasting the electoral outcomes of the 2018 mid-term election in the United States, again using only Twitter Data and opinion polls conducted during that period. The primary differences in our attempt from earlier works, are (i) to combine the opinion polls conducted during the campaigns with the Twitter data, and (ii) to incorporate the dynamic network structure in our time series model. For (i), we note that the opinion polls, representing the traditional source of information, suffer from time-lags in tracking effects of influence campaigns in quickly evolving scenarios and are often less sensitive in predicting electoral outcomes. On the other end, social media data, such as Twitter posts, lack fair representation of the entire voting population and may not be very effective by themselves for prediction purposes. However, we show that by cleverly combining the two sources of information, it is possible to generate very accurate predictions of the election outcome, even in cases where the races are very close. We do this by building suitable statistical models based on some crucial network features of the Twitter data. For campaigns where the margins of wins are reasonably large, we show that predictions based on a simple model referred to as the THOS (Twitter Hashtag-based Opinion Survey) predictions adequately predict the campaign outcomes. However, the simple model is ineffective for predicting the outcomes when the margin of win is small. For such cases, we develop a more complex model and the corresponding predictor, referred to as the THANOS (Twitter Hashtag and Network-based Opinion Survey) model/predictor incorporates additional network features to yield better prediction accuracy. An essential contribution of the chapter is constructing the feature variables, including one that uses the Harmonic centrality measures of the associated networks. Applied to the 2018 US Elections, the THANOS model correctly predicted the winning candidates in 11 out of the top 12 closest Senate races; The remaining one (Florida) was the closest race where the margin of the winning candidate was only 0.12% and the THANOS prediction was ”undecided.” In particular, the predictions generated by the THANOS model compare favorably with competing predictive modeling approaches, including those of FiveThirtyEight.com. See Section 5 for more details.

The rest of the chapter is organized as follows. Section 4.2 describes the data sets used in the chapter for predicting election campaign outcomes. It also describes relevant background information on Network analysis used in our modeling approach. Section 4.3 develops the two predictive models, the THOS and the THANOS, combining opinion polls with Twitter data. Sections 4.4 and 4.5 respectively describe predictions from our models for the Irish referendum on 36th constitutional amendment and the 2018 US Senate elections, respectively. Section 4.6 gives a brief discussions of the results and some concluding remarks.

## 4.2 Background

### 4.2.1 Descriptions of Data Examples

To illustrate the predictive modeling approaches of the chapter, we considered two recent electoral campaigns where results of several opinion polls and social media data were readily available. The first of these is the Ireland Vote on the 36th constitution referendum. Abortion has been subjected to criminal penalty in Ireland by the Offences against the Person Act since 1861. The Eighth Amendment of the Constitution Act 1983 inserted a subsection recognizing the equal right to life of the pregnant woman and the unborn. The Pro-Life Amendment Campaign instigated this for fear that liberal legislators might weaken the 1861 prohibition. In 1992, the X case ruled that abortion is permitted where pregnancy threatens a woman's life, including by risk of suicide. However, no regulatory framework existed till 2013, until the introduction of the Protection of Life During Pregnancy Act 2013, which defined the circumstances and processes within which abortion in Ireland could be legally performed. Two famous events drove this: The A, B, and C vs. Ireland case in the European Court of Human Rights in 2010 and the death of Savita Halappanavar after a miscarriage in 2012. The 1983 referendum is commonly viewed to equate the life of a pregnant woman with the fetus, making abortion unavailable in almost all circumstances. It is also seen as affecting maternal healthcare because a woman loses her right to refuse consent to medical treatment during pregnancy. Even though the Protection of Life During Pregnancy Act 2013 defined the circumstances under which abortion could be legally performed, it still prohibited legal abortion in many cases. The proposed $36^{th}$ amendment, also known as the repeal of the $8^{th}$ amendment, allows the government to legislate on abortion. The proposed legislation aligned Ireland with most European



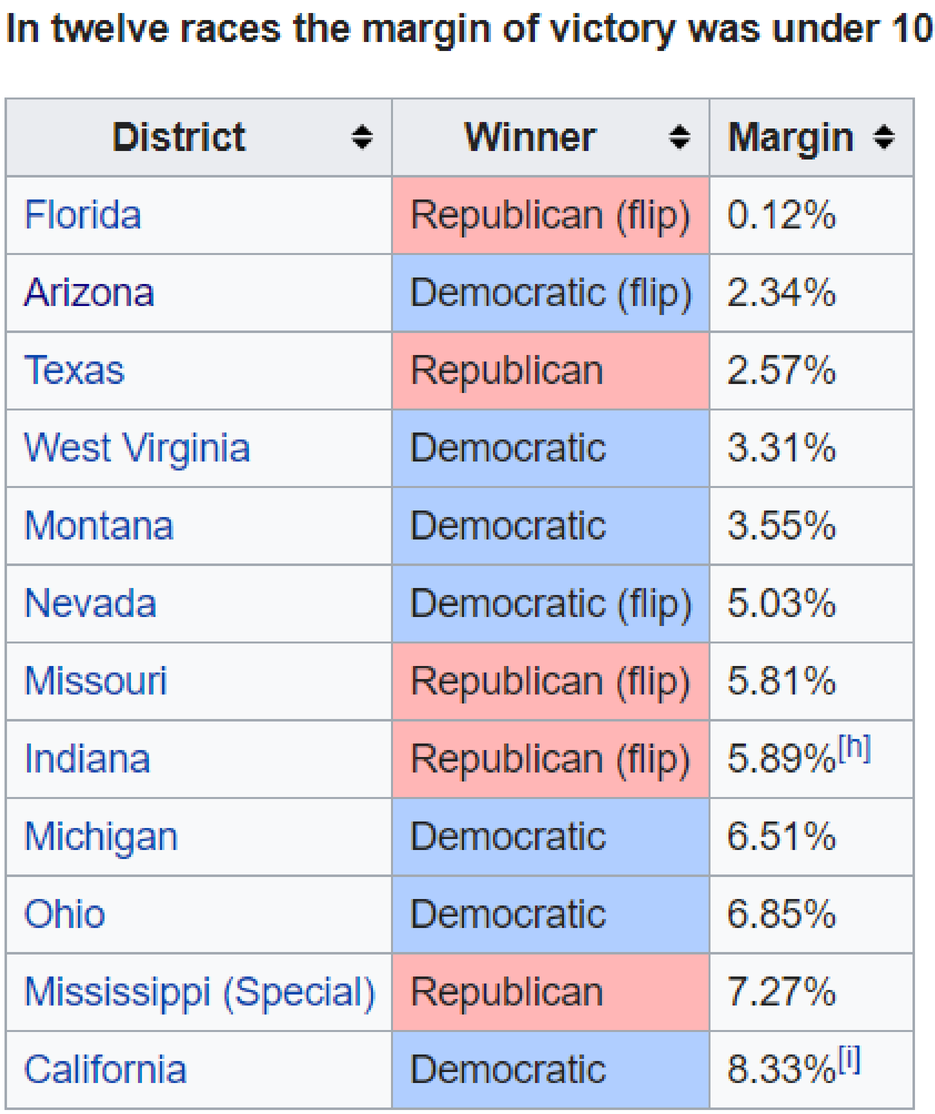
| Date(s) conducted | Polling organisation/client | Sample size | Yes | No | Undecided |
|---|---|---|---|---|---|
| 10–16 May 2018 | Red C/Sunday Business Post | 1,015 | **56%** | 27% | 17% |
| 14–15 May 2018 | Ipsos MRBI/Irish Times | 1,200 | **44%** | 32% | 24% |
| 3–15 May 2018 | Behaviour & Attitudes/The Sunday Times | 935 | **52%** | 24% | 19% |
| 18–30 Apr 2018 | Millward Brown/Sunday Independent | 1,003 | **45%** | 34% | 18%[note 1] |
| 19–25 Apr 2018 | Red C/Sunday Business Post[91] | 1,000 | **53%** | 26% | 19% |
| 5–17 Apr 2018 | Behaviour & Attitudes/The Sunday Times | 928 | **47%** | 29% | 21% |
| 16–17 Apr 2018 | Ipsos MRBI/Irish Times[note 2] | 1,200 | **47%** | 28% | 20% |
| 15–22 Mar 2018 | Red C/Sunday Business Post | 1,000 | **56%** | 26% | 18% |
| 6–13 Mar 2018 | Behaviour & Attitudes/The Sunday Times | 900 | **49%** | 27% | 20% |
| 1–13 Feb 2018 | Behaviour & Attitudes/The Sunday Times | 926 | **49%** | 30% | 21% |
| 18–25 Jan 2018 | Red C/Sunday Business Post[93] | 1,003 | **60%** | 20% | 20% |
| 25 Jan 2018 | Ipsos MRBI/Irish Times | N/A | **56%** | 29% | 15% |
| 4–5 Dec 2017 | Ipsos MRBI/Irish Times | 1,200 | **62%** | 26% | 13% |

Figure 4.1: Opinion Polls For Ireland Voting

countries, allowing for abortion on request up to the $12^{th}$ week of pregnancy (subject to medical regulation). After 12 weeks, abortion would only be available in cases of fatal fetal anomaly, if the pregnant woman's life was at risk, or if her health was at risk of serious harm. Cases after 12 weeks would have to be approved by two doctors. The dataset for this example consists of Twitter data collected during the campaign and opinion polls conducted by various organizations from December 2017 through May 2018, as shown in Figure 4.1. We will combine the two data sources to predict the referendum's outcome using the (simpler) THOS model, as described in Section 3. In this case, the actual proportion of votes in favor of the amendment was 66.4%.

The second example we consider here involves the outcomes of the 2018 US Election. The election was conducted on November 6, 2018, for 35 of the 100 seats in the US Senate and all 435 seats in the US House of Representatives. Figure 4.2 gives the list of states/senate seats, the winning party, and the margins of win for the top 12 closest races.

The Twitter Data in this study was obtained from [143] and consisted of $1,387,688$ tweets from $708,916$ users, over the time period of November $20, 2017$ to March $1, 2019$. We also used opinion polls from several US organizations for predictive modeling. For races where



In twelve races the margin of victory was under 10%.

| District | Winner | Margin |
|---|---|---|
| Florida | Republican (flip) | 0.12% |
| Arizona | Democratic (flip) | 2.34% |
| Texas | Republican | 2.57% |
| West Virginia | Democratic | 3.31% |
| Montana | Democratic | 3.55% |
| Nevada | Democratic (flip) | 5.03% |
| Missouri | Republican (flip) | 5.81% |
| Indiana | Republican (flip) | 5.89%[h] |
| Michigan | Democratic | 6.51% |
| Ohio | Democratic | 6.85% |
| Mississippi (Special) | Republican | 7.27% |
| California | Democratic | 8.33%[i] |

Figure 4.2: Close Races in US 2018 Election

the margin of win was more than 5%, the simpler THOS model performed well. As a result, we primarily concentrated on the senate races with winning margins below 5% to illustrate the predictive accuracy of the THANOS model. See Section 5 for the detailed analyses and results.

### 4.2.2 Construction of a Network for Twitter data and its relation to influence campaigns

Social media has emerged as a potential platform for political campaigns to gain momentum. Propaganda from political campaigns can reach a vast audience, primarily owing to mentions and retweets on Twitter. To better understand how influence propagates through social media, the growth of the underlying network is of primary interest. Figure 4.3 exhibits the network structure developing over time for one particular case (Ireland 2018 Campaign). The network is created using retweets and with four "influencers" (Users with top retweets), two from each campaign ("yes" and "no"). The nodes in the graph represent the users in the data, with a directed edge existing from node A to node B if a tweet from user A was retweeted by user B. The "Influencers" form the central nodes in the four clusters (or cliques). As we can see, the network initially started with four clusters and dynamically expanded with interactions among all four clusters.

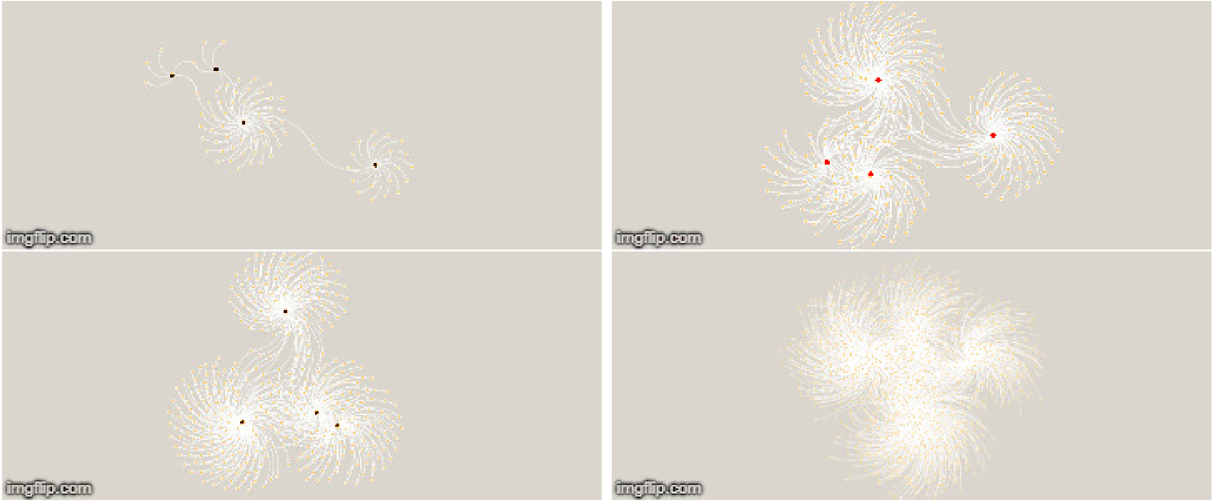


Figure 4.3: Network Growth in Ireland Dataset

The growth pattern of the network in the bottom panel of Figure 4.3 suggests copious interactions within and among the four cliques, which indicates that the influencers from both campaigns influence the common audience. The next important question from this discussion is, therefore, how much an "influencer influences a common user", and what is the social significance of an "influencer," i.e., the extent to which they can influence a campaign. A well-studied approach is to use centrality measures to define the amount of influence a user has on a network. A centrality measure can be defined as a function that assigns a score to a user, indicating the number of interactions the user has with other nodes in the network. For example, the four influencers in Figure 4.3 will have high centrality measures, while those on the fringes will have low scores. We can also obtain a measure of the centrality score of the network using the graph centrality score ([48]), which calculates the interaction level of a graph based on the centrality scores of its nodes. This might indicate campaign effectiveness or enable classification between campaign and non-campaign Twitter datasets. Here we will consider two popular measures of centrality scores:

- *Degree Centrality Score:* The Degree Centrality Score of a given node $v$, denoted by $C_D(v)$, is the number of edges incident upon $v$. It measures the first level interaction experienced by user $v$. The graph centrality in this case is given by

  $$\frac{\sum_{i=1}^{|V|}\{C_D(v^*) - C_D(v_i)\}}{|V|^2 - 3|V| + 2},$$

  where $|V|$ is the size of the Network and $v^*$ is the node with the highest degree centrality score.

- *Harmonic Centrality ($H(v)$)* [ [86]]: Let $d(y,v)$ denotes the distance between nodes $y$ and $v$. The Harmonic Centrality of a node $v$ is defined as $H(v) = \frac{1}{\sum_{y \neq v} d(y,v)}$,

In the next section, we will use the (dynamic) network features to build our statistical models.

## 4.3 Models

The statistical models developed in this chapter are time series regression models based on appropriately chosen features from the Twitter database. The dependent variable ($y_t$) is a transformation of the outcome from the opinion polls conducted during the campaign period. For our modeling purpose, we will only use a two-party system, which is a reasonable assumption for both examples. For the Ireland case, there were only two parties under consideration while, in the US election, the Democrats and the Republicans are the only two major competitors of the elections, with the other political parties accounting for negligible vote shares. Our first model is given by:

$$\log \frac{y_t}{1-y_t} = \beta_0 + \beta_1 \bar{x}^{(1)}_{t,h,l} + \beta_3 \bar{x}^{(2)}_{t,h,l} + \beta_4 \frac{\bar{x}^{(2)}_{t,h,l}}{\bar{x}^{(1)}_{t,h,l}} + \epsilon_t, \tag{4.3.1}$$

where

- $y_t$ := proportion of Votes obtained by Party 1 with respect to Party 2 in the opinion polls,

- $\bar{x}^{(k)}_{t,h,l}$ := proportion of 10 most popular Hashtags corresponding to Party $k$ averaged over the period of $(t-h, t-l)$, $k = 1, 2$, where $h \geq l$ determine the lagged time-window, and

- $\epsilon_t$ := error variable at time t.

In this model, we use hashtag frequency to define the covariates. The underlying idea is that tweets from users leaning towards a particular party will contain more hashtags related to the corresponding party. The choice of top 10 hashtags is arbitrary and can be tuned to get better predictions. The third feature $\frac{\bar{x}^{(2)}_{t,h,l}}{\bar{x}^{(1)}_{t,h,l}}$ gives the relative number of Party 2 hashtags with respect to Party 1 hashtags that occur in the campaign dataset. We avoided delving deeper into the tweets' texts, primarily owing to the lack of reliable sentiment analysis dictionaries in languages other than English, which is not always the most dominant language of the datasets. For example, the most dominant language in the Ireland dataset was French. However, if we consider the frequency of the most popular hashtags, we can avoid the language barrier.

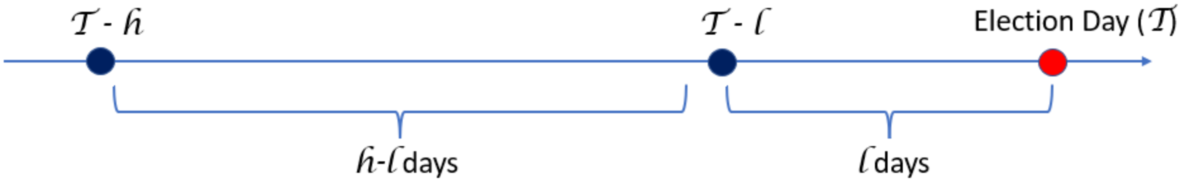


Figure 4.4: Choice of h and l

The logit transform is applied to the poll predictions to convert the support of the dependent variables to the entire real line. The most popular hashtags are classified into two groups from either campaign and are used in the model, along with their ratio. The parameter $h$ signifies how much of the history we need to predict the final outcome accurately, and $l$ determines how far ahead we can predict the outcome. A pictorial representation is given in Figure 4.4. We can provide predictions based on different choices of $h$ and $l$. Suppose that the prediction of the proportion of votes obtained by Party 1 obtained from (4.3.1) for one choice of $h$ and $l$ in Model 1 is defined as $\hat{p}^{TH}_{h,l}$, where $TH$ stands for Twitter Hashtag. The prediction is obtained by fitting the linear regression model and using the inverse logit transform on the obtained prediction $\left(\frac{exp(x)}{1+exp(x)}\right)$. Using this model, we can propose a prediction of the outcome, which will be henceforth called the THOS (Twitter Hashtag-based Opinion Survey) forecast, defined as

$$\bar{p}^{TH}_{\mathbf{h},\mathbf{l}} = \frac{1}{|(\mathbf{h},\mathbf{l})|}\sum_{h,l}\hat{p}^{TH}_{h,l}.$$

Thus, the final forecast $\bar{p}^{TH}_{\mathbf{h},\mathbf{l}}$ is the average of all the predictions obtained by using different choices of $h$ and $l$. The averaging step makes the predictor stable and guards against the effects of large and sudden changes in the polling numbers at isolated time points on the final prediction.

The above model seems to work well for elections with wide victory margins, such as the Ireland 36th amendment vote on legalizing abortion and the election for the senate seat in Connecticut in US 2018 mid-term elections. In both of these cases, the winning party got more than 60% of votes. However, for closely fought elections (e.g. the Arizona senate seat in the 2018 US mid-term election), the THOS forecast seemed inadequate. Hence, we introduced additional network features of the Twitter data into our model. We used the number of retweets of the top influencer from either campaign (based on degree centrality) and also the harmonic centrality for the most influential nodes in the Twitter dataset. The introduction of centrality scores seemed to drastically improve our model predictions, and also the adjusted $R^2$ of the model. The modified model is given by:

$$\log\frac{y_t}{1-y_t} = \beta_0 + \beta_1\bar{x}^{(1)}_{t,h,l} + \beta_3\bar{x}^{(2)}_{t,h,l} + \beta_4\frac{\bar{x}^{(2)}_{t,h,l}}{\bar{x}^{(1)}_{t,h,l}} + \beta_5 h_t + \beta_6 r^{(1)}_t + \beta_7 r^{(2)}_t + \epsilon_t. \tag{4.3.2}$$

Here, $y_t$, $\bar{x}_{t,h,l}^{(k)}$ and $\epsilon_t$ are as in (4.3.1), and the other quantities are defined as follows:

- $h_t$ : Harmonic centrality score of the most influential user in the network at time t (for example, for the 2018 US election, it was *@realDonaldTrump*);
- $r_t^{(k)}$ : Retweet Proportion of the most influential user from Party $k$ at time $t$, $k = 1, 2$ ( for example, for the 2018 US election, $@realDonaldTrump$ for Republicans and $@krassenstein$ for Democrats ) .

The aim of the second model is to improve upon the first with the use of additional network features. To provide some insight, we now discuss consideration and motivation behind the specific choices of features made in formulating model (4.3.2). Since the edges in the network based on the Twitter data are formed by retweets and mentions, it readily gives us an overview of the election campaign activities of the supporters of the two parties. It was observed that the top "influencers" from the two parties formed a clique around them, the size of which dynamically increased over time. The *proportion* at which (the size of) a clique around an influencer from Party 1 increases with respect to that of an influencer from Party 2 would give an indication to the public sentiment during the course of the campaign. In this project, we considered a scaled degree centrality score (based only on 1-step retweets) of the most influential user from either campaigns. Additionally, we also took the Harmonic Centrality Score of the most retweeted "influencer" of the entire dataset, which is expected to capture effectiveness of the lead influencer in influencing the entire electorate (including the undecided voters). The time series using these network features, added to the earlier hashtag features, are fitted against the time series of opinion polls, to create the THANOS (Twitter Hashtag and Network based Opinion Survey) model, given in (4.3.2).

Next, suppose that the prediction obtained from this model is defined as $\hat{p}_{h,l}^{THN}$, where we use the superscript THN (a shorthand for Twitted Hashtags and Network) to refer to the THANOS model. As in the case of the THOS model, we can provide a prediction for the probability of winning the election (by Party 1) using $\hat{p}_{h,l}^{THN}$-s for different $h, l$, as

$$\bar{p}_{\mathbf{h},\mathbf{l}}^{THN} = \frac{1}{|(\mathbf{h},\mathbf{l})|} \sum_{h,l} \hat{p}_{h,l}^{THN}.$$

In the next two sections, we will apply the models to the Ireland and US elections data, and consider their accuracy in well-separated and closely fought races. We also compare the our results with existing predictions put forward by some well known election prediction websites.

## 4.4 Ireland Vote on the $36^{th}$ constitution referendum

### 4.4.1 Data Description

In this section, we combine opinion polls and Twitter data to predict the outcome of Ireland vote on the Abortion Rights amendment. Opinion polls conducted by different organizations over the period of December 2017 through May 2018 are summarized in Figure 4.1. The sample sizes ranged from 900 to 1200, with one opinion poll having no sample size information. However, for all the opinion polls, the date and the percentage of 'yes'/'no' votes were recorded. The Twitter data was collected for one and a half months, from April 13, 2018 to May 25, 2018, during the campaign for the $36^{th}$ amendment (Source: [143]) and the vote was conducted on May 25, 2018. The dataset contained 2,279,396 tweet ids, which reduced to 1,933,397 tweets after passing through the Hydrator to obtain the actual tweets; It was not possible to collect tweets from users who have deleted their twitter accounts after the collection of the tweet ids.

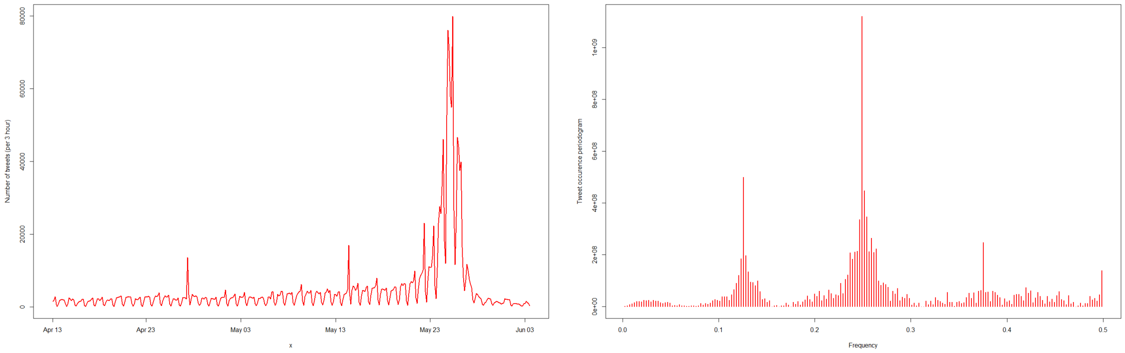


Figure 4.5: Analysis of Time Series of Tweets. The first plot indicates the time distribution of the tweets with a granulation level of 3 hours. The second plot is a periodogram of the differenced time series, indicating a peak at a frequency of 8, and thereby exhibiting the daily periodic pattern of the tweets.

The time series of the times of tweets, aggregated over a 3-hour window, is given in the top panel of Figure 4.5. The plot shows a sharp peak around May 25, the day of the vote, and a couple of small peaks during the campaign. One such peak is seen around May 13, after Facebook and Google disclosed that pro-life agencies from outside Ireland were trying to influence the campaign. Additionally, the time series of the tweets exhibited a periodic pattern. A periodogram analysis (cf. the bottom panel of Figure 4.5) reveals a sharp peak around a frequency of 0.125, giving the period to be 8. Since we have taken a granularity of 3 hours, this suggests a periodic pattern every 24 hours.

### 4.4.2 Prediction of Outcome

For prediction of the outcome of the May 25 vote. we extracted information only on the hashtags of the Twitter data. Although it might be tempting to use the actual text of the tweets, it is extremely difficult to conduct a meaningful sentiment analysis. This is primarily due to the fact that even though we can take bigrams and trigrams (i.e., two- and three-word patterns) from the texts, it still might not give a clear indication as to which side the user is leaning towards. Also, as indicated earlier, the absence of a “good” sentiment dictionary covering different languages used in the tweets makes it quite difficult to use the text information; The most frequently occurring language of the tweets was French, and surprisingly, not English. As a result, it is difficult to use standard sentiment analysis tools for this case. By suitably choosing the hashtag information alone, we could generate good predictions. An advantage of this approach is that the proposed methodology can be applied more widely even in situations where the tweets are not in English.

To construct relevant network features using the hashtags, we collected the most frequently occurring hashtags for the two campaigns and extracted the number of times they occurred over time in the twitter dataset. The set of hashtags used are given in Table 1 - with the “yes” hashtags being associated with the “vote yes” campaign and “no” hashtags with the other campaign. Note that the referendum was in effect a choice between repeal (“yes” campaign) and continuation “no” campaign) of the provisions of the 8th Amendment of the constitution of Ireland which explains the hashtags containing ‘8th’ in Table 1.

Figure 4.6 shows that the time series of the number of hashtags for the two campaigns and also the proportion of tweets in support of the “yes” campaign in 3 hour intervals over the campaign period. Although the number of hashtags for the “yes” campaign was

Table 4.1: Hashtags Chosen

| "Yes" Hashtags | "No" Hashtags |
|---|---|
| *repealthe8th* | *savethe8th* |
| *togetherforyes* | *lovebothvoteno* |
| *voteyes* | *voteno* |
| *repeal* | *loveboth* |
| *repealedthe8th* | *votenotoabortion* |
| *yes* | *prolife* |
| *together4yes* | *no* |

overwhelmingly larger for the "yes" campaign than the "no" campaign, there was an indication of "mood swing" in the Twitter database. The right panel shows that there was a sharp decrease in the proportion of "yes" hashtags for the initial time period, which then increased again from around the end of April. This is also seen in the proportion of supporters of the "yes" campaign in the opinion polls. Thus, it seems logical to build a model using this feature, i.e., the proportion of hashtags for the "yes" campaign in the Twitter database.

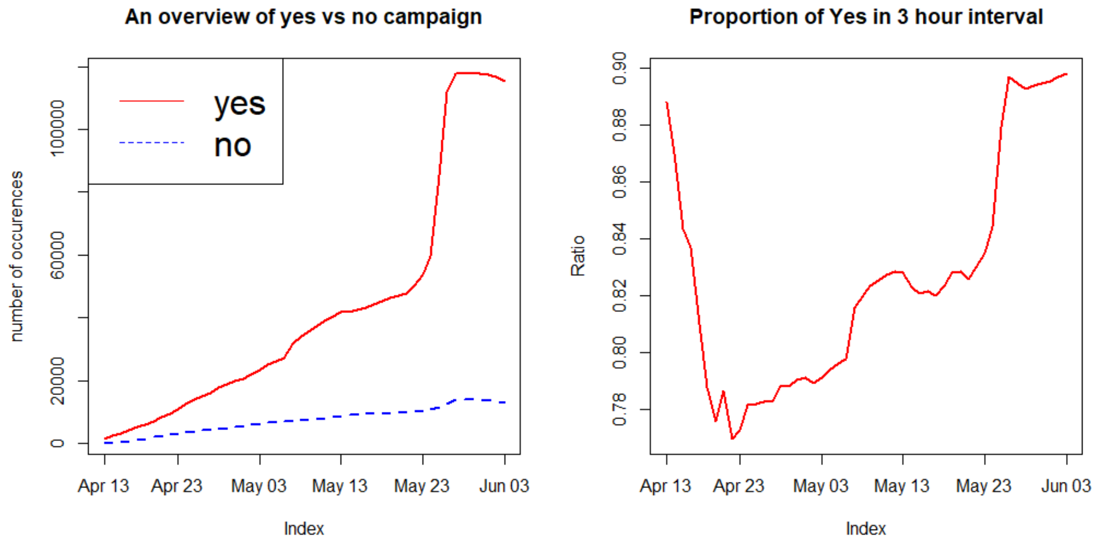


Figure 4.6: Hashtag Frequency of Two Parties

In an exploratory data analysis, the THOS model was fitted for different lags, ranging from past 3 hours to the entire history, i.e. from the day the data was available. The adjusted $R^2$ can be seen to flatten out after about 15 days. Thus, it was enough to consider the Twitter data for past 15 days, since we wouldn't lose any information if we ignored the data before that. In other words, to predict today's vote's outcome using Twitter, it is enough to look at the data for the last 15 days. The fit summary gives a high positive $\beta_3$ which implies that the proportion of "yes" hashtags is an important feature. The predicted THOS proportion of the final vote using this model is **0.6725**, in comparison with the actual value of **0.664**. The close prediction shows that it might actually be possible to predict the outcome of the vote using the Twitter data.

## 4.5 US Election 2018

In this section, we will use our models to predict the outcomes of the US 2018 election using Twitter Data and opinion polling. The election was conducted on November 6, 2018 for 35 of the 100 seats in the US Senates and all 435 seats in the US House of Representatives. As indicated in Section 2, we will primarily concentrate on the senate election for our prediction purposes using the THOS and the THANOS models and compare their performance with existing predictions.

The Twitter Data in this study consist of $1,387,688$ tweets from $708,916$ users, over the time period of November $20,2017$ to March $1,2019$. For each senate election, we have subsampled further to select only those tweets that are relevant to the voting campaigns for that seat. In the Ireland example, the victory margin was substantial but that was not always the case for the US election. Hence, we would need a better model for efficient predictions.

For space consideration, we will report the results of our predictive analysis for 4 states in particular – Connecticut (Democrats 59.53%), Montana (Democrats 50.33%), Arizona (Democrats 50%) and Florida (Republicans 50.06%). The last three of these appear in the list of top 12 close races in the US 2018 election as shown in Figure 4.2, Section 2. Connecticut was a one-sided contest, similar to the Ireland referendum.

### 4.5.1 Connecticut

The Connecticut election experienced a clearly separated race, with the incumbent Democrat Senator Chris Murphy winning with 59.53% votes. This election outcome is similar to the Ireland votes, and we applied Model 1 to see the performance. In this case, the model can be written as:

$$\log \frac{y_t}{1-y_t} = \beta_0 + \beta_1 \bar{x}^{(d)}_{t,h,l} + \beta_3 \bar{x}^{(r)}_{t,h,l} + \beta_4 \frac{\bar{x}^{(r)}_{t,h,l}}{\bar{x}^{(d)}_{t,h,l}} + \epsilon_t \tag{4.5.1}$$

where, as in (4.3.1), $y_t$ is the proportion of votes obtained by Democrats in the opinion polls and $\bar{x}^{(d)}_{t,h,l}$ and $\bar{x}^{(r)}_{t,h,l}$ are respectively the proportions of Democrat and Republican Hashtags averaged over the period of $(t-h, t-l)$. The opinion polls used in this model are form multiple sources, conducted prior to the election. The hashtags used in $\bar{x}^{(r)}_{t,h,l}$ and $\bar{x}^{(d)}_{t,h,l}$ are the top 10 hashtags from Republican and Democrat campaigns, respectively. The predictions obtained from the model obtained from different choice of $h$ and $l$ are given in Table 4.2.

Here the prediction target is the relative percentage of Democrat votes over Republican votes, i.e. $\frac{v_d}{v_d+v_r}$ ($v_d(v_r)$: Votes obtained by Democrats (Republics)). In this case, the true value was **0.6002**. As we can see the predictions for all choices of $h$ and $l$ are pretty close to the actual value. The THOS model forecast (i.e., the average of the entries in the table) in this case is **0.5928**, with relative error ( defined as []prediction - true]/true) as

Table 4.2: $\hat{p}^{TH}$ Forecast for Connecticut Senate in US 2018 Midterm Election for different choices of h and l

| | $l = 1$ | $l = 2$ | $l = 5$ | $l = 10$ | $l = 20$ |
|---|---|---|---|---|---|
| $h = 5$ | 0.5953305 | 0.5945630 | - | - | - |
| $h = 10$ | 0.5954857 | 0.5957060 | 0.5980478 | - | - |
| $h = 15$ | 0.6059773 | 0.6077476 | 0.6063503 | 0.5646253 | - |
| $h = 20$ | 0.6126172 | 0.6095569 | 0.5724295 | 0.5376829 | - |
| $h = 50$ | 0.6133662 | 0.5975079 | 0.5889431 | 0.5826463 | 0.5912204 |

Table 4.3: $\hat{p}^{TH}$ Forecast for Montana Senate in US 2018 Midterm Election for different choices of h and l

| | $l = 1$ | $l = 2$ | $l = 5$ | $l = 10$ | $l = 20$ |
|---|---|---|---|---|---|
| $h = 5$ | 0.5495463 | 0.5473994 | - | - | - |
| $h = 10$ | 0.5318327 | 0.5268676 | 0.4984408 | - | - |
| $h = 15$ | 0.5219245 | 0.5148890 | 0.4898930 | 0.4736951 | - |
| $h = 20$ | 0.5085644 | 0.4990497 | 0.4594693 | 0.4555957 | - |
| $h = 50$ | 0.4927631 | 0.4970696 | 0.510152 | 0.5011334 | 0.5128071 |

1.23%. FiveThirtyEight.com, the popular website for predictions, gave a prediction of 0.609, which is comparable to our Model 1 predictions.

### 4.5.2 Montana

In Montana, incumbent Democrat Senator Jon Tester ran against Republican candidate Matt Rosendale, keeping the seat with 50.33% votes compared to 46.78% for Rosendale. The prediction outcomes for Model 1 in this case are given in Table 4.3. As we can see, for many choices of $h$ and $l$, we obtain correct predictions but there are some cases where the individual predictions are wrong. The THOS forecast in this case is **0.5051**, with a relative error of 2.55%, the true value being **0.5183**. Next, we test the performance of our second model, using network structure of the Twitter data, to improve the prediction.

Table 4.4: $\hat{p}^{THN}$ Forecast for Montana Senate in US 2018 Midterm Election for different choices of h and l

| | $l = 1$ | $l = 2$ | $l = 5$ | $l = 10$ | $l = 20$ |
|---|---|---|---|---|---|
| $h = 5$ | 0.5222159 | 0.5227301 | - | - | - |
| $h = 10$ | 0.5190015 | 0.5172068 | 0.5097920 | - | - |
| $h = 15$ | 0.5201558 | 0.5196510 | 0.5018776 | 0.5261745 | - |
| $h = 20$ | 0.5195608 | 0.5191532 | 0.5019673 | 0.4899208 | - |
| $h = 50$ | 0.4950174 | 0.5031124 | 0.5090309 | 0.4859780 | 0.5123795 |

Table 4.5: $\hat{p}^{TH}$ Forecast for Arizona Senate in US 2018 Midterm Election for different choices of h and l

| | $l = 1$ | $l = 2$ | $l = 5$ | $l = 10$ | $l = 20$ |
|---|---|---|---|---|---|
| $h = 5$ | 0.4943676 | 0.4932947 | - | - | - |
| $h = 10$ | 0.4997875 | 0.4997739 | 0.5043522 | - | - |
| $h = 15$ | 0.4962760 | 0.4939905 | 0.4847742 | 0.4615345 | - |
| $h = 20$ | 0.4800426 | 0.4708316 | 0.4360083 | 0.4455570 | - |
| $h = 50$ | 0.4962980 | 0.4934741 | 0.4974655 | 0.5107796 | 0.5011974 |

Model 2 predictions are given in Table 4.4 for different choices of h and l. The averaged THANOS forecast in this case turns out to be **0.5118** with a relative error of 1.25%. In comparison, the prediction given by FiveThirtyEight.com was 0.5246. Also, in the second case, almost all choices of h and l predicts Democrats to be the winner in Montana, an improvement from our first model. Hence, incorporation of network features has led to an improvement of the forecast.

### 4.5.3 Arizona

In Arizona, incumbent republic senator Jeff Flake did not run for re-election. The Democratic candidate was Kyrsten Sinema, while the Republic candidate was Martha McSally. It was a close fought election, with Sinema winning with 50% vote, compared to McSally's 47.61%. The predictions from Model 1 and Model 2 are given given in Table 4.5 and Table 4.6, respectively.

Table 4.6: $\hat{p}^{THN}$ Forecast for Arizona Senate in US 2018 Midterm Election for different choices of h and l

| | $l = 1$ | $l = 2$ | $l = 5$ | $l = 10$ | $l = 20$ |
|---|---|---|---|---|---|
| $h = 5$ | 0.5048349 | 0.5112638 | - | - | - |
| $h = 10$ | 0.5116958 | 0.5131329 | 0.5092351 | - | - |
| $h = 15$ | 0.5080966 | 0.5109581 | 0.5111811 | 0.5246686 | - |
| $h = 20$ | 0.5073115 | 0.5055461 | 0.5021870 | 0.5474827 | - |
| $h = 50$ | 0.5033984 | 0.5035030 | 0.5150762 | 0.5121114 | 0.5071346 |

As we can see, Model (4.3.1) is not sensitive enough in this case, giving wrong predictions for many values of $h$ and $l$. The averaged THOS forecast here is **0.4867** (true value: **0.5122**), with a relative error of 4.98%. However, Model 2 gives a much better prediction, with $\bar{p}^{THN}_{h,l}$ being **0.5116**, with a relative error of 0.117%. The FiveThirtyEight.com prediction in this case was 0.5087. Hence, in this case, Model 2 is performing decidedly better than Model 1. In fact, the THANOS forecast performs better than existing election prediction algorithms as shown in Figure 4.7.

### 4.5.4 Florida

The Florida election was the closest election in the US 2018 election, with incumbent Democratic Senator Bill Nelson (49.33% votes) narrowly defeated by Republican Candidate Rick Scott (50.06% votes). Both our models gave very close predictions to the actual outcome (**0.4994**). The value of $\bar{p}^{TH}_{h,l}$ for Model 1 was **0.5035** (relative error: 0.821%), while that for Model 2 was **0.5025** (relative error 0.621%). Since every prediction has some variability associated with it, it is sensible to mark a contest Neutral if the predicted value is so close to 0.5. FiveThirtyEight.com gave a close wrong prediction, with the prediction value being 0.516 (relative error: 3.32%). Hence, our prediction method is in fact out-preforming in this particular scenario.

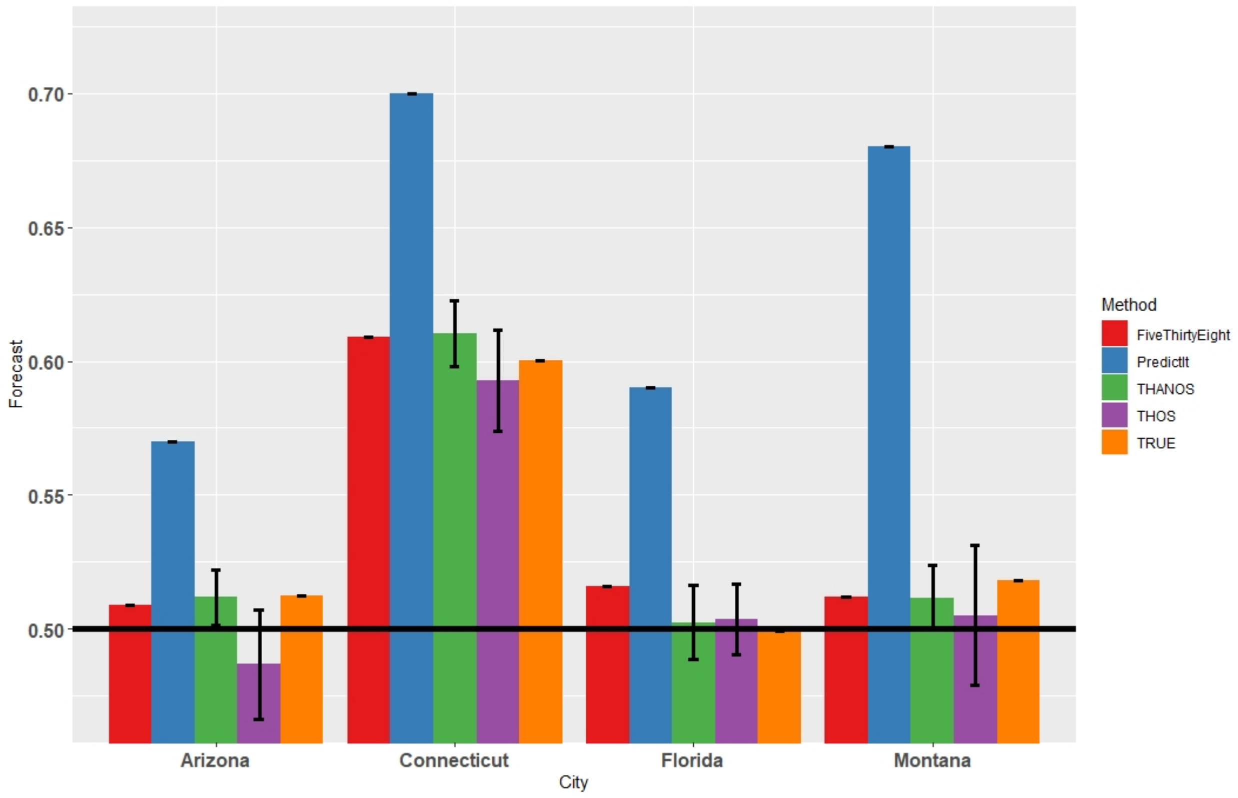


Figure 4.7: Comparison of the result of our model with other existing prediction platforms

Figure 4.7 provides a graphical illustration comparing the results of the two models with those of existing platforms, including FiveThirtyEight.com. It is evident that the THANOS forecast provides closer approximation of the final outcome, compared to existing platforms, including the FiveThirtyEight.com and PredictIt predictions.

### 4.5.5 Discussion

As we saw in the earlier section, the THOS and THANOS model both performed well in the Connecticut election, where the difference in the victory margin was quite significant. However, as the victory margin decreases, the advantage of including the network structure in the model becomes apparent, with the THANOS model performing significantly better than the THOS in the Arizona and Montana elections. Also both the models seem to perform better than existing poll based election prediction methods (cf. Figure 4.7), indicating that the addition of social media data along with polling data can lead to remarkably improved predictions.

## 4.6 Concluding Remarks

In this chapter, we have proposed two methods of forecasting elections, combining Twitter data and opinion polls conducted during the campaigns. The basic idea is that social media data can be used efficiently to improve upon traditional opinion polls, and can provide an accurate forecast of election results. The simple THOS forecast, where we only use the hashtag frequency of the Twitter data, proved to be effective for elections where the victory margin was wide (viz., Ireland vote, Connecticut Election). However, for close races with small victory margins, the THOS forecast seemed inadequate (e.g. Arizona). In this scenario, we proposed the THANOS forecast, which also incorporates network features of the Twitter dataset, in particular the centrality of the top influential persons. This second model proved to be very efficient in predicting elections, even when the victory margins were very small. For example, in Arizona election, the victory margin was as low as 2.4%, which the second model predicted with a relative error of only 0.117%, a prediction that proved to be even better than the widely acclaimed FiveThirtyEight.com prediction.

A key contribution of our predictive modeling approach is to identify network features that are critical for high accuracy of predictions using social media data. In particular, the network centrality features used in our analysis indicate that the users who occupy central or pivotal positions in a campaign, have significant influence and provide important clues to the final outcome of an election.

One potential concern with the use of social media data for forecasting purposes is whether the social media users provide a representative sample of the voting community. [142] discusses the possibility of an inherent bias among the Twitter users towards the Democratic ideology. While the issue of presence or absence of such bias among Twitter users needs further investigation, by combining Twitter data with opinion polls, our prediction approach seems to be robust against such perceived bias and is able to provide accurate predictions of the election results that are comparable to the best available competing methods based on alternative data sources including site-specific historical election data.

There is significant scope of further research in this area. For example, finding an optimum choice of the vector $\mathbf{h}$ and $\mathbf{l}$ is of theoretical interest. One can also consider adding other features of the Twitter dataset that might improve the predictive power of the model. Some such potential choices are sentiment analysis, other centrality measures, local network features, etc. To summarize the contribution at this point, we have devised a methodology to predict election outcomes by coupling opinion polls with social media data which proves to be quite accurate in the real data applications considered in this work.

# Chapter 5

# Bot Detection in Twitter

## 5.1 Introduction

The rise of social media has been one of the most critical events in this century. Social media has given the common people a platform to share their views with the public. Information is now readily available to every corner of the world. Not long after its inception, social media started to prove influential in business, politics, society, and so on. Political campaigns nowadays all over the world are heavily dependent on social media. For example, the US presidential campaign of Barrack Obama extensively used different social media platforms, like Twitter, Facebook, Myspace, etc. Inorganic accounts, also known as bots, are an integral aspect of social media. Many companies and organizations use bots to enable smooth functioning and better customer services. However, inorganic accounts can also have negative consequences. They can be used to spread fake rumors or to publicize malicious propaganda. Undisclosed bots are also a security threat to organizations, which can cause severe monetary and publicity losses. In a social network, inorganic accounts controlled by a group can create and manipulate public opinions and can play a prominent role in significant events, like elections or military conflicts. According to recent findings, about 19% of Twitter traffic is created by bots. Bots can significantly magnify some topics while downplaying other voices and influencing people to a considerable extent. Inorganic accounts often rely on the norm of reciprocity. Usually, these bots gain a significant amount of followers through the follow-refollow policy and then start producing propaganda that real users then start to retweet, relying on the source to be accurate due to the high number of followers. Often, there also exist other inorganic accounts that start the retweeting process, and when a post has been retweeted or shared many times, social media platforms tend to display them on the timeline of real users, who then start to share or retweet them, resulting in exponential growth of the reach of the original tweet. Bots that produce misinformation, or try to influence people's political ideology, can have harmful impacts on society. In this chapter, we will discuss two main bot detection methods. The first section is on static time bot detection, where we will use the tweet history of users and apply machine learning algorithms on features extracted from them to classify them as organic or inorganic accounts. The second section develops a dynamic classification algorithm, where we will rely on a tweet-based classification to obtain a bot score of users, which will give us an indication of the bot behavior of every user over time. Finally, we will briefly scrutinize the network structures developed by inorganic accounts compared to real-life networks.

Inorganic accounts or bots in social media has been extensively studied ([47], [122]). There have been multiple attempts in the literature to identify inorganic social media accounts to stop malicious propaganda campaigns as soon as possible. The BotOrNot [39] software was made public in 2014 and uses several features from a Twitter user's tweet history to classify them as bots or non-bots. [115] used Random Forest Classifier on thousands of extracted features, [87] proposed RTBust in order to use the temporal distribution of retweets for unsupervised bot classification, [37] introduced the concept of social fingerprinting for bot detection techniques. However, bot detection remains an unsolved problem [79], with a plethora of research still going on to perfect social bot detection techniques.

## 5.2 Data Description

The data is obtained from the Bot Repository at Botometer [36] The repository provides a list of identified organic and inorganic users, the tweet history of which users were scraped from Twitter through R. The dataset used in the analysis contains the tweet history of 470 verified organic users and 373 verified inorganic users. The average number of tweets by verified organic users was 3121.47, while that of inorganic users was 2598.21.

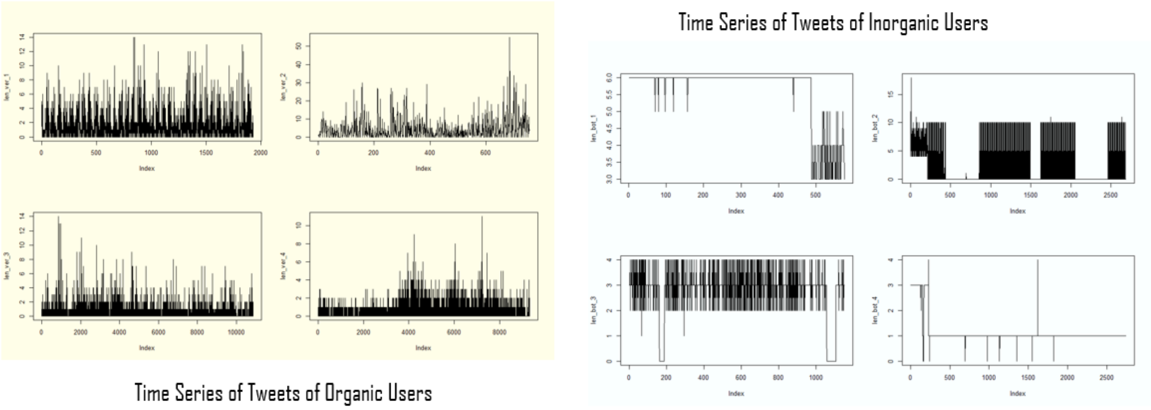


Figure 5.1: Time Series of Tweet History of Organic and Inorganic Users

Figure 5.1 gives the time series of tweet history of selected organic and inorganic users, using a granularity of three hours. The above sample demonstrates that the temporal activity of inorganic users significantly differs from that of organic users. One noticeable difference is that the inorganic users tend to be active uniformly for some fixed periods in the day, while the activity of organic users is relatively random. Hence, taking temporal features into our classification algorithm is essential. Following time-series features were included in the analysis:

- Periodicity of the Time series of tweet activity. A fundamental characteristic of time series is how frequently the observations are spaced in time, which is demonstrated by the periodicity property. For example, in this particular case, a periodicity of 8 would indicate a period of 24 hours, i.e., one day. As we can see from Figure 5.2, for most organic users, the periodicity was 8 or 4, indicating a daily or 12-hour recurring activity. However, for inorganic users, the periodicity was relatively low, indicating that the activity pattern drastically differed from the inorganic users.

- We also fitted the data to an ARIMA (Autoregressive Integrated Moving Average) process and used the fit features in our classification model. In particular, the features taken from the ARIMA fit were: log-likelihood, sum square of coefficients, error variance, and fit length.

- Finally, we also included features to capture the shape of the periodogram. Periodicity only identifies the peak of the periodogram. However, we might also be interested in the shape of the periodogram, and hence we used local maxima to incorporate those properties.

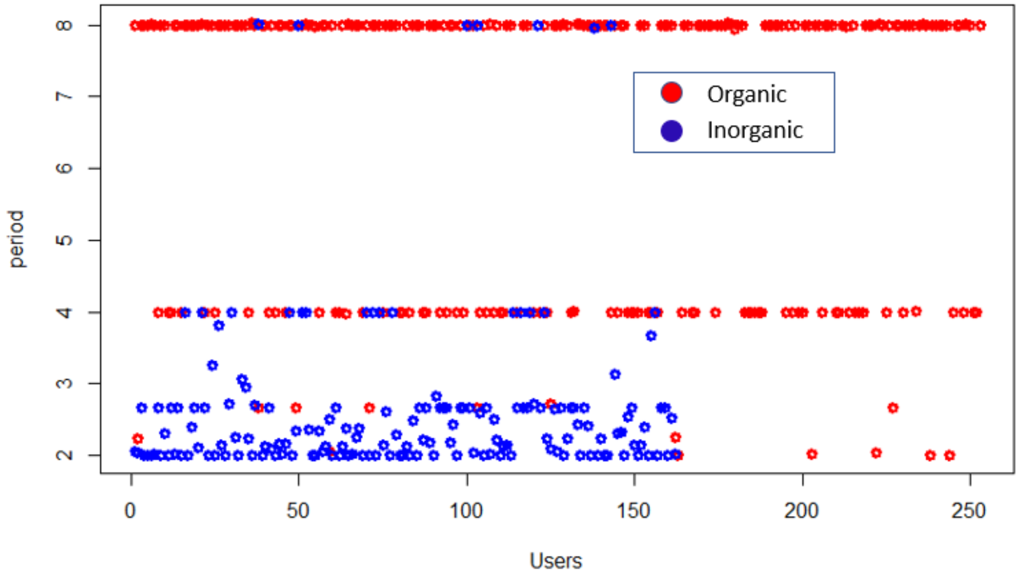


Figure 5.2: Periodicity of Tweet Activity of Organic and Inorganic accounts

Apart from the temporal features mentioned above, we also included semantic features from the tweet history. On inspection of the tweets, we found that apart from the significant difference in the time series, there also appears to be a stark contrast in words used by the organic and inorganic users. Inorganic users appear to use similar words repetitively, a trait that was expectantly absent in the tweeting pattern of organic users. Following are the semantic features used in the classification algorithm:

- Average number of words per tweet and it's variance
- Number of unique words used (see Figure 5.3)
- Relative frequency of most used words
- Relative Hashtag Frequency
- Sentiment Score (the sentiment score is calculated using the three dictionaries available at R, namely, afinn, bing and nrc).

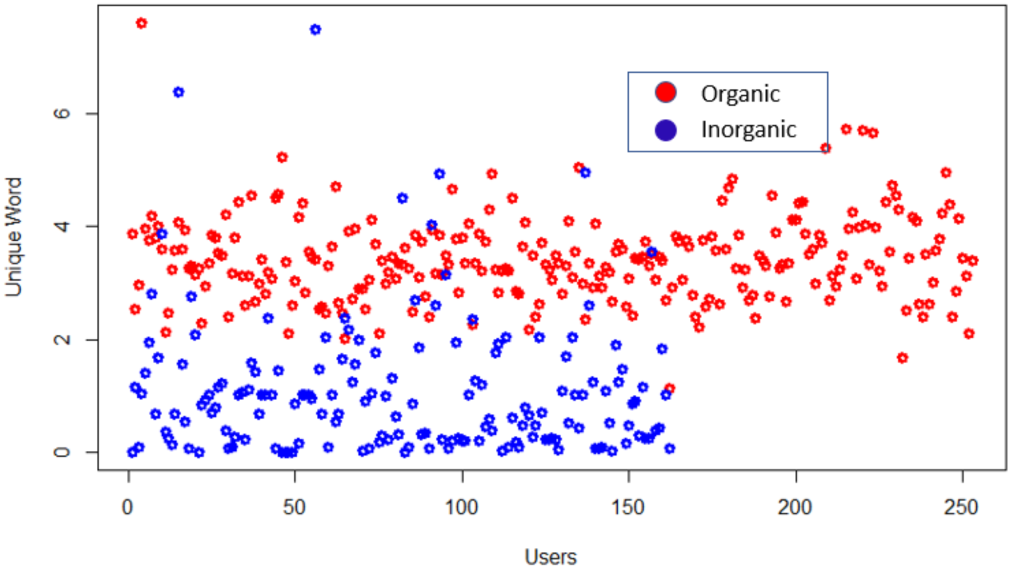
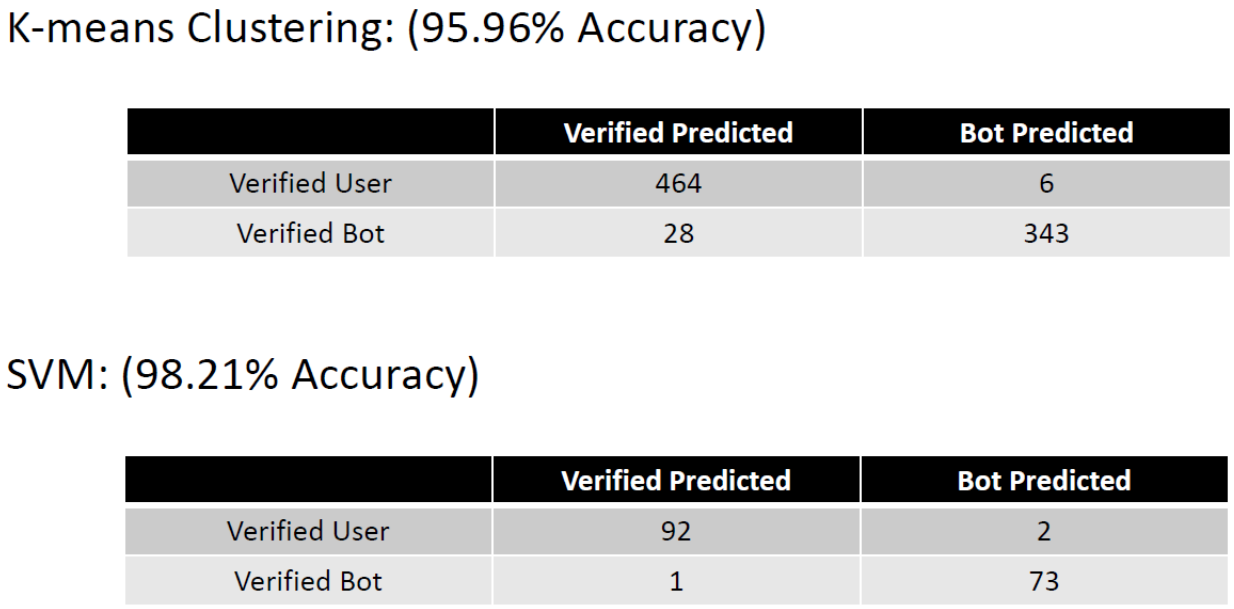


Figure 5.3: Number of Unique Words used by Organic and Inorganic Accounts

## 5.3 Methodology

The features described above contain 19 features, a mixture of temporal and semantic features. However, one or more of these features might be linearly related, resulting in multicollinearity in the data, which is a nuisance in the classification problem. Hence, our classification algorithm uses VIF (Variance Inflation Factor) to select 11 of these features. We have used three main classification algorithms: k-means clustering, SVM, and logistic regression-based clustering. K-means clustering is an unsupervised classification algorithm that aims to cluster a given data into partitions of a given number of clusters based on the features provided. The SVM (or Support Vector Machine) is a supervised algorithm that creates a classifier based on a given training data, which can then be applied to get the classification in the test dataset. Finally, the logistic regression-based algorithm involves fitting a logistic regression model to our data, the dependent variable being the labels of the training dataset and the independent variables being the extracted features.

The prediction of the test dataset would then give a score between 0 and 1, which will be clustered into two or more partitions. The output of the k-means clustering and SVM are given in Figure 5.4.



K-means Clustering: (95.96% Accuracy)

| | Verified Predicted | Bot Predicted |
|---|---|---|
| Verified User | 464 | 6 |
| Verified Bot | 28 | 343 |

SVM: (98.21% Accuracy)

| | Verified Predicted | Bot Predicted |
|---|---|---|
| Verified User | 92 | 2 |
| Verified Bot | 1 | 73 |

Figure 5.4: Clustering Accuracy

We can see that both k-means and SVM provided moderately well classification accuracy. SVM performs marginally better since it is a supervised machine learning algorithm compared to k-means clustering. The Logistic regression-based clustering also gave around 97% accuracy.

Now that we have achieved significant classification accuracy, it might be interesting to find out the importance of the features provided to the clustering algorithm. There are multiple possible ways to measure the importance of the features. Here, we have defined an Accuracy Score (AS) for the features:

$$AS = 100 \times \frac{Accuracy - Accuracy_i}{\sum_i (Accuracy - Accuracy_i)} \tag{5.3.1}$$

where $Accuracy_i$ is the accuracy of the classification algorithm when the $i^{th}$ feature is removed.

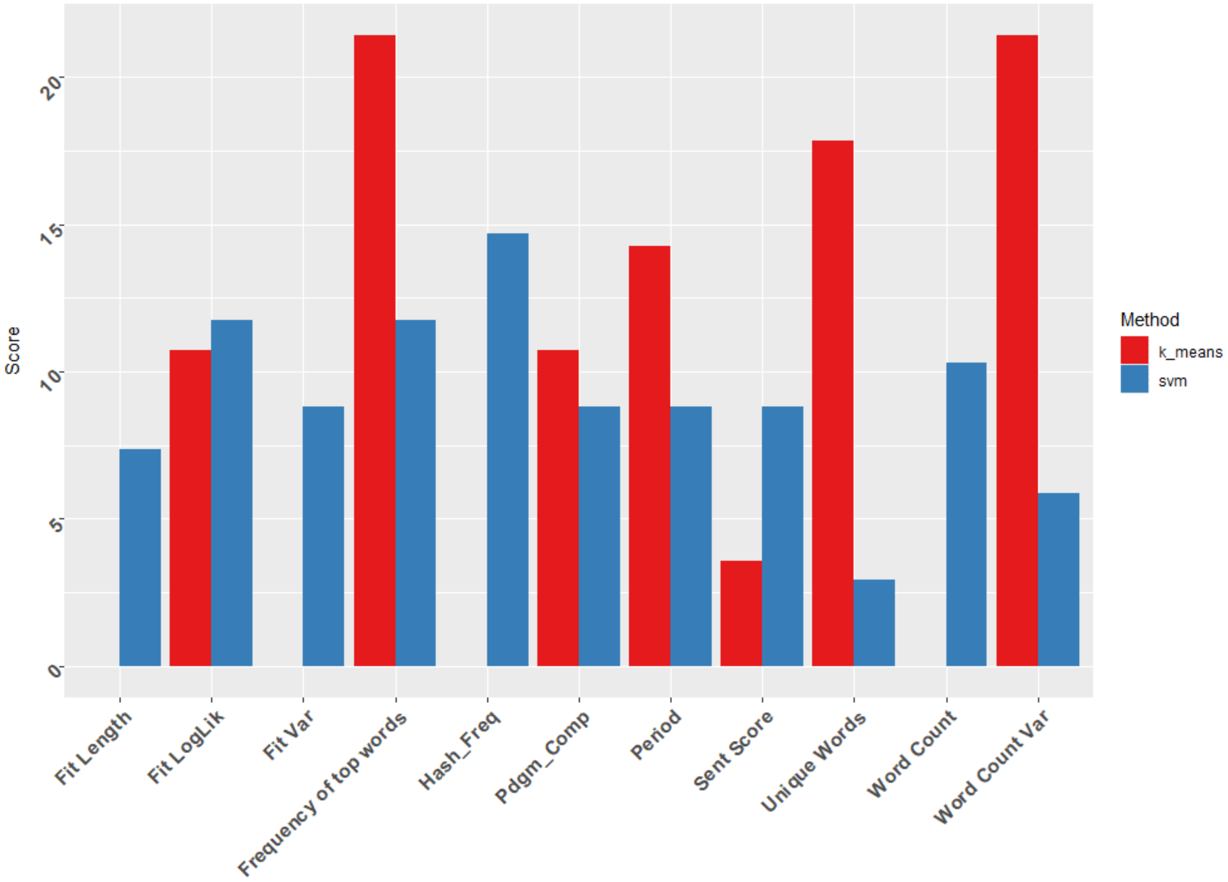


Figure 5.5: Accuracy Score of Features

Figure 5.5 gives the accuracy score of the features under the two classification algorithms. As we can see, for k-means, the relative frequency of words and word count variance are the most significant features, while for SVM, the Hashtag and word count are the most important.

We can also assess the importance of the features by other machine learning techniques, for example, Bagging and Boosting or LASSO. Figure 5.6 gives the output for these two methods.

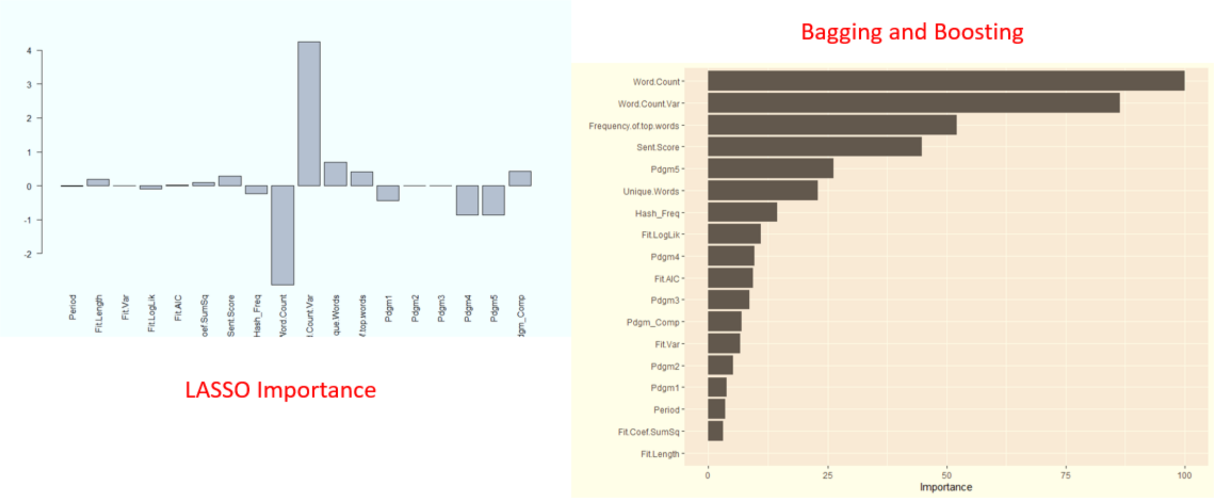


Figure 5.6: LASSO and Bagging And Boosting

## 5.4 Dynamic Time Bot Detection

The bot detection algorithm discussed above uses a static time algorithm, meaning that we take a user's tweet history at a particular time point and use that entire history to classify them as an organic or inorganic account. However, it is of significant importance to identify these inorganic accounts as soon as possible in order to deactivate the accounts and stop them from propagating malicious propaganda. Additionally, it is often noticed that the same account might not always behave like an inorganic account. In other words, "bot behavior" might be observed in an account for a specific period of time, so it is important to identify that account in that window and deactivate it. In order to do that, we have developed a dynamic algorithm that uses tweet-specific information of a group of users and assigns them a "bot" score, and we update them over time. The algorithm is given in Algorithm 1.

The algorithm provides a time series of "bot scores" for every user. The higher the score, the higher the probability of an account being inorganic or paid. As we saw earlier, inorganic accounts have a different way of tweeting, like using similar words a lot, using the same tweets over and over again, etc. Using these distinctive features, we hope to identify the inorganic accounts as soon as possible. We also update the dictionary every 30 days. All users with a score higher than 0.7 are classified as bots. All users with a score lower than 0.3 are classified as verified accounts, and the dictionary is updated accordingly.

We also used two scores: Score 1: $y = \frac{p-\bar{p}}{\sqrt{n_{test}}}$ and Score 2: $y = \frac{p-0.5}{\sqrt{n_{test}}}$. Here $\bar{p}$ is the average score of all users in the test set, and $n_{test}$ is the number of users in the test set. It is of importance that the two scores give different measures. For instance, the first score measures the deviation of the score of a tweet from the sample mean, i.e., how likely is the tweet coming from a bot based on the sample we have in hand, while the second score measures the deviation from a pre-specified score of 0.5. Figure 5.7 and 5.8 gives the time series of bot scores of three known bots and three verified users for the two scores. As we can see, the score gradually decreases to zero for verified users, while for the bots, the score increases gradually over time. It is also noticeable that for Score 2, the pattern is steeper compared to score 1. It would be important to analyze the effect of different scores on this algorithm.

```
Algorithm 1: Dynamic Time Bot Detection
Data: Tweet Activity of several users, inorganic or organic over time
Result: A time series of “Bot Score” of a user that tracks the “botness” of the user
        over time
score ← 0.5;
$N$ ← Total Remaining Time Period;
Initialization:
    • Start with a base history of tweets, already labeled as 0 or 1. The base history is the
      tweets of known organic and inorganic users, the users having tweeted at least 20
      tweets.
    • Obtain following features:
        – Length of Tweet
        – Number of Words Used
        – Relative Frequency of most used Word.
        – Number of hashtags
    • Fit a logistic regression model using the features, the dependent variable being the
      binary class.
while $N \neq 0$ do
    • Take the first 5 days, and provide a prediction (say, $bot_p$) of ”BOT” score for all the
      tweets using features obtained in this period. Set $N = N - 5$.
    • Higher this prediction, higher would be the chance of the tweet coming from an
      inorganic user.
    • Create a scaled variable, say $score_p$ for every tweet (e.g. $\frac{bot_p - \bar{bot}_p}{\sqrt{n}}$, where the average
      is over all predictions obtained in that 5-day period, and $n$ is the total number of
      tweets considered in that period).
    • Take the mean of the above score of all tweets by an user to be the score of an user
      in the time period.
    • Add the score for that user to that of previous 5 days (“score” defined in the
      beginning).
end
```

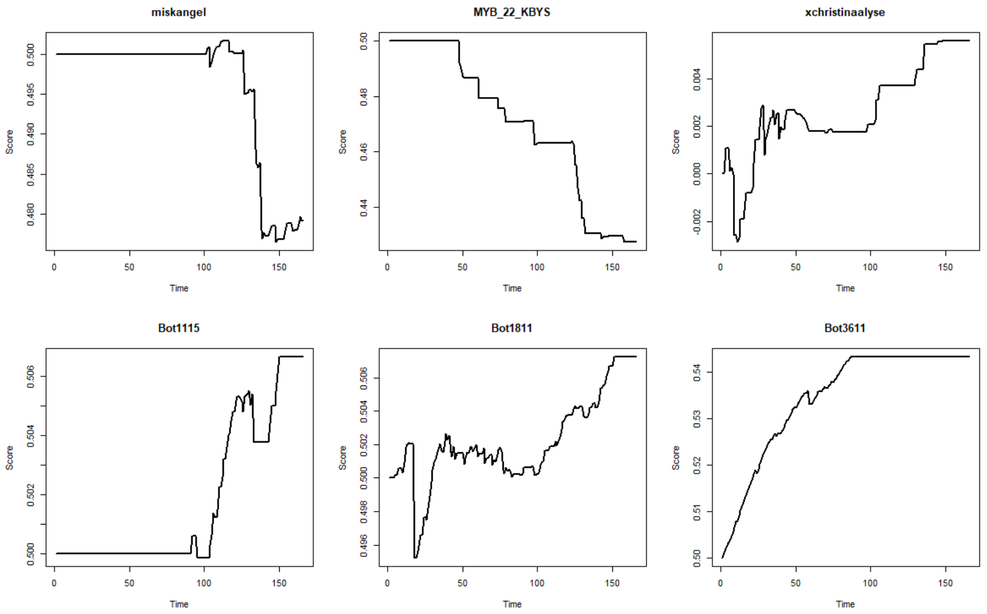


Figure 5.7: Bot Score Of Six Users for Score 1

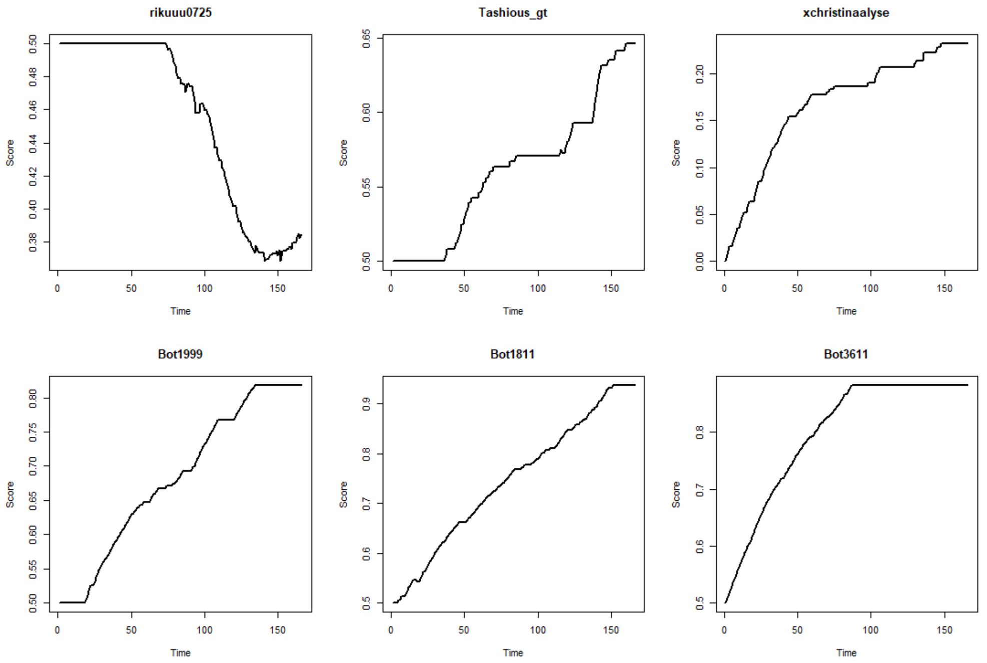


Figure 5.8: Bot Score Of Six Users for Score 2

## 5.5 Conclusion

We have discussed two main types of bot detection, static and dynamic. While static bot detection is more reliable since it uses many tweets from every user, it is time-consuming and needs at least some tweet activity from the user before it can identify them as bots or non-bots. On the other hand, the dynamic tweet classification creates a time series of bot scores, which will indicate the bot behavior of a user from their very first tweet. We can then devise a threshold, above or below which we can classify the user as a bot or non-bot. Future works will include devising a score indicating how many inorganic accounts have infiltrated a network. Also another approach would be to use graph partitioning to identify different sources of these inorganic account activities so that the social media platform can take corresponding actions. We are also working on a project to analyze to what extent these inorganic bots in social media influence political campaigns worldwide.

# Chapter 6

# Rate of convergence and optimal choice of $m$ in the $m$ out of $n$ Bootstrap for sample extremes

## 6.1 Introduction

Theory of Extreme Value Distributions is quite popular in many real-life applications, where the events of interest involve extremes and/or rare events. Some such applications include financial risk modeling ([110]), weather anomalies like earthquake problems ([72], [109]), Environmental Sciences ([113]) and so on. These widespread applications generate complicated analytical questions on unknown parameters of the underlying random process. The Bootstrap and its variants provide a computer-intensive simple mechanism to answer these questions. However, it is well-known that Efron's ([45]) Bootstrap with a resample size $m$ that is equal to the sample size $n$ fails drastically for the sample extremes (cf. [13], [7]). A remedy is given by the $m$ out of $n$ Bootstrap ([14], [50]). The choice of $m$ is of significant interest in the $m$ out $n$ Bootstrap as it determines the accuracy of the resulting Bootstrap approximation. Although several authors have studied the problem of choosing the resample size $m$ and put forward various methods for its choice ([14], [15]), a definitive answer to the issue of optimal choice of $m$ has been elusive. In this chapter, we first derive rates of convergence results

This chapter proposes an optimal block size m as an order of the sample size n, in univariate i.i.d. framework, in order to minimize the convergence rate of the Bootstrap distribution to the actual distribution of the extremes. Additionally, we will give a small extension of our work in univariate framework to multivariate set-up. In summary, this chapter provides the convergence rate of non-parametric Bootstrap, proposes an optimal Bootstrap sample size, and gives a small extension of the work to multivariate framework. Section 6.2 is devoted to literature review in the field under consideration. The main results of this chapter has been segregated into two different sections. Section 6.3 deals with the $m$ out of $n$ non-parametric Bootstrap, it's rate of convergence under known (Section 6.3.1) and unknown (Section 6.3.2) normalizing constants, and a proposal for an optimal block size (Section 6.3.3) as a function of the sample size n. In Section 6.4 will provide a comparison of the performance of $m$ out of $n$ Bootstrap with the traditional approach of fitting by limit distribution. Section 6.6 gives the simulation results, and Section 6.9 contains the proof of the main theorem of the chapter. Finally, Appendix 6.10 will contain further discussions, and proofs of some results given in the chapter.

## 6.2 Background

[55] first proposed that the extremes of univariate independent random variables can converge to only three types of distributions, under suitable standardization. In particular, suppose we have $n$ independent and identically distributed random variables $X_1, \ldots, X_n$ generated from some underlying distribution function $F$. Then there exists $b_n > 0$ and $a_n \in \mathbb{R}$, such that the limiting distribution of $P\left(b_n^{-1}\left(X_{n:n} - a_n\right) \leq x\right) = F^n(a_n + b_n x)$ is one of the three types: Type I or Gumbel ($\Lambda(x) = exp(-e^{-x}), x \in \mathbb{R}$), Type II or Fréchet ($\Phi_\alpha(x) = exp(-x^{-\alpha})\mathbb{I}(x > 0)$ for some $\alpha > 0$) and Type III or Weibull ($\Psi_\alpha(x) = exp(-(-x)^\alpha)\mathbb{I}(x \leq 0)$ for some $\alpha > 0$). Here $X_{i:n}$ denotes the $i^{th}$ order statistics from a sample of size $n$. The three types of extreme valued distribution can also be written in a compact form, $G_\gamma(x) = exp\left\{-\left(1 + \gamma x\right)^{-\frac{1}{\gamma}}\right\}$ if $\gamma \neq 0$, and $exp(-exp(-x))$ if $\gamma = 0$. The $\gamma = 0$ case provides the Type I or Gumbel limit law distribution, while $\gamma > 0$ and $\gamma < 0$ indicates the Frechet and Weibull limit laws respectively. It is to be noted that the $\alpha$ in $\Phi_\alpha(x)$ and $\Psi_\alpha(x)$ are related to $\gamma$ through the relation $\gamma = \frac{1}{\alpha}$ and $\gamma = -\frac{1}{\alpha}$ respectively. Mathematically, we can write Gnedenko's theorem as:

$$P(b_n^{-1}(X_{n:n} - a_n) \leq x) = F^n(a_n + b_n x) \to G_\gamma(x) \tag{6.2.1}$$

This central limit theorem like result helps in attaining asymptotic properties of statistical methodologies concerned with extremes of random variables. Some such examples of applications include Financial Risk Modeling ([54]), Public Health ([129]), Reliability Theory ([58]), Environmental Data Analysis ([131]), etc. There are well established works on determining the normalizing constants and the limit law of a distribution function. Suppose $\omega(F) = sup\{x : F(x) < 1\}$. The following guidelines for determining the limit law of the maximum of random variables is by [52].

- $F \in D(\Lambda)$ iff the following happens:

  Suppose, for some finite $a$,

$$\int_{a}^{\omega(F)} (1 - F(y))dy < +\infty$$

  For $\alpha(F) < t < \omega(F)$, define:

$$R(t) = (1 - F(t))^{-1} \int_{t}^{\omega(F)} (1 - F(y))dy$$

  Then, we must have:

$$\lim_{t \to \infty} \frac{1 - F(t + xR(t))}{1 - F(t)} = e^{-x} \tag{6.2.2}$$

- $F \in D(\Phi_\alpha)$ iff $\omega(F) = +\infty$ and there exists constant $\gamma > 0$ such that, for all $x > 0$, as $t \to +\infty$,

$$\lim_{t \to \infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-\gamma} \tag{6.2.3}$$

- $F \in D(\Psi_\alpha)$ iff $\omega(F) < \infty$ and there exists constant $\gamma > 0$ such that, for all $x > 0$, as $t \to +\infty$,

$$\lim_{t \to \infty} \frac{1 - F^*(tx)}{1 - F^*(t)} = x^{-\gamma} \tag{6.2.4}$$

  where $F^*(x) = F(\omega(F) - \frac{1}{x})$

For example, the Normal distribution belongs to $D(\Lambda)$, the Cauchy distribution belongs to $D(\Phi_\alpha)$ and the Uniform distribution belongs to $D(\Psi_\alpha)$. The normalizing constants $a_n$ and $b_n$ depend on the underlying distribution of the random variables, and are unique to an extent of perturbation . In particular, if the the maximum of random variables belong in the domain of attraction of $G_\gamma(x)$ for some choice of normalizing constants $a_n$ and $b_n$, then we can also choose $a'_n$ and $b'_n$ such that:

$$\lim_{n\to\infty} \frac{a'_n - a_n}{b_n} = 0 \qquad \text{and} \qquad \lim_{n\to\infty} \frac{b'_n}{b_n} = 1 \tag{6.2.5}$$

[55] proposed a choice for the normalizing constants for three types of distributions:

- **Type I:** $a_n = \gamma_n$ and $b_n = F^{-1}(1 - \frac{1}{en}) - \gamma_n$.
- **Type II:** $a_n = 0$ and $b_n = \gamma_n$.
- **Type III:** $a_n = \omega(F)$ and $b_n = \omega(F) - \gamma_n$.

where $\gamma_n = F^{-1}(1 - \frac{1}{n})$. Hence, if we know the underlying distribution, we can approximate the distribution of the standardized maximum of random variables generated from the distribution, and can provide significant insights into the behaviour of the extremes (known as Extreme Value Analysis). For example, if we consider the Air Quality Index (AQI) of Mumbai, under known normalizing constants, we can provide the probability of the maximum of the index over a period of time crossing a certain threshold. However, in real life, these normalizing constants are unknown, and hence needs to be estimated from the data. Multiple methods have been proposed, with some working better than others in some scenario. [71] in his book discusses three such methods, regression estimate, minimum variance unbiased estimate, and maximum likelihood estimates. [30] proposes a simple estimate based on moments. However, all these methods is reliant on estimation of $\gamma$, the tail index, when it is non-zero. [40] gives a detailed description of several methods of obtaining an estimate for the tail index, under the second-order condition $\lim_{t\to\infty} \frac{\log U(tx) - \log U(tx) - \gamma \log x}{b(t)} = \frac{x^\rho - 1}{\rho}$, where $\rho \leq 0$, $U = \left(\frac{1}{1-F}\right)^{-1}$, and $\gamma$ the tail-index. The popular methods for estimating $\gamma$ include the Hill estimator ([62], $\gamma > 0$), Pickland Estimator ([100], $\gamma \in \mathbb{R}$), Maximum Likelihood Estimator ($\gamma > -\frac{1}{2}$), Moment Estimator ($\gamma \in \mathbb{R}$), and so on. However, all these estimator are $\sqrt{k(n)}$ consistent, i.e. $\sqrt{k(n)}(\hat{\gamma}-\gamma)$ follows a normal distribution, where $k(n)$ is such that $k(n) \to \infty$ and $\frac{k(n)}{n} \to 0$. An optimal choice of $k(n)$, as given in [40], is $n^{-\frac{2\rho}{1-2\rho}}$, with suitable constants, and hence the convergence rate is always slower than $\sqrt{\frac{1}{n}}$. When $\gamma = 0$, i.e. for Type III limiting law distributions, the moment estimators of $a_n$ and $b_n$ are $\sqrt{n}$ consistent. Hence, in all three cases, estimation of the normalizing constants would always have a convergence rate slower than $\sqrt{\frac{1}{n}}$.

In order to approximate the distribution of maximum of random variables, one would also require the rate of convergence of the standardized extremes to the corresponding limiting law. [41] determined the rate of convergence of the maximum of random variables from distribution functions satisfying second order Von-Mises condition, and second order regular variation condition.

**Definition 6.2.1.** A function $f:(0,\infty)\to\mathbb{R}$ satisfies a second order Von-Mises condition with first order parameter $\gamma\in\mathbb{R}$ and second order parameter $\rho\leq 0$, in other words, $f\in$ 2-von Mises$(\gamma,\rho)$ if $f$ is twice differentiable, $f'$ is eventually positive and the function $A(t)=\frac{tf''(t)}{f'(t)}-\gamma+1$ has constant sign near infinity and satisfies $\lim_{t\to\infty}A(t)=0$ and $|A|\in RV_\rho$.

**Definition 6.2.2.** Second-Order Regular Variation Condition:

- $\frac{f'(tx)/f'(t)-x^{\gamma-1}}{A(t)} \to x^{\gamma-1}(\frac{x^\rho-1}{\rho}) = K'_\gamma(x)$
- If $\gamma\geq 0$, $\frac{[f(tx)-f(t)]/tf'(t)-(x^\gamma-1)/\gamma}{A(t)} \to \int_1^x u^{\gamma-1}(\frac{u^\rho-1}{\rho})du = K_\gamma(x)$

Suppose

$$f=(\frac{1}{-\log(F)})^{\leftarrow}, \qquad h=\frac{1}{-\log F}, \qquad S=-\log(-\log F) \tag{6.2.6}$$

Consider the Uniform Distance $(d_n)$ and the Total Variance Distance $D_n$, defined by $d_n = \underset{x\in\mathbb{R}}{sup}|F^n(a_nx+b_n)-G_\gamma(x)|$ and $D_n = \underset{A\in\mathcal{B}(\mathbb{R})}{sup}|P(a_n^{-1}(M_n-b_n)\in A)-G(A)| = \frac{1}{2}\int_\infty^\infty\left|\frac{d}{dx}F^n(a_nx+\right.$

$b_n) - G'(x)\Big| dx$. If $f$ satisfies second order Von-Mises Condition, and second order regular variation conditions, [41] showed that $A(t) = (\frac{1}{S'})'(f(t)) - \gamma$ and both uniform and total variation metric converges to 0 at a rate of $O(A(n))$. Hence, computation of $A(n)$ is of significant interest. It can be shown that for absolutely continuous distributions,

$$A(t) = \frac{c(t)^2 \left(1 + f'(t)\right) - c''(t) f(t)^2}{\left(c(t) + c'(t) f(t)\right)^2} \frac{e^{-\frac{1}{t}}}{t\left(1 - e^{-\frac{1}{t}}\right)} - (1+\gamma) + \frac{1}{t} \tag{6.2.7}$$

where,

$$1 - F(t) = c(t) exp\left(-\int_{t_0}^{t} \frac{ds}{f(s)}\right) \tag{6.2.8}$$

Any distribution function in the domain of convergence of $G_\gamma$ for any $\gamma \in \mathbb{R}$ must be of the form given in equation 6.2.8 ([42]), with:

- $\lim_{t \to \infty} \frac{f(t)}{t} = \gamma$ for $\gamma > 0$
- $\lim_{t \to x^*} \frac{f(t)}{(x^* - t)} = -\gamma$ for $\gamma < 0$
- $f'(t) \to 0$ and $f(t) \to 0$ if $x^* < \infty$ for $\gamma = 0$.

One possible choice of the function $f(t)$ is $\frac{1-F(t)}{F'(t)}$, or the Mill's Ratio. The rate of decay of $A(n)$ is the same as the rate at which $\frac{1-F}{F'}$ and it's derivative converges. This rate can be as fast as $\frac{1}{n}$ (eg. Exponential, Pareto, etc.) or as slow as $\frac{1}{\log n}$ (Normal, Skew Normal, etc). Hence, estimating the extreme distribution using the limit law involves complicated procedures, and the convergence rate is always slower than $\frac{1}{\sqrt{n}}$.

Another possible alternative method of estimating the distribution of the extreme is using Bootstrap, which uses resampling procedure to provide estimates of higher order statistics. [45] in 1979 introduced the Bootstrap method to estimate the sampling distribution of statistics. Suppose, $X_1, \ldots . X_n$ be iid random variables with cdf $F$, and $\chi_n = (X_1, \ldots, X_n)$.

Suppose we wish to estimate the sampling distribution of a specified random variable $R(\chi_n, F)$. Then, we first need to construct empirical distribution function from the available data. With the empirical distribution function $F_n$ fixed, we have to draw a random sample of size $m = n$ with replacement from $F_n$, i.e. $X_i^* \sim F_n$, also known as the Bootstrap sample, $\chi_n^* = (X_1, \dots, X_n)$. The final step is to approximate the sampling distribution of $R(\chi_n, F)$ by the conditional distribution of $R(\chi_n^*, F_n)$. [45] considered the Bootstrap sample size to be the same as the original sample size. [13] first proposed the possibility of taking $m$ out of $n$ samples, i.e choosing m Bootstrap samples from the empirical distribution $F_n$. [13] showed that the Efron's Bootstrap is inconsistent for maximum of i.i.d. uniform random variables. [6] showed that the Bootstrap distribution of the maximum converges in distribution to a random distribution, the form of which was given by [50]. However, [50] showed that $m$ out of $n$ Bootstrap is consistent for extreme valued distribution, when $m = o(n)$, i.e. $\frac{m}{n} \to 0$.

Given $\underline{X}_n = (X_1, X_2, \dots X_n)$, let $X_1^*, X_2^*, \dots, X_m^*$ be conditionally i.i.d. random variables with the distribution:

$$P(X_1^* = X_j | \underline{X}_n) = \frac{1}{n}, \quad j = 1, 2, \dots, n$$

i.e. a random sample generated from the empirical distribution of X. Now, let us define:

$$G_n(x) = P\{b_n^{-1}(X_{n:n} - a_n) \leq x\} \tag{6.2.9}$$

$$H_{n,m}(x, \omega) = P\{b_m^{-1}(X_{m:m}^* - a_m) \leq x | \underline{X}_n\} \tag{6.2.10}$$

$H_{n,m}(x, \omega)$ is the Bootstrap distribution of $b_n^{-1}(X_{n:n} - a_n)$, n and m are called the sample size and resample size respectively. The convergence of $F^n(a_n + b_n x)$ to $G_\gamma(x)$ is equivalent to $n\{1 - F(a_n + b_n x)\} \to c(x) \equiv -\log G_\gamma(x)$ for each $x \in C_G$ . Suppose

$$T_{n,m}(A, \omega) = \#\{j : 1 \leq j \leq n, b_m^{-1}(X_j - a_m) \in A\} \tag{6.2.11}$$

It is easy to see that: $T_{n,m}((x, \infty), \omega) = n\{1 - F_n(a_m + b_m x)\}$. [50] proved that if

$$P\left(b_n^{-1}(X_{n:n} - a_n) \leq x\right) \to G(x)$$

for each $x \in C_G$, and G is non-degenerate, then

$$\sup_{x \in \mathbb{R}^d} |H_{n,m}(x, \cdot) - G(x)| \xrightarrow{p} 0 \tag{6.2.12}$$

if $m = o(n)$. In addition, if $\sum_{n=1}^{\infty} \lambda^{\frac{n}{m}} < \infty$ for each $\lambda \in (0, 1)$, the convergence is with probability 1.

Even though Bootstrap provides a simple method of approximating the extreme of random variables, the convergence rate depends significantly on the choice of the Bootstrap sample size $m$. In this work, we will derive the convergence rate of $m$ out of $n$ Bootstrap distribution of a sample $X_1, \ldots, X_n$, and propose an optimal choice of sample size. We will show that the convergence rate is of the order $n^{-\frac{1}{3}}$ under optimal choice of $m$, which in some cases are better than the actual convergence rate, even for known normalizing constants. Under unknown normalizing constants, the convergence rate of the actual normalized extremes is always slower than $\sqrt{\frac{1}{n}}$, with bias often arising with wrong choice of estimation method. These hindrances can be easily overcome by applying Bootstrap, which will always give a convergence rate of the order of $n^{-\frac{1}{3}}$. A convergence rate in multivariate set-up is also provided, under certain assumptions on the dependence structure. Simulation study is conducted on three distributions belonging to one of each three types of extreme value distribution, i.e. Gumbel, Frechet and Weibull. The three distributions were chosen such that the unknown parameters of the distribution are also embedded in the parameters of the limiting extreme value distributions, after normalizing with suitable constants $a_n$ and $b_n$. The three chosen distributions are:

- **Type I (Gumbel):** SN($\alpha$) (SN implies Skewed Normal).
- **Type II (Frechet):** Pareto($\alpha$).
- **Type III (Weibull):** Beta(1,$\beta$).

As shown in [41], to determine the rate of convergence for the extreme of these random variables, it is enough to compute $A(n)$, where $A(t) = \frac{tf''(t)}{f'(t)} - \gamma + 1$. The $A(n)$ for the three distributions under discussions are as follows:

- $SN(\alpha) : O(\frac{1}{\log n})$
- $Pareto(\alpha) : O(n^{-1})$
- $Beta(1, \beta) : O(n^{-1})$

In the subsequent sections, we will analyze the rates of convergence for these distributions, and compare the rates under different choices of Bootstrap sample size.

## 6.3 Non-Parametric Bootstrap

The practice of using $m$ out of $n$ Bootstrap is particularly popular in extreme valued random variables, due to the inconsistency of Efron's Bootstrap in these scenario. [50] showed that non-parametric Bootstrap works for extremes of random variables when $m = o(n)$.

However, using such low Bootstrap sample size can have detrimental effect to the rate of convergence. Hence, it is of particular importance to find how well the $m$ out of $n$ Bootstrap performs, in other words the rate of convergence of non parametric Bootstrap in extreme valued random variables. Another important question in this area would be determining an optimal Bootstrap sample size, in order to minimize the rate of convergence, which has been discussed in Section 6.3.3.

### 6.3.1 Known Normalizing Constants

One important aspect of limiting distribution of extreme valued random variables is the choice of normalizing constants $a_n$ and $b_n$. Theorem 6.3.1 gives the convergence rate for non-parametric Bootstrap for extreme valued random variables under known normalizing constants.

**Theorem 6.3.1.** *Suppose (6.2.1) holds. Then, for the non-parametric Bootstrap case, we can write:*

$$F_n^m(a_m + b_m x) - H(x) = \zeta_1(x,n,m) + \zeta_2(x,n,m) + \zeta_3(x,n,m) \tag{6.3.1}$$

*where, $\zeta_1(x,n,m)$ is $O\left(\frac{1}{m}\right)$, $\zeta_2(x,n,m)$ is $O_p\left(\sqrt{\frac{m}{n}}\right)$, and $\zeta_3(x,m)$ is $O\left(A(m)\right)$ under second-order regular variation condition. If we are only concerned with the rate of convergence to the true distribution, then we can only concentrate on $\zeta_1$ and $\zeta_2$, in which case the rate of convergence would be $O_p\left(\max\left(\frac{1}{m}, \sqrt{\frac{m}{n}}\right)\right)$. Additionally, $E\zeta_2^2 = O\left(\frac{m}{n}\right)$.*

The detailed proof of Theorem 6.3.1 can be found in Section 6.9. Here we will discuss an overview of the proof and the potential implications. The expression $F_n^m(a_m + b_m x) - H(x)$ can be broken down into 3 parts, as given below:

$$F_n^m(a_m + b_m x) - H(x) = \underbrace{\Xi(z_m) + \Xi'(z_m)(z_m^* - z_m) + R_2^{\Xi}(z_m)}_{\zeta_1^*(x,m,n)} + \underbrace{e^{-z_m^*} - e^{-z_m}}_{\zeta_2^*(x,m,n)} + \underbrace{e^{-z_m} - H(x)}_{\zeta_3(x,m,n)} \tag{6.3.2}$$

In the above equation, $\Xi(z_m)$ is given by $e^{-z_m^*}R_2^g(x)$ where, $z_m^* = m(1 - F_n(a_m + b_m x))$ and $R_2^g(x)$ is a remainder term of a Taylor series expansion (see proof for details). Calculations reveal that the first term is of the order $\frac{1}{m}$, while [41] has proved that the last term goes to 0 at a rate of A(m), as given in equation 6.2.7. Finally, the second part can be proven to be of the order $\frac{m}{n}$ (see Proof). In particular,

$$P(|z_m^* - z_m| > t) \leq \frac{mc(x)}{nt^2} \tag{6.3.3}$$

which in turn gives $|z_m^* - z_m| = O_p(\sqrt{\frac{m}{n}})$. In the coefficient term, $c(x)$ is the limit of $mp_m$, which according to Lemma 2.1, is $-\log(G(x))$ where $G(x)$ is the limiting distribution. The coefficient term $c(x)$ will be of particular importance in determining the optimal resample size. It can also be further proved that $E(\hat{z}_n - z_n)^2$ is of the order $O(\frac{m}{n})$, i.e.

$$E(\hat{z}_n - z_n)^2 = O(\frac{m}{n}) \tag{6.3.4}$$

Hence, we also have a $L_2$ convergence rate for non-parametric Bootstrap in extreme valued random variables.

Thus, we have seen that while $\zeta_1^*(x, m, n)$ is of the order $\frac{1}{m}$, $\zeta_2^*$ and $\zeta_3^*$ are of the order $O_p(\sqrt{\frac{m}{n}})$ and $A(m)$ respectively. In practice, only the rate of convergence of the Bootstrap distribution to the true value is of real interest , i.e. the rate of convergence of $H_{n,m}(x)$ to

$G_n(x)$ This rate of convergence can be given by $O_p\left(\max\left(\frac{1}{m}, \sqrt{\frac{m}{n}}\right)\right)$, which would always be slower than $n^{-\frac{1}{3}}$, the upper bound being attained under the optimal choice of sample size, as given in Section 6.3.3. Furthermore, since this rate of convergence does not require any distributional properties, the rate of convergence would be same for all three limit distributions, differing only in the coefficients.

*Remark* 6.3.1. Thus, we can see that for non-parametric case $F_n^m(a_m + b_m x)$ converges to $H(x)$ at a rate of $\max\left(\frac{1}{m}, \sqrt{\frac{m}{n}}, A(m)\right)$. Let's consider the rates for $Pareto(\alpha)$ and $Beta(1, \beta)$ distribution. For both these cases, $A(m) = m^{-1}$, and hence, $\max\left(\frac{1}{m}, \sqrt{\frac{m}{n}}, A(m)\right)$ becomes, $\max\left(\frac{1}{m}, \sqrt{\frac{m}{n}}\right)$. For the $SN(\alpha)$ case, the rate of convergence turns out to be $\max\left(\sqrt{\frac{m}{n}}, \frac{1}{\log m}\right)$. Again, if we consider convergence to $F^n(a_n + b_n x)$ and not the limiting distribution, then for all three distributions, the convergence rate would be $\max\left(\frac{1}{m}, \sqrt{\frac{m}{n}}\right)$.

### 6.3.2 Unknown Normalizing Constants

The previous section dealt with only known normalizing constants. However, in practice, the normalizing constants are often unknown, and we might need to use estimates of the normalizing constants. Suppose, $l_n = \lfloor \frac{n}{m} \rfloor$ and $l'_n = \lfloor \frac{n}{em} \rfloor$. A possible estimate for the normalizing constants for the subsamples, as given by [50] are:

- **Type I:** $\hat{a}_m = F_n^{-1}(1 - \frac{1}{m}) = X_{(n-l_n)}$ and $\hat{b}_m = F_n^{-1}(1 - \frac{1}{em}) - F_n^{-1}(1 - \frac{1}{m}) = X_{(n-l'_n)} - X_{n-l_n}$.
- **Type II:** $\hat{b}_m = F_n^{-1}(1 - \frac{1}{m}) = X_{(n-l_n)}$.
- **Type III:** $\hat{a}_m = \omega(F_n) = X_{(n)}$ and $\hat{b}_m = \omega(F_n) - F_n^{-1}(1 - \frac{1}{m}) = X_{(n)} - X_{(n-l_n)}$.

[50] showed that the Bootstrap is consistent for choices of $\hat{a}_n$ and $\hat{b}_n$ such that $\frac{\hat{b}_m}{b_m} \to 1$ and $b_m^{-1}(\hat{a}_m - a_m) \to 0$. For the corresponding rate of convergence, it is enough to find the rate of

convergences of $\frac{\hat{b}_m}{b_m}$ and $b_m^{-1}(\hat{a}_m - a_m)$ to their corresponding limits. We will do this separately for the three cases.

- **Type III:** In this case, we know:

  - $a_n = \omega(F)$, $b_n = \omega(F) - \gamma_n$, where $\gamma_n = F^{-1}(1 - \frac{1}{n})$
  - $\hat{a}_m = \omega(F_n)$ and $\hat{b}_m = X_{(n)} - X_{(n-l_n)}$, where $l_n = \lfloor \frac{n}{m} \rfloor$.

  [50] already showed that $\frac{\hat{b}_m}{b_m} \to 1$ and $b_m^{-1}(\hat{a}_m - a_m) \to 0$. In the proof, it was shown that $\frac{\omega(F)-\gamma_n}{\omega(F)-\gamma_m} \to 0$, and $\frac{X_{(n-l_n)}-\omega(F)}{\omega(F)-\gamma_n} \to -1$, and since $\frac{\hat{b}_m}{b_m} = \frac{X_{(n)}-X_{(n-l_n)}}{\omega(F)-\gamma_m} = \frac{X_{(n)}-\omega(F)}{\omega(F)-\gamma_n} . \frac{\omega(F)-\gamma_n}{\omega(F)-\gamma_m} - \frac{X_{(n-l_n)}-\omega(F)}{\omega(F)-\gamma_n}$, the first condition follows. Here, we will compute the convergence rate. Building up on the proof by [50]:

  $$P\left(\left|\frac{X_{(n-l_n)} - \omega(F)}{\omega(F) - \gamma_n} - (-1)\right| > \epsilon\right) = P(X_{(n-l_n)} > u_m^{(1)}(\epsilon)) + P(X_{(n-l_n)} \leq u_m^{(2)}(\epsilon))$$

  where $u_m^{(1)}(\epsilon) = \omega(F) - (1-\epsilon)\left(\omega(F) - \gamma_m\right) = a_m + b_m(\epsilon - 1)$, and $u_m^{(2)}(\epsilon) = \omega(F) - (1 + \epsilon)\left(\omega(F) - \gamma_m\right) = a_m + b_m(-\epsilon - 1)$. Suppose, $S_{n,m}(x) = \#\{i | 1 \leq i \leq n, X_i \leq x\}$. Then, we can write:

$$
\begin{aligned}
P\left(X_{(n-l_n)} > u_m^{(1)}(\epsilon)\right) &= P\left(S_{n,m}(u_m^{(1)}(\epsilon)) > l_n\right) \\
&\sim P\left(\frac{m}{n}\sum_{i=1}^{n}\mathbb{I}(X_i > u_m^{(1)}(\epsilon)) > 1\right) \\
&= P\left(m(1 - F_n(a_m + b_m(\epsilon - 1))) > 1\right) \\
&= P\left(z_m^*(\epsilon - 1) > 1\right), \qquad \text{where } z_m^*(x) = m(1 - F_n(a_m + b_m x)) \\
&= P\left(z_m^*(\epsilon - 1) - z_m(\epsilon - 1) > 1 - z_m(\epsilon - 1)\right) \\
&\le P\left(\left|z_m^*(\epsilon - 1) - z_m(\epsilon - 1)\right| > 1 - z_m(\epsilon - 1)\right) \\
&\le \frac{E\left|z_m^* - z_m\right|^2}{(1 - z_m)^2}
\end{aligned}
$$

We already know, that $z_m(\epsilon - 1) \to c(\epsilon - 1) = (1 - \epsilon)^\alpha < 1$. Hence, the denominator is always positive, the rate of convergence being the same as of $E(z_m^* - z_m)^2$ which is $\frac{m}{n}$ (See Proof). Similarly, we will obtain the same rate of convergence for $P(X_{(n-l_n)} \le u_m^{(2)}(\epsilon))$. In that case, $z_m(-1 - \epsilon) \to (1 + \epsilon)^\alpha > 1$.

Now, since $F \in D(\Psi_\alpha)$, $1 - F(\omega(F) - t) = t^\alpha L(t)$ where $L$ is a slowly varying function at 0. Under this criterion, $1 - \frac{1}{n} \sim F(\gamma_n)$ ([50]). Hence:

$$
\begin{aligned}
\frac{m}{n} &\sim \frac{1 - F(\gamma_n)}{1 - F(\gamma_m)} \\
&= \left(\frac{\omega(F) - \gamma_n}{\omega(F) - \gamma_m}\right)^\alpha \frac{L(\omega(F) - \gamma_n)}{L(\omega(F) - \gamma_m)} \\
\Rightarrow \frac{\omega(F) - \gamma_n}{\omega(F) - \gamma_m} &= \left(\frac{m}{n}\right)^{\frac{1}{\alpha}} \left(\frac{L(\omega(F) - \gamma_m)}{L(\omega(F) - \gamma_n)}\right)^{\frac{1}{\alpha}} \qquad (6.3.5)
\end{aligned}
$$

Hence, we can state that the rate of convergence of $\frac{\omega(F)-\gamma_n}{\omega(F)-\gamma_m}$ to 0 is $\left(\frac{m}{n}\right)^{\frac{1}{\alpha}}$. Finally, the rate of convergence of $\frac{\hat{b}_m}{b_m}$ is $\max\left(\frac{m}{n}, \left(\frac{m}{n}\right)^{\frac{1}{\alpha}}\right)$. Hence, if $\alpha > 1$, the rate would be $\left(\frac{m}{n}\right)^{\frac{1}{\alpha}}$, while if it is a fraction, the rate would be simply $\frac{m}{n}$.

Next, we need the convergence rate of $b_m^{-1}(\hat{a}_m - a_m) = \frac{X_{(n)}-\omega(F)}{\omega(F)-\gamma_m} = \frac{X_{(n)}-\omega(F)}{\omega(F)-\gamma_n}\frac{\omega(F)-\gamma_n}{\omega(F)-\gamma_m}$. The rate at which this converges to 0 is same as the rate at which $\frac{\omega(F)-\gamma_n}{\omega(F)-\gamma_m}$ goes to 0. Hence, the final rate of convergence coming from unknown normalizing coefficients is $\max\left(\frac{m}{n}, \left(\frac{m}{n}\right)^{\frac{1}{\alpha}}\right)$.

- **Type II:** In this case, we know:

  - $a_n = 0$, $b_n = \gamma_n$, where $\gamma_n = F^{-1}(1 - \frac{1}{n})$
  - $\hat{a}_m = 0$ and $\hat{b}_m = X_{(n-l_n)}$, where $l_n = \lfloor \frac{n}{m} \rfloor$

  We need to compute the convergence rate of $\frac{X_{(n-l_n)}}{\gamma_m}$ to 1.

$$P\left(\left|\frac{X_{(n-l_n)}}{b_m} - 1\right| > \epsilon\right) = P(X_{(n-l_n)} > u_m^{(1)}(\epsilon)) + P(X_{(n-l_n)} \leq u_m^{(2)}(\epsilon))$$

where $u_m^{(1)}(\epsilon) = \gamma_m(1+\epsilon) = a_m + b_m(1+\epsilon)$, and $u_m^{(2)}(\epsilon) = \gamma_m(1-\epsilon) = a_m + b_m(1-\epsilon)$. Suppose, $S_{n,m}(x) = \#\{i | 1 \leq i \leq n, X_i \leq x\}$. Then, we can write:

$$
\begin{aligned}
P(X_{(n-l_n)} > u_m^{(1)}(\epsilon)) &= P(S_{n,m}(u_m^{(1)}(\epsilon)) > l_n) \\
&\sim P\left(\frac{m}{n}\sum_{i=1}^{n}\mathbb{I}(X_i > u_m^{(1)}(\epsilon)) > 1\right) \\
&= P\left(m(1 - F_n(a_m + b_m(1+\epsilon))) > 1\right) \\
&= P\left(z_m^*(1+\epsilon) > 1\right), \qquad \text{where } z_m^*(x) = m(1 - F_n(a_m + b_m x)) \\
&= P\left(z_m^*(1+\epsilon) - z_m(1+\epsilon) > 1 - z_m(1+\epsilon)\right) \\
&\leq P\left(\left|z_m^*(1+\epsilon) - z_m(1+\epsilon)\right| > 1 - z_m(1+\epsilon)\right) \\
&\leq \frac{E\left|z_m^* - z_m\right|^2}{(1 - z_m)^2}
\end{aligned}
$$

We already know, that $z_m(1+\epsilon) \to c(1+\epsilon) = (1+\epsilon)^{-\alpha} < 1$. Hence, the denominator is always positive, the rate of convergence being the same as of $E(z_m^* - z_m)^2$ which is $\frac{m}{n}$ (See Proof). Similar process follows with $P(X_{(n-l_n)} > u_m^{(2}(\epsilon)$. Hence, the rate of convergence for Type I random variables is $\frac{m}{n}$.

- **Type I:** In this case, we know:
  - $a_n = \gamma_n$, $b_n = F^{-1}(1 - \frac{1}{en})$, where $\gamma_n = F^{-1}(1 - \frac{1}{n}) - \gamma_n$
  - $\hat{a}_m = X_{(n-l_n)}$ and $\hat{b}_m = X_{(n-l'_n)} - X_{(n-l_n)}$, where $l_n = \lfloor \frac{n}{m} \rfloor$ and $l'_n = \lfloor \frac{n}{em} \rfloor$.

  Now, we have to find the rate of convergence for $\frac{\hat{b}_m}{b_m}$ and $b_m^{-1}(\hat{a}_m - a_m)$. We can write $\frac{\hat{b}_m}{b_m} = \frac{X_{(n-l'_n)} - X_{(n-l_n)}}{F^{-1}(1-\frac{1}{em}) - \gamma_m} = \frac{X_{(n-l'_n)} - \gamma_m}{F^{-1}(1-\frac{1}{em}) - \gamma_m} - \frac{X_{(n-l_n)} - \gamma_m}{F^{-1}(1-\frac{1}{em}) - \gamma_m}$. We can write:

$$
P\left(\left|\frac{X_{(n-l'_n)} - \gamma_m}{F^{-1}(1 - \frac{1}{em}) - \gamma_m} - 1\right| > \epsilon\right) = P\left(X_{(n-l'_n)} > u_m^{(1)}(\epsilon)\right) + P\left(X_{(n-l'_n)} \leq u_m^{(2)}(\epsilon)\right)
$$

where $u_m^{(1)}(\epsilon) = \gamma_m + \big(F^{-1}(1-\frac{1}{em}) - \gamma_m)(1+\epsilon)\big) = a_m + b_m(1+\epsilon)$. Similarly, $u_m^{(2)}(\epsilon) = a_m + b_m(1-\epsilon)$. Then,

$$\begin{aligned}
P(X_{(n-l'_n)} > u_m^{(1)}(\epsilon)) &= P(S_{n,m}(u_m^{(1)}(\epsilon)) > l'_n) \\
&\sim P\left(\frac{m}{n}\sum_{i=1}^{n}\mathbb{I}(X_i > u_m^{(1)}(\epsilon)) > \frac{1}{e}\right) \\
&= P\left(m(1-F_n(a_m + b_m(1+\epsilon))) > \frac{1}{e}\right) \\
&= P\left(z_m^*(1+\epsilon) > \frac{1}{e}\right), \qquad \text{where } z_m^*(x) = m(1-F_n(a_m+b_mx)) \\
&= P\left(z_m^*(1+\epsilon) - z_m(1+\epsilon) > \frac{1}{e} - z_m(1+\epsilon)\right) \\
&\leq P\left(\left|z_m^*(1+\epsilon) - z_m(1+\epsilon)\right| > \frac{1}{e} - z_m(1+\epsilon)\right) \\
&\leq \frac{E\left|z_m^* - z_m\right|^2}{(1-z_m)^2}
\end{aligned}$$

Now, $z_m(1+\epsilon) \to e^{-1-\epsilon} < \frac{1}{e}$, and hence $\frac{1}{e} - z_m > 0$. Thus, the rate of convergence is $\frac{m}{n}$. The part for $u_m^{(2)}(\epsilon)$ is similar. We can again find the convergence rate of $\frac{X_{(n-l_n)}-\gamma_m}{F^{-1}(1-\frac{1}{em})-\gamma_m}$.

Now, for the convergence rate of $b_m^{-1}(\hat{a}_m - a_m) = \frac{X_{(n-l_n)}-\gamma_m}{F^{-1}(1-\frac{1}{en})-\gamma_m}$. This we already found out to be $\frac{m}{n}$. Hence the extra addition for using $\hat{a}_m$ and $\hat{b}_m$ in this case is $\frac{m}{n}$.

*Remark* 6.3.2. Hence, we have obtained the effect of using $\hat{a}_m$ and $\hat{b}_m$ in place of $a_m$ and $b_m$ on the rate of convergence of the distribution of extremes. While the rate of convergence doesn't change for Type I and Type II random variables, i.e. random variables in the domain of attraction of Gumbel and Fréchet distribution, the rate changes for the Type III (Weibull) random variables, particularly when $\alpha > 1$ (if $\alpha \leq 1$, the rate will remain the same). For $\alpha > 1$ in Type III cases, the rate of convergence will be $\left(\frac{m}{n}\right)^{\frac{1}{\alpha}}$. However, the rate of convergence we calculated for known $a_m$ and $b_m$, the rate was $\max(\frac{1}{m}, \sqrt{\frac{m}{n}})$. The additional rate obtained from the estimation of $a_m$ and $b_m$ is $\left(\frac{m}{n}\right)^{\frac{1}{\alpha}}$ which is slower than $\sqrt{\frac{m}{n}}$ iff $\alpha > 2$. Hence we can finally conclude, the rate of convergence for non-parametric Bootstrap for extreme valued random variables, under unknown normalizing constants, is $\max(\frac{1}{m}, \sqrt{\frac{m}{n}})$ for Type I, Type II and Type III ($\alpha \leq 2$) cases, and $\max(\frac{1}{m}, \left(\frac{m}{n}\right)^{1/\alpha})$ for Type III case when $\alpha > 2$.

*Remark* 6.3.3. We saw earlier, the rate at which the distribution of extremes converges to the limit law is of the order $n^{\frac{\rho}{2\rho-1}}$, $\rho < 0$, which is always slower than $\sqrt{\frac{1}{n}}$. Moreover, using the limit law to approximate the extreme distribution involves estimating the normalizing coefficients, and the tail index $\gamma$. Even though there exists multitude of methods for these estimation procedures, none of them are consistently dominant over others, if we take both efficiency and computational simplicity into perspective. Furthermore, most of the estimation methods are applicable only to a restricted domain of the parameters. For example, the Hill's estimator works only when $\gamma > 0$. The moment estimator of the normalizing coefficient works for Type II distribution only when $\alpha > 2$. Hence, approximation by limit distribution involves efficient choice of estimation procedure, which is a complicated task in itself. Conversely, the $m$ out of $n$ Bootstrap procedures provides an easily computable approximation for the distribution of the extremes, with a convergence rate of $\frac{1}{n^{\frac{1}{3}}}$ for known normalizing constants. For unknown normalizing constants, this rate would prevail for Type I, Type II and Type III($\alpha \leq 2$) cases. For Type III case with $\alpha > 2$, the fastest rate would be $n^{-(1+\alpha)}$. In many cases (for example, Normal, Skew Normal), this rate will actually be faster than the rate of convergence to the limit law, and hence it is always better to use Bootstrap in these cases. For other cases, where $n^{\frac{\rho}{2\rho-1}}$ is faster than the Bootstrap convergence rate, it might be better to use the limiting distribution. But even in these cases, the problem of choosing the appropriate estimation procedure persists, and can give drastic errors under incorrect choice of estimates. Hence, even for these cases, it might be easier to use Bootstrap approximation, which would be of the rate $n^{-\frac{1}{3}}$, compared to the maximum attainable rate of estimating the tail-index, which is $\sqrt{\frac{1}{n}}$.

*Remark* 6.3.4. Let's consider the rate of convergence for the three example introduced in Section 6.1 for unknown normalizing constants. As we saw in this section, the rates would still be the same for Type I, Type II and Type III case with $\alpha \leq 2$, i.e. same as given in Remark 6.3.1. However, if $\beta > 2$ in $Beta(1, \beta)$ distribution, the rates of convergence will be different. The rate of convergence in this case to both the real distribution and the limit would be $O_p\left(\max\left(\frac{1}{m}, \left(\frac{m}{n}\right)^{\frac{1}{\beta}}\right)\right)$.

### 6.3.3 Optimal Resample Size

As we have seen till now, for the non-parametric case, the rate of convergence to the limit turns out to be $\max(\frac{1}{m}, \sqrt{\frac{m}{n}}, A(m))$. If we consider the convergence to the real distribution of the sample maximum for a given sample size, i.e. with $e^{-z_n}$, then we can ignore the $\zeta_3$ term, and hence the $A(m)$ term. This would reduce the rate of parametric Bootstrap to $\sqrt{\frac{1}{n}}$ and $min(\frac{1}{m}, \sqrt{\frac{m}{n}})$. In either case, we can see that the rate of convergence depends heavily of the Bootstrap sample size $m$. Hence, it is of significant interest to obtain an optimal Bootstrap sample size, keeping in mind that we have to satisfy $m = o(n)$ ([50]).

Since convergence to the actual distribution of the extremes is of primary importance, let us at first only consider the rate of convergence to the true distribution, which would be $\max\left(\frac{1}{m}, \sqrt{\frac{m}{n}}\right)$. As we can see in the proof (Section 6.9), the $P\left(|z_m^* - z_m) > t\right) \leq \frac{mc(x)}{nt^2}$, where $c(x) = -\log G(x)$, $G(\cdot)$ being the limiting distribution. The second term $\zeta_2^*(x, m, n)$ consisted of $e^{-z_m^*} - e^{-z_m}$, which would give the convergence rate to be $e^{-c(x)}\sqrt{\frac{mc(x)}{n}}$. Next, we have to concentrate on the coefficients for the $\frac{1}{m}$ term. Calculations show this to be $2e^{-c(x)}$ (see proof for details). Hence, an optimal sample size can be obtained by equating these two, which would give:

$$\begin{aligned} e^{-c(x)}\sqrt{\frac{mc(x)}{n}} &\sim \frac{2e^{-c(x)}}{m} \\ c(x)\frac{m}{n} &\sim \frac{4}{m^2} \\ \Rightarrow m &\sim \left(\frac{4n}{c(x)}\right)^{\frac{1}{3}} \end{aligned} \tag{6.3.6}$$

Hence, we have obtained an optimal Bootstrap sample size, depending only on the limiting distribution of the extremes, which is of the order $n^{\frac{1}{3}}$. However this rate depends on x, the point at which the distance is being minimized. To get a universal choice of m, we might approach in the following manner:

$$\begin{aligned} \sup_{x\in\mathbb{R}}\sqrt{\frac{mc(x)}{n}}e^{-c(x)} &\sim \sup_{x\in\mathbb{R}}\frac{2e^{-c(x)}}{m} \\ \left(\sup_{x\in\mathbb{R}}\sqrt{c(x)}e^{-c(x)}\right)^2\frac{m}{n} &\sim \frac{4}{m^2}, \qquad (\text{ since } \sup_{x\in\mathbb{R}} e^{-c(x)} = 1) \\ \Rightarrow m &\sim \left(\frac{4n}{\left(\sup_{x\in\mathbb{R}}\sqrt{c(x)}e^{-c(x)}\right)^2}\right)^{\frac{1}{3}} \end{aligned} \tag{6.3.7}$$

*Lemma* 6.3.1. Suppose $H(\cdot)$ be a distribution function with density $f(\cdot)$, and $\chi$ be it's support. Then:

$$\sup_{x\in\chi} H(x)\sqrt{-\log H(x)} = \sqrt{\frac{1}{2e}} \tag{6.3.8}$$

*Proof.* Let $g(x) = H(x)\sqrt{-\log H(x)}$. Then we can write:

$$g'(x) = -\frac{f(x)}{2\sqrt{-\log H(x)}} + f(x)\sqrt{-\log H(x)}$$

Equating to 0, we can write:

$$\begin{aligned} f(x)\sqrt{-\log H(x)} &= f(x)\frac{1}{2\sqrt{-\log H(x)}} \\ \Rightarrow -\log H(x) &= \frac{1}{2} \\ \Rightarrow H(x) &= e^{-\frac{1}{2}} \end{aligned}$$

The function $g'(x)$ clearly increases when $-\log H(x) > \frac{1}{2}$ and decreases for $\log H(x) < \frac{1}{2}$. Since $H(x)$ is a distribution function and hence non-decreasing, $-\log H(x)$ is a decreasing function of x. Hence, the function $g(x)$ increases till $\{x|H(x) = e^{-\frac{1}{2}}$, and then decreases, thereby giving an unique maximum. The value of the maximum would be: $e^{-\frac{1}{2}}\sqrt{\frac{1}{2}} = \sqrt{\frac{1}{2e}}$. Hence Proved. □

Using Lemma 6.3.1, we can now provide an unique optimal block size for all three types of distributions. The expression would be (keeping in mind $c(x) = -\log G(x)$):

$$m \sim \left(\frac{4n}{\frac{1}{2e}}\right)^{\frac{1}{3}} = (8en)^{\frac{1}{3}} \sim 2.8n^{\frac{1}{3}}$$

It is to be noted that this expression of optimal sample size is true for all three cases when the normalizing constants are known. However, when they are unknown, the optimal choice would be different only in Type III cases with $\alpha > 2$. In this particular, the calculation of the optimal Bootstrap sample size is a little complicated, since we have to compare the coefficient of $\frac{1}{m}$ with that of $\left(\frac{m}{n}\right)^{\frac{1}{\alpha}}$, as given in equation 6.3.5. In this case, the coefficient of $\left(\frac{m}{n}\right)^{\frac{1}{\alpha}}$ is $\frac{L(\omega(F)-\gamma_m)}{L(\omega(F)-\gamma_m)}$. Since L is a slowly varying function at 0, this would in turn converge to 1, and we can give an optimal sample size proposal by equating $\frac{2}{m}$

with $\left(\frac{m}{n}\right)^{\frac{1}{\alpha}}$. This would give an optimal sample size of $2^{\frac{\alpha}{1+\alpha}} n^{\frac{1}{1+\alpha}} = (2^\alpha n)^{\frac{1}{1+\alpha}}$. Hence, higher the $\alpha$, smaller will be the sample size and slower will be the convergence rate.

If we want to obtain the optimal sample size to minimize the convergence rate to the limit, we have to bring in $A(n)$ as defined in equation 6.2.7. In this case, we have to use the following result given by [41]:

$$\lim_{n\to\infty} \frac{F^n(a_n + b_n x) - H_\gamma(x)}{A(n)} = (c(x))^{1+\gamma} G_\gamma(x) f_\gamma(-\log c(x)) \tag{6.3.9}$$

where $G_\gamma(x)$ is the limiting distribution $G_\gamma(x) = exp\{-(1+\gamma x)^{\frac{1}{\gamma}}\}$ with $\gamma$ being the parameter for the generalized extreme value distribution, and

$$f_\gamma(x) = \begin{cases} \int_0^x e^{\gamma u} \int_0^u e^{\rho s} ds du, & \text{for } \gamma \geq 0 \\ -\int_x^\infty e^{\gamma u} \int_0^u e^{\rho s} ds du, & \text{for } \gamma < 0 \end{cases}$$

where $\lim_{t\to\infty} \frac{A(tx)}{A(t)} = x^\rho$ for $x > 0$. Let us denote $(c(x))^{1+\gamma} G_\gamma(x) f_\gamma(-\log c(x))$ as $\psi(x)$. Then:

- If $A(m)$ is slower than $\frac{1}{m}$, we have to equate $e^{-c(x)}\sqrt{\frac{mc(x)}{n}}$ with $\psi(x)$ to obtain an optimal value of $m$ as a function of n.
- If $A(m)$ is of the same order as $\frac{1}{m}$, then we have compare the coefficients $2e^{-c(x)}$ and $\psi(x)$ to see which one is smaller, and then equate $e^{-c(x)}\sqrt{\frac{mc(x)}{n}}$ with that multiplied by $\frac{1}{m}$.
- If $A(m)$ decays faster than $\frac{1}{m}$, then we can use the earlier obtained value of m as an optimal Bootstrap sample size.

However, the convergence rate to the true distribution is of primary significance, and hence it is advisable to use the optimal Bootstrap size of $2.8n^{\frac{1}{3}}$ for all three cases, except for Type II case with $\alpha > 2$, in which case the optimal size would be $(2^\alpha n)^{\frac{1}{1+\alpha}}$. Finally, we can summarize the discussion of this section in the following theorem:

**Theorem 6.3.2.** *Suppose $X_1, \ldots, X_n$ are random variables from a distribution function $F$, in the domain of attraction of $G$, one of the three types of the extreme value distribution. Then an optimal choice of $m$ for conducting $m$ out of $n$ Bootstrap is $(8en)^{\frac{1}{3}}$, for all three types of distributions, except for distribution in the domain of convergence of Type III or Weibull ($\Psi_\alpha$) with $\alpha > 2$, in which case, the optimal sample size would be $(2^\alpha n)^{\frac{1}{1+\alpha}}$.*

*Remark* 6.3.5. Let us now obtain the optimal sample size for the three distributions chosen for example. Since, we are primarily interested in the optimal sample size for maximum rate of convergence of the Bootstrap distribution to the real distribution of the extremes, an optimal sample size for all three, Pareto, Beta and Skew Normal would be $2.8n^{\frac{1}{3}}$, except the case when the shape parameter of the Beta random variable is greater than 2. In that case, i.e. for $Beta(1, \beta)$ distribution with $\beta > 2$, the optimal Bootstrap sample size would be $(2^\alpha n)^{\frac{1}{1+\alpha}}$. However, if we want to choose Bootstrap sample size such that the convergence rate to the limiting law is minimum, then we have to take the rate $A(n)$ into consideration. As given in Appendix, the rate $A(n)$ for Skew Normal is $\frac{1}{\log n}$, which is slower than $\frac{1}{m}$, and hence the optimal sample size would be $2.8n^{\frac{1}{3}}$. For Pareto and Beta, $A(n) = \frac{1}{n}$, and hence, we need to compare $\psi(x)$ with $2e^{-c(x)}$.

## 6.4 Comparison with Approximating by Limit Distribution

The traditional approach in Extreme Value Analysis involves approximating the distribution of the maximum by the limit distribution, which is one of three types as proved by [55]. The $m$ out of $n$ non-parametric Bootstrap proposes an alternative method, and hence comparison of the two procedure is of significant importance. [41] showed that under known normalizing constants, the extremes of random variables converges to the limit distribution at a rate of $A(n) = \frac{nf''(n)}{f'(n)} - \gamma + 1$, where $f(\cdot) = \left(\frac{1}{-\log F}\right)^{-1}$, and $\gamma$ is the extreme value index. Theorem 6.4.1 provides an outline of this rate of convergence.

**Theorem 6.4.1.** *Suppose $M(t) = \frac{1-F(t)}{F'(t)}$ be the Mill's ratio of the underlying distribution. Then the convergence rate $A(t)$ can be summarised for the three cases as follows:*

- *Type I (Gumbel):*
    - $\omega(F) < \infty$ : *In this case, the Mill's ratio will converge to* $0$, *and so will it's derivative.* $A(n)$ *will be the maximum of these two rates.*
    - $\omega(F) = \infty$ : *The expression is complicated in this casewith* $A(n)$ *being the rate at which* $\frac{c(t)^2 - c''(t)M(t)^2}{(c(t)+c'(t)M(t))^2}$ *converges to* $1$, *where* $c(t)$ *is as in* $1-F(t) = c(t)exp\left(-\int_{t_0}^{t}\frac{ds}{f(s)}\right)$ *([40]).*
- *Type II (Frechet,* $D(\Phi_\alpha)$*): The rate of convergence would be the maximum of the rate of convergence of the derivative of Mill's ratio, and that of* $\frac{L(tx)}{L(t)}$, *where* $1 - F(t) = t^{-\alpha}L(t)$
- *Type III (Weibull,* $D(\Psi_\alpha)$*): The rate of convergence would be the maximum of the rate of convergence of the derivative of Mill's ratio, and that of* $\frac{L(tx)}{L(t)}$, *where* $1 - F(\omega(F) - t) = t^{\alpha}L(t)$.

*Therefore, the non-parametric Bootstrap using the proposed optimal block size* $m_{opt} = 2.8n^{\frac{1}{3}}$ *will always perform better than the traditional method whenever the above mentioned rates are slower than* $n^{-\frac{1}{3}}$.

In practical applications, the normalizing constants are often unknown. There exists numerous estimation procedures of these normalizing constants, all of which involves estimation of the tail-index $\gamma$, in Type II and Type III cases. There are multiple works on the estimation of $\gamma$, some of them being Hill Estimator ([62]), Pickand's Estimator ([100]), MLE estimator, Moment Estimator, etc. Each of these estimators are consistent only in some domain of $\gamma$, and choice of the most efficient method is still unknown. The rate of convergence is given by $n^{\frac{\rho}{1-2\rho}}$, where $\rho$ is the constant in the second order regular variation condition, i.e. $\lim_{t\to\infty}\frac{\log U(tx)-\log U(t)-\gamma\log x}{b(t)} = \frac{x^{\rho}-1}{\rho}$, $U = \left(\frac{1}{1-F}\right)^{-1}$, $\rho \leq 0$ and $\gamma$ is the tail index. Hence the rate of convergence of the estimator is always slower than $\sqrt{\frac{1}{n}}$. The following theorem can be proposed:

**Theorem 6.4.2.** *Suppose* $\lim_{t\to\infty}\frac{\log U(tx)-\log U(t)-\gamma\log x}{b(t)} = \frac{x^{\rho}-1}{\rho}$, $U = \left(\frac{1}{1-F}\right)^{-1}$, $\rho \leq 0$ *and* $\gamma$ *is the tail index. Let the normalizing constants be unknown. Then:*

- *Type I, Type II and Type III* ($\alpha > 2$)*: In these cases, the rate of convergence of non-parametric Bootstrap is* $n^{-\frac{1}{3}}$*. The rate of convergence of the conventional method of approximating by limiting distribution is the maximum of* $n^{\frac{\rho}{1-2\rho}}$ *and* $A(n)$*, which was discussed in Theorem 6.4.1 for the three cases. Hence* $m$ *out* $n$ *non-parametric Bootstrap with optimal Bootstrap sample size* $m = 2.8n^{\frac{1}{3}}$ *performs better than approximating by limiting distribution, whenever the maximum of* $n^{\frac{\rho}{1-2\rho}}$ *and* $A(n)$ *is slower than* $n^{-\frac{1}{3}}$*.*

- *Type II* ($\alpha > 2$)*: In this case rate of convergence of non-parametric Bootstrap is* $n^{-\frac{1}{1+\alpha}}$*. Hence, the* $m$ *out* $n$ *Bootstrap will perform better if the maximum of* $n^{\frac{\rho}{1-2\rho}}$ *and* $A(n)$ *is slower than* $n^{-\frac{1}{1+\alpha}}$

It is to be noted that approximating by limiting distribution has a convergence rate, which is always slower than $\sqrt{\frac{1}{n}}$. Contrarily, the m out n Bootstrap provides a convergence rate of $n^{-\frac{1}{3}}$ in most cases, except for Type III with $\alpha > 2$. Additionally, the traditional method involves the efficient choice of estimation procedure, and also testing procedure for correct choice of limiting distribution. The $m$ out of $n$ Bootstrap, is much simpler to apply, and doesn't involve any parameter estimation.

## 6.5 Multivariate Extreme Values

Let $\mathcal{X} = (X^{(1)}, \dots, X^{(d)})$ be from a d-dimensional random variable. Suppose, we have n observations $\mathcal{X}_1, \dots, \mathcal{X}_n$. Then, the distribution function can be written as:

$$F(\mathbf{x}) = P(X^{(1)} \leq x_1, \dots, X^{(d)} \leq x_d) = C(F_1(x_1), \dots, F_d(x_d))$$

where $F_i$ is the $i^{th}$ marginal distribution, and C is the dependence function, or the copula.

In this work, we will investigate the asymptotic thoery of non-parametric Bootstrap of component-wise maximum of random variables. Let $H_n(\mathbf{z}) = P(Z_{1,n} < z_1, \dots Z_{d,n} < z_d) = F^n(\mathbf{z})$, $Z_{i,n}$ be the maximum of the $i-th$ marginal, and $\mathbf{z} = (z_1, \dots, z_d)$. Galambos proved that if we consider $\mathbf{a}_n$ and $\mathbf{b}_n$ to be the vectors $(a_n^{(1)}, \dots, a_n^{(d)})$ and $(b_n^{(1)}, \dots, b_n^{(d)})$, where $a_n^{(i)}$ and $b_n^{(i)}$ are the scalings corresponding to the $i^{th}$ marginal distribution, then:

$$P(B_n^{-1}(\mathbf{z}_n - \mathbf{a}_n) \leq z) \to H(z) \tag{6.5.1}$$

where $B_n$ is the diagonal matrix with $i^{th}$ diagonal elements $b_n^{(i)}$, $\mathbf{z}$ the d-dimensional random variable with maximum of the marginals, and H is the non-degenerate limiting distribution. The above result is true if and only if each marginal of $F$ belongs to the domain of attraction of one of the three types of distributions (Gumbel, Frechet, Weibull), and

$$C_F^n(y_1^{1/n}, y_2^{1/n}, \dots, y_d^{1/n}) \to C_H(y_1, \dots, y_d)$$

It can be shown that the marginals of $H$ are the limiting distribution of the corresponding marginal of F. Additionally, if F(x) be such that, with some sequence $a_n$ and $b_n$, $H_n(a_n + b_n x) \to H(x)$, then

$$C_H^k(y_1^{1/k}, \dots, y_m^{1/k}) = C_H(y_1, \dots, y_d), \qquad \text{for any } k \geq 0 \tag{6.5.2}$$

Let $F_n$ be the empirical cdf of $\{X_1, \dots, X_n\}$. Let $Y_i^*, i = 1, \dots, m$ be m iid random vectors in $\mathbb{R}^d$ with the cdf $F_n$ and let $M_Y = \left(Y_{(m)}^{1*}, \dots, Y_{(m)}^{d*}\right)$. Define

$$H_{m,n}(\mathbf{x}) = P\left(b_m^{-1}(M_Y - a_m) \leq x | X_1, \dots, X_n\right)$$

. Then [50] showed that $\sup_{x \in \mathbb{R}^d} \left| H_{m,n}(x) - G(x) \right| \to 0$ in probability. In this section, we will determine the rate of convergence.

Empirical copula, as defined by Deheuvels, can be written as:

$$C_n(u) = F_n(F_{n1}^{-1}(u_1), \dots, F_{nd}^{-1}(u_d)), u \in [0,1]^d \tag{6.5.3}$$

where $F_n$ is the empirical cdf and $F_{ni}$ is the empirical cdf of the $i^{th}$ marginal. In other words, we can write the empirical cdf of the multivariate distribution F as

$$F(\mathbf{x}) = C_n(F_1(x_1), \dots, F_d(x_d))$$

[117] proves that, under some regularity conditions, the process $\mathbb{C}_n = \sqrt{n}(C_n - C)$ converges weakly to $\alpha(u) - \sum_{j=1}^{d} C'_j(u)\alpha_j(u_j)$ where $\alpha_j$ is a C-Brownian Bridge, i.e. a tight centered Gaussian process with covariance function: $Cov(\alpha(u), \alpha(v)) = C(u \wedge v) - C(u)C(v)$. Here, we will assume the regularity conditions proposed by [117]. If we further assume continuity of second order partial derivatives of C, then [135] showed that:

$$\sup_{u\in[0,1]^d} \left|\mathbb{C}_n(u) - \tilde{\mathbb{C}}_n(u)\right| = O(n^{-\frac{1}{4}}\sqrt{\log n}(\log\log n)^{\frac{1}{4}}) \tag{6.5.4}$$

where $\tilde{\mathbb{C}}_n(u) = \alpha_n(u) - \sum_{j=1}^{d} C'_j(u)\alpha_{j,n}(u_j)$, $\alpha_n$ converging to $\alpha$, the C-Brownian Bridge.

To obtain the rate of convergence in multivariate set-up, we would also need the rate of convergence for dependence functions (or Copula), as given by [97]. Suppose

$$r_n = \sup_{u\in[0,1]^k} |C^n(u_1^{\frac{1}{n}}, \ldots, u_k^{\frac{1}{n}}) - C_H(u_1, \ldots, u_k)| \tag{6.5.5}$$

By using transformation $x = -\frac{1}{\log u}$, we can rewrite $r_n$ as:

$$\sup_{x\in[0,\infty)^k} |K^n(nx_1, \ldots, nx_k) - G(x_1, \ldots, x_k)| \tag{6.5.6}$$

where each of the marginals of K and G has the distribution function $\phi_1(x) = e^{-\frac{1}{x}}(x \geq 0)$, and G is max-stable. [97] defined weighted Kolmogorov metric between two random variables V and W as follows:

$$\rho_\psi(V, W) = \underset{x\in[0,\infty)^k}{sup} \psi(x)|F_V(x) - F_W(x)| \tag{6.5.7}$$

where $\psi$ is a continuous function, increasing to $\infty$ in each component. Let $\phi(x) = \psi(x, \ldots, x)$, and $\rho_\phi(V, W) = \sup_{\mathbf{x}\in[0,1]^k} \psi(\|x\|)|F_V(\mathbf{x}) - F_W(\mathbf{x})$. Suppose, $g(a) = \sup_{x\geq 0}\frac{\phi_1(xa)}{\phi(x}$, $R(n) = ng(\frac{1}{n-1})$, and V and W be random variables with distribution function $K$ and $G$ respectively, then

[97] proved that if $\rho_\psi(V, W) < \infty$, then for each $n \geq 2$, $r_n \leq R(n)\rho_\psi(V, W)$. Also, if $\rho_\phi(V, W) < \infty$, then for each $n \geq 2$, $r_n \leq R(n)\rho_\phi(V, W)$.

Returning to our problem, we have to find the convergence rate of:

$$\begin{aligned} F_n^m(\mathbf{a}_m + B_m\mathbf{x}) - H(\mathbf{x}) &= C_n^m(F_1(a_m^{(1)} + B_m^{(1)}x_1), \dots, F_d(a_m^{(d)} + B_m^{(d)}x_d)) - C_H(H_1(x_1), \dots, H_d(x_d)) \\ &= C_n^m(y_1^{\frac{1}{m}}, \cdots, y_d^{\frac{1}{m}}) - C_H(H_1(x_1), \dots, H_d(x_d)) \qquad (6.5.8) \end{aligned}$$

where $y_i = F_1^m(a_m^{(i)} + b_m^{(i)}x_i)$ for $i = 1, \dots, d$. We know from our results from univariate framework, $y_i \to H_i(x_i)$. 6.5.8 can be further divided into 3 parts, say $\nu_1, \nu_2$ and $\nu_3$, where:

- $\nu_1 = C_n^m(y_1^{\frac{1}{m}}, \dots, y_d^{\frac{1}{m}}) - C^m(y_1^{\frac{1}{m}}, \dots, y_d^{\frac{1}{m}})$
- $\nu_2 = C^m(y_1^{\frac{1}{m}}, \dots, y_d^{\frac{1}{m}}) - C_H(y_1, \dots, y_d)$
- $C_H(y_1, \dots, y_d) - C_H(H_1(x_1), \dots, H_d(x_d))$

Thus, we can write:

$$F_n^m(\mathbf{a}_m + B_m\mathbf{x}) - H(\mathbf{x}) = \nu_1 + \nu_2 + \nu_3 \qquad (6.5.9)$$

The first term $\nu_1$ goes to zero at a rate of $\sqrt{\frac{1}{n}}$, according to the discussion above, based on the results of [117]. The convergence rate of the second term has been discussed in [97]. Finally, rate of convergence of $\nu_3$ is same as the slowest rate of convergence of the marginals to the limit distribution, multiplied by the derivative of the copula of the limiting distribution. Thus, the final rate of convergence would be the slowest of the three rates. We can summarize the findings of this section in the following theorem:

**Theorem 6.5.1.** *Suppose* $= (X^{(1)}, \dots, X^{(d)})$ *be a d-dimensional random vector. Suppose, we have n observations* $_1, \dots, \mathcal{X}_n$. *Then,* $P(B_m^{-1}(\mathbf{z}_n^* - \mathbf{a}_m) \leq z)$ *converges to* $H(x)$ *at a rate* *of* $\max\left(\sqrt{\frac{1}{n}}, \frac{1}{m}, \sqrt{\frac{m}{n}}, R(m)\right)$ *where* $Z_n$ *is the component-wise maximum of the marginals,* $a_n$ *is the vector containing the* $a_n^{(i)}$ *of the* $i^{th}$ *marginal, and* $B$ *is the diagonal matrix with diagonal elements the* $b_n^{(i)}$ *of* $i^{th}$ *marginal,* $H(x)$ *is the limiting distribution, where each marginal is the limiting distribution of the maximum of the corresponding component. The rate* $R(n)$ *is as described by [97].*

## 6.6 Simulation

In this section, we will conduct simulation studies on one example from each of the three types of distribution: Type I (SN($\alpha$)), Type II (Pareto($\alpha$)) and Type III (Beta($1, \beta$)).

### 6.6.1 Pareto($\alpha$):

We know, the distribution function for Pareto($\alpha$) is given as: $F(x) = 1 - x^{-\alpha}$ for $x > 1$. It can be shown that under suitable normalization, it belongs to the domain of attraction of $D(\Phi_\alpha)$, i.e. the distribution of maximum converges to the distribution $H_{1,\alpha}(x) = e^{-x^{-\alpha}}$ for $x > 0$. The normalizing constants can be easily derived to be:

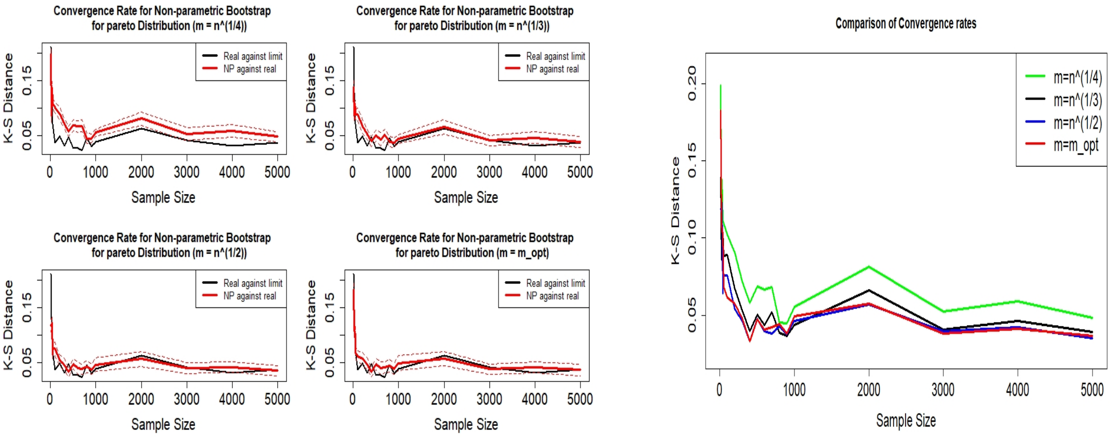
- $a_n = 0$
- $b_n = n^{\frac{1}{\alpha}}$

The maximum likelihood estimator of $\alpha$ based on n iid random variables $X_1, \ldots, X_n$ from Pareto($\alpha$) distribution is given by:

$$\hat{\alpha}_{MLE} = \frac{n}{\sum\limits_{i=1}^{n} \log X_i}$$

Four our simulation study, we considered:

- $\alpha = 4$
- Sample size from 20 to 10000
- Number of Bootstraps: 1000



(a) Performance of Parametric and Non-Parametric Bootstrap for Pareto

(b) Comparison of Bootstrap for Pareto under different Bootstrap sample size

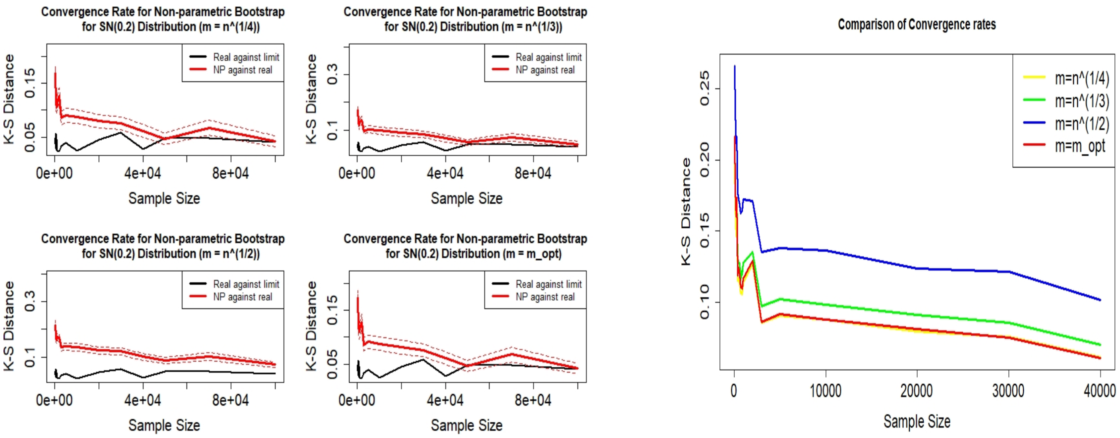
Figure 6.1: Comparison of convergence rate of Non-Parametric Bootstrap under Pareto distribution for different sample sizes

Since $A(n)$ is of the order of $\frac{1}{n}$ (6.10.1), Bootstrap will not provide an improvement in convergence rate under known $a_n$ and $b_n$. However, if we consider the normalizng constants to be unknown, the rate of convergence of extremes to their limit law will always be slower than $\sqrt{\frac{1}{n}}$. Figure 6.1 exhibits that the rate of convergence is quite close to the Bootstrap rate of convergence attained under choice of optimal sample size, $m_{opt}$, as proposed in Section 6.3.3. Figure 6.1(b) gives the comparison of convergence rate under different choice of sample size $m = m(n)$, and we can see that the best rate of convergence is attained under the optimal block size, $(8en)^{\frac{1}{3}}$.

### 6.6.2 Beta(1,$\beta$):

The distribution function of Beta(1,$\beta$) is given by: $F(x) = 1 - (1-x)^\beta$ for $x \in (0,1)$. It can be shown that under suitable normalization, it belongs to the domain of attraction of $D(\Psi_\alpha)$, i.e. the distribution of maximum converges to the distribution $H_{2,\alpha}(x) = e^{-(-x)^{-\alpha}}$ for $x \leq 0$. The normalizing constants can be easily derived to be:

- $a_n = 1$
- $b_n = n^{-\frac{1}{\beta}}$



(a) Performance of Parametric and Non-Parametric Bootstrap for Left Skewed Beta

(b) Comparison of Bootstrap for Beta under different Bootstrap sample size

Figure 6.2: Comparison of convergence rate of Parametric and Non-Parametric Bootstrap under Beta distribution

The maximum likelihood estimator of $\beta$ based on n iid random variables $X_1, \ldots, X_n$ from Beta(1,$\beta$) distribution is given by:

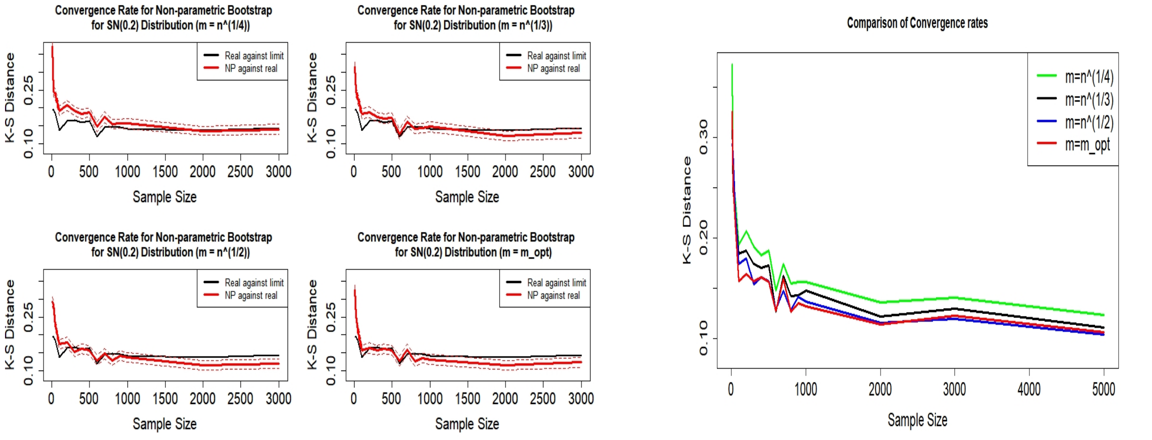
$$\hat{\beta}_{MLE} = -\frac{n}{\sum_{i=1}^{n} \log(1 - X_i)}$$

We saw earlier that the optimal choice of $m = m_{opt}$ in this case depends on the value of $\beta$. For $\beta \leq 2$, we can choose the optimal sample size to be $(8en)^{\frac{1}{3}}$ as in other cases. However, for $\beta > 2$, the optimal sample size would be $(2^{\alpha}n)^{\frac{1}{1+\alpha}}$. In this simulation study, we will consider $\alpha = 4$, to observe the rate of converge when $\alpha > 2$. As we can see in Figure 6.2, in this case, the convergence rate for the optimum sample size is the fastest amongst the chosen sample sizes. In this case, the convergence rate to the limiting law in this case is $\frac{1}{n}$ (6.10.2), so Bootstrap will not be able to provide better approximation than the limiting distribution, even though it performs rather similar to the limit law when the sample size is relatively large. However, in the simulation we used maximum likelihood estimator of $\beta$. The non-parametric estimates discussed in Section 6.3.3 brings more complexities in using the limiting distribution as an approximation. Hence, Bootstrap might be able to provide an easier to use alternative even in this scenario.

### 6.6.3 SN($\alpha$):

The density of SN($\alpha$) can be written as $f(x) = 2\phi(x)\Phi(\alpha x)$, with the distribution function being $F(x) = \Phi(x) - 2T(x,\alpha)$, where $T(\cdot)$ is the Owen's T-Function ($T(x,\alpha) = \frac{1}{2\pi}\int_0^{\alpha} \frac{e^{-\frac{1}{2}x^2(1+y^2)}}{1+y^2}dy$, which same as the probability of the event $X > x$ and $0 < Y < \alpha X$ where X and Y are independent standard normal random variables). It can be shown ([80]) that the extreme of random variables from the skewed normal family belongs in the domain of attraction of Type I (D($\Lambda$)) distributions. In this simulation study, we will only use $\alpha < 0$, for which the normalizing constants are given by (see 6.10.3):

- $a_n = (1+\alpha^2)^{-\frac{1}{2}}(2\log n)^{-\frac{1}{2}}$
- $b_n = \left(\frac{2\log n}{1+\alpha^2}\right)^{\frac{1}{2}} - \frac{\log\log n + \log\pi}{(1+\alpha^2)^{\frac{1}{2}}(2\log n)^{\frac{1}{2}}}$



(a) Performance of Parametric and Non-Parametric Bootstrap for Skew Normal Distribution

(b) Comparison of Bootstrap for Skew Normal under different Bootstrap sample size

Figure 6.3: Comparison of convergence rate of Parametric and Non-Parametric Bootstrap under Skew Normal distribution

There is no closed form expression for the maximum likelihood estimator of $\alpha$. The MLE would be a solution of the equation:

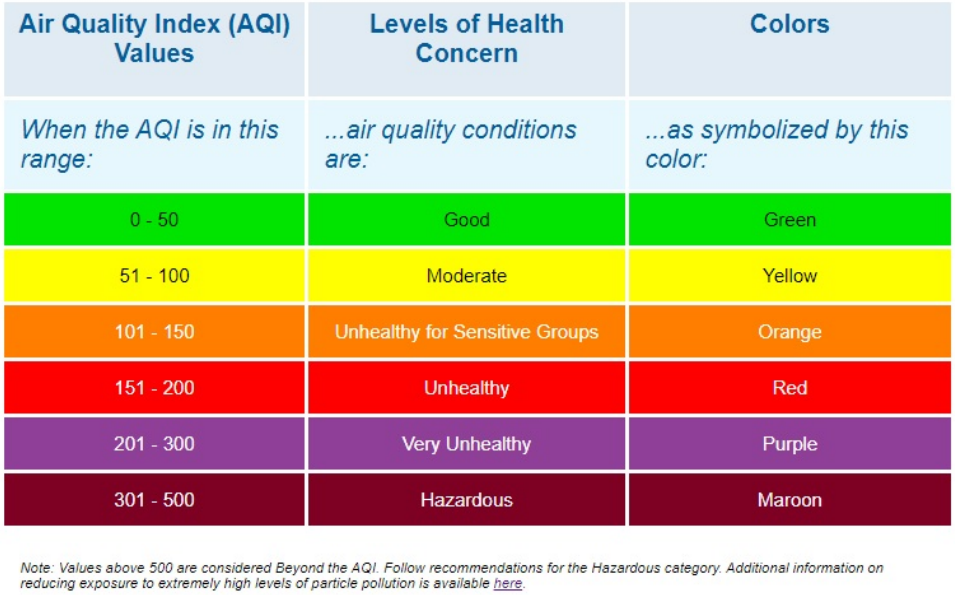
$$\sum_{i=1}^{n} \frac{\phi(\alpha X_i)}{\Phi(\alpha X_i)} X_i = 0$$

We can find the MLE numerically in R, and use it for our simulation. In this case, we considered $\alpha$ to be 0.2, and all other parametres are same as the earlier examples.

As we can see, in this case, the non-parametric Bootstrap actually performs better than the actual case. This is owing to the fact that under the optimal Bootstrap sample size, the rate at which non-parametric Bootstrap distribution converges to the actual distribution is of the order of $\left(\frac{1}{8en}\right)^{\frac{1}{3}}$, which is significantly faster than the convergence rate of $\frac{1}{\log n}$ (6.10.3). Also, we can see from the Fig 6.3 that the optimal choice of Bootstrap sample size gives better convergence rate than other choices of $m$ taken in this example. Hence, in this case, Bootstrap always provides a better approximation of the distribution of the extreme than the limiting distribution.

## 6.7 Real Data Analysis

Air Quality is an essential aspect of life. The quality of air we breathe in day to day has immense influence on our health. We know that the air around us contain various pollutants, ranging from completely harmless to particles that can be fatal above a certain threshold. Some such examples of harmful pollutants present in our atmosphere include particle pollution (or particulate matter), Ground-level ozone, Carbon monoxide, Sulfur oxides, Nitrogen oxides and Lead. These pollutants can be extremely harmful on excessive exposure, and abundance of these particles in the atmosphere can cause harm to the inhabitants of an entire city. For example, excessive exposure to Ozone can cause varying levels of health problems, like irritation of respiratory system, reduced lung function, lung infection, and may even cause permanent lung damage. Particulate Matters in high concentration are known to be fatal for patients of heart or lung diseases and older adults, and may lead to cardiac arrhythmia and heart attacks even in healthy individuals. [144] demonstrated that particulate matters significantly increase the possibility of cardiovascular diseases, particularly in people above the age of 40. [123] discusses about the effect of air pollution in children's health. [77] gives an overview of the risk of air pollution in human population. Hence, it is imperative to control the level of these air pollutants in the atmosphere, and keep their concentration under tolerable limits. The US Environmental Protection Agency (EPA) collaborates with local environmental agencies in different countries, in order to provide information on the concentration of these air pollutants in cities around the world. The Air Quality Index (AQI) provides an indicator of the concentrations of four major air pollutants (ground level ozone, particle pollution, carbon monoxide, and sulfur dioxide), according to the Clean Air Act of 1963. EPA provides the tolerable ranges of AQI, as given in Figure 6.4. As we can see, AQI level above 100 is considered unhealthy for sensitive groups, for example patients of lung and heart diseases. AQI level above 200 is considered extremely unhealthy for all groups, while that above 300 is declared as hazardous. However, in India, the AQI level goes as high as 1000 (New Delhi in September 2019), which can cause significant health problem for the people living in the area.



| Air Quality Index (AQI) Values | Levels of Health Concern | Colors |
|---|---|---|
| *When the AQI is in this range:* | *...air quality conditions are:* | *...as symbolized by this color:* |
| 0 - 50 | Good | Green |
| 51 - 100 | Moderate | Yellow |
| 101 - 150 | Unhealthy for Sensitive Groups | Orange |
| 151 - 200 | Unhealthy | Red |
| 201 - 300 | Very Unhealthy | Purple |
| 301 - 500 | Hazardous | Maroon |

*Note: Values above 500 are considered Beyond the AQI. Follow recommendations for the Hazardous category. Additional information on reducing exposure to extremely high levels of particle pollution is available here.*

Figure 6.4: Tolerable levels of AQI

AQI level should always be below a certain threshold. In other words, the maximum of the AQI level should never exceed a certain limit, for which one would require an extreme value analysis. This chapter provides an Extreme Value Analysis using Bootstrap, as proposed in this work, to determine the probability of the extreme of the AQI level during a certain period crossing a given threshold. The extreme value analysis using m out of n bootstrap is conducted on the AQI data of 13 cities around the world, 5 Indian and 8 outside India, to obtain the probability of the AQI level going above ceratin threshold in these cities. The primary aim is to demonstrate the exigency of the air pollution crisis in Indian cities, compared with other "highly polluted" cities in the world.

### Data Description

We have used Air Quality Index (AQI) data from five metropolitan cities in India: Delhi, Kolkata, Mumbai, Chennai and Hyderabad, and eight metropolitan cities outside India (collected from airnow.gov). The aim is to compare the air pollution crisis in India with other highly polluted metropolitan cities around the world. The data consists of hourly Air Quality Indices collected over the year, for last 5 years. Since, this is a time series data, we might expect temporal dependence within the data. To ameliorate issues arising from the dependence structure, we only considered the daily maximum of the AQI values in a city, and have assumed that the maximum AQI per day follows a weakly dependent structure, i.e. the dependence between the daily maximum Air Quality Index decays exponentially with time. It is sensible to take the daily maximum, since we are mainly concerned about the AQI crossing a particular threshold. Additionally, the data is divided into two parts, the summer data (March to September) and winter data (August to February). This is because in winter, air pollutants get trapped inside a warm layer of air, created by temperature inversion. Additionally, the daily maximum AQI, is undergone a transformation, primarily for computational simplicity.

The cities selected around the world for comparison with the Indian cities are: Lima (Peru), Bogota (Colombia), Kampala (Uganda), Manama (Bahrain), Chengdu (China), Colombo (Sri Lanka), Dhaka (Bangladesh) and Jakarta (Indonesia). Lima ranks first in the list of top polluted cities of South America (according to WHO), while Bogota is currently at $15^{th}$, and is the second most polluted city in Colombia. Both cities experience high concentration of Particulate Matter in their atmosphere, which are extremely harmful for human health. Air pollution in China has been a topic of discussion for a long time. While Beijing, which used to be the most polluted city in China, has improved the air condition recently owing to Air Pollution Prevention and Control Laws, the situation in Chengdu has not improved over the years, with reports suggesting that the situation now is even worse than that of Beijing. Dhaka and Jakarta also face major crisis in terms of presence of particulate matters and ozone concentration in the air, and are among the top polluted cities in Asia. The city of Manama is selected due to it's location, being near the largest industrial area in the world, the Jubail industrial city in Saudi Arabia. Manama is the nearest city to Jubail where EPA collects the Air Quality Index Data. Finally, the five metropolitan cities in India were selected, in order to assess the severity of the air pollution problem in India compared to other regions of the world.

Figure 6.5 and Figure 6.6 provides an overview of the AQI levels in two cities (Bogota and New Delhi). As we can see the AQI level in Bogota remains almost steady throughout the year at a level of about 120 – 150. However, the AQI in New Delhi experiences a periodic pattern, the peaks occurring mainly during the time of November or December, i.e. the time of Diwali. The AQI levels is seen to be drastically rising during this period, the highest peak going as high as 1100. Also, the AQI throughout remains significantly high, posing serious threat to the living population in the region.

### Results and Discussion

Table 6.1, 6.2, 6.3 provides the estimates of the probability of the daily maximum AQI crossing threshold: 200, 300 and 500 respectively. As we saw in Figure 6.4, AQI values of above 200 are considered to be unhealthy for all people, above 300 are considered to hazardous, and above 500 are not even listed in the AQI scale. Ideally, the probability of the

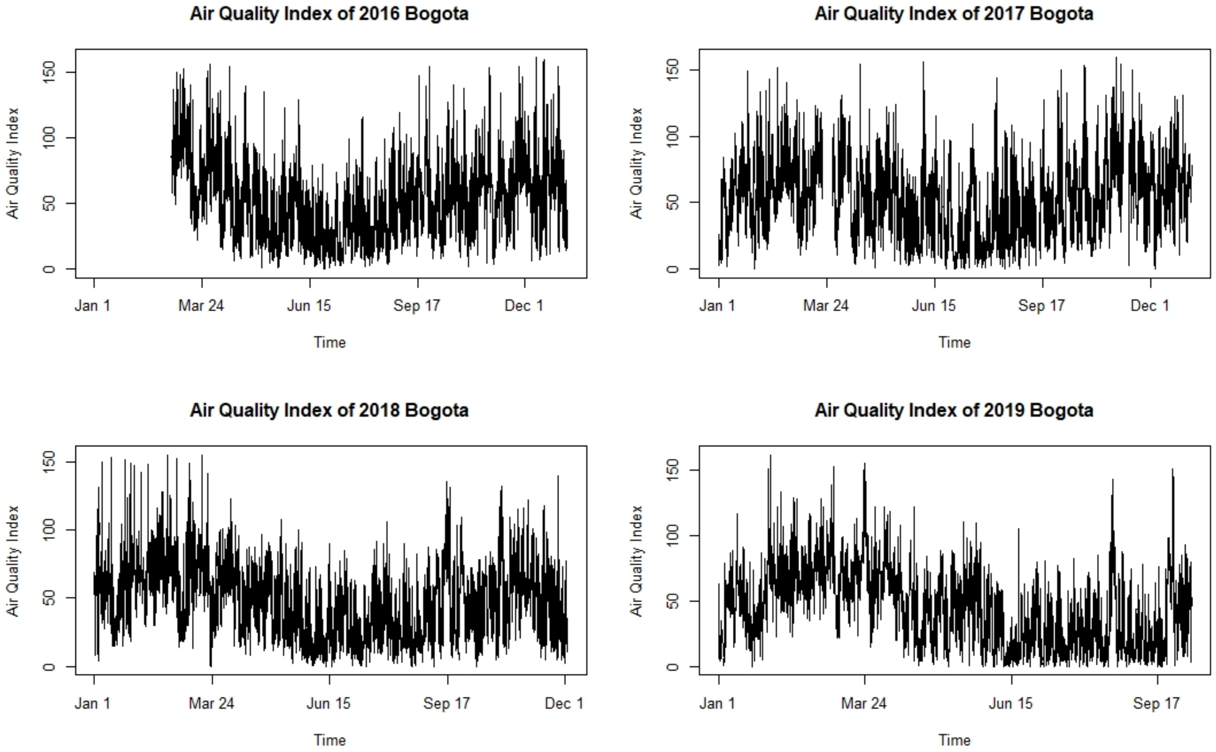


Figure 6.5: Yearly AQI in Bogota

Air Quality Index crossing the threshold of even 200 should be extremely low, if not 0. As we can see in Table 6.4, only Bogota, the second most polluted city in Colombia, achieves a 0% chance of the AQI crossing the value of 200, in either season. In other words, the Air Quality Index in Bogota, is always expected to be below the threshold of 200. Colombo also fares relatively well, with only a 14% chance in Summer to have an AQI above 200. The situation is relatively worse in Lima and Jakarta, with respectively 10% and 28.3% chance of AQI over 200 in Summer, and 44.09% and 13.65% chance in winter. However, all other cities are in terrible condition, with the predictions reaching as high as 100% in Dhaka and New Delhi. In other words, the maximum daily Air Quality Index in Dhaka and New Delhi are almost certain to be unhealthy for the entire population.

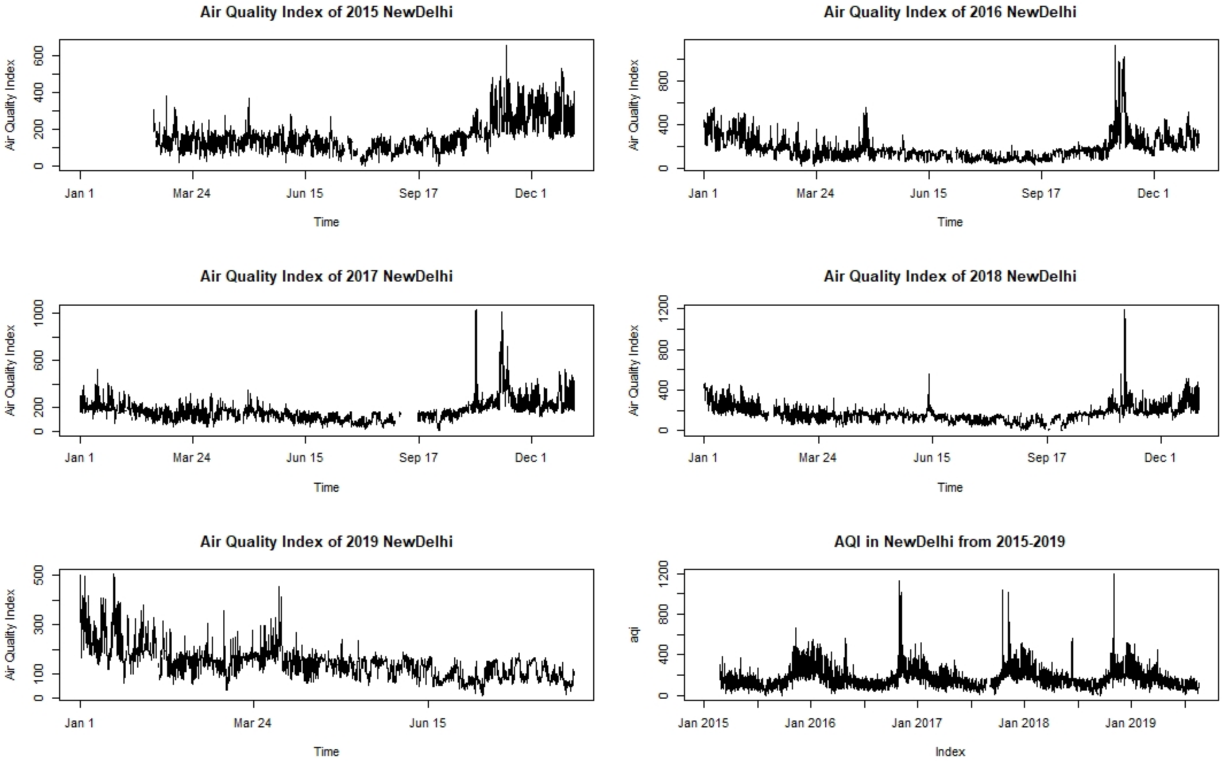


Figure 6.6: Yearly AQI in New Delhi

Table 6.1: Estimate of the Probability of AQI going above 200 in some cities

| City (Country) | Estimated Probability of AQI > 200 in Summer | Estimated Probability of AQI > 200 in Winter |
|---|---|---|
| Lima (Peru) | 10.39% | 44.09% |
| Bogota (Colombia) | 0% | 0% |
| Kampala (Uganda) | 99.06% | 99.85% |
| Manama (Bahrain) | 99.78% | 98.87% |
| Chengdu (China) | 85.18% | 100% |
| Colombo (Sri Lanka) | 14.62% | 0% |
| Dhaka (Bangladesh) | 100% | 100% |
| Jakarta (Indonesia) | 28.3% | 13.65% |
| Chennai (India) | 59.15% | 81.41% |
| Mumbai (India) | 94.26% | 100% |
| Hyderabad (India) | 88.67% | 100% |
| Kolkata (India) | 88.18% | 100% |
| New Delhi (India) | 100% | 100% |

Table 6.2: Estimate of the Probability of AQI going above 300 in some cities

| City (Country) | Estimated Probability of AQI > 300 in Summer | Estimated Probability of AQI > 300 in Winter |
|---|---|---|
| Lima (Peru) | 0% | 0% |
| Bogota (Colombia) | 0% | 0% |
| Kampala (Uganda) | 0% | 21.92% |
| Manama (Bahrain) | 90.94% | 50.59% |
| Chengdu (China) | 0% | 78.82% |
| Colombo (Sri Lanka) | 0% | 0% |
| Dhaka (Bangladesh) | 98.83% | 100% |
| Jakarta (Indonesia) | 0% | 0% |
| Chennai (India) | 36.04% | 39.04% |
| Mumbai (India) | 45.28% | 96.82% |
| Hyderabad (India) | 10% | 78.64% |
| Kolkata (India) | 10.4% | 100% |
| New Delhi (India) | 94.15% | 100% |

Table 6.2 provides the estimates of the probabilities of the AQI going above 300, the hazardous zone according to EPA. Here, we can see only Dhaka and New Delhi show almost 100% chance of having the daily maximum value of AQI to be over 300. Kolkata, Mumbai Hyderabad and Chengdu also display high chance of having AQI values higher than 300. The probability of getting AQI above than 300 in Kolkata changes from 10% in Summer to 100% in Winter. In table 6.3, we can see most of the cities show very small chance of getting an AQI above 500, as it should be. However, even in this case, New Delhi attains an estimate of 97.8% in winter, i.e. there is 97.8% chance that the AQI value in New Delhi would exceed 500 in Winter, which is even beyond the hazardous zone in the EPA regulation. *Dhaka* and *Kolkata* also indicates high probability of air pollution exceeding the hazardous zone during the winter.

It is clear from the tables that cities like New Delhi and Dhaka experience considerable air crisis throughout the year. If we compare only amongst the Indian cities, only Chennai in Summer has a less than 50% chance of having AQI above 200. However, in winter, all

Table 6.3: Estimate of the Probability of AQI going above 500 in some cities

| City (Country) | Estimated Probability of AQI > 500 in Summer | Estimated Probability of AQI > 500 in Winter |
|---|---|---|
| Lima (Peru) | 0% | 0% |
| Bogota (Colombia) | 0% | 0% |
| Kampala (Uganda) | 0% | 0% |
| Manama (Bahrain) | 30.15% | 13% |
| Chengdu (China) | 0% | 0% |
| Colombo (Sri Lanka) | 0% | 0% |
| Dhaka (Bangladesh) | 9.75% | 77.23% |
| Jakarta (Indonesia) | 0% | 0% |
| Chennai (India) | 0% | 27.74% |
| Mumbai (India) | 0% | 34.35% |
| Hyderabad (India) | 0% | 21.72% |
| Kolkata (India) | 0% | 81.2% |
| New Delhi (India) | 32.19% | 97.98% |

Table 6.4: Empirical mean of the extreme value distribution

| City (Country) | Empirical Mean in Summer | Empirical Mean in Winter |
|---|---|---|
| Lima (Peru) | 190.32(9.83) | 210.78(30.12) |
| Bogota (Colombia) | 150.41(6.98) | 156.09(4.25) |
| Kampala (Uganda) | 252.25(26.67) | 279.95(42.18) |
| Manama (Bahrain) | 449.46(106.1) | 364.05(134.79) |
| Chengdu (China) | 228.01(21.15) | 332.79(43.38) |
| Colombo (Sri Lanka) | 172.95(14.62) | 187.19(6.39) |
| Dhaka (Bangladesh) | 422.74(54.54) | 550.95(62.87) |
| Jakarta (Indonesia) | 196.81(4.59) | 188.52(8.25) |
| Chennai (India) | 256.81(83.52) | 342.58(180.37) |
| Mumbai (India) | 312.96(92.47) | 450.58(96.98) |
| Hyderabad (India) | 243.61(53.93) | 434.21(179.03) |
| Kolkata (India) | 237.47(34.29) | 540.86(41.91) |
| New Delhi (India) | 424.01(89.87) | 859.87(235.76) |

the cities exhibit very high probability, with 4 of the 5 cities attaining the 100% level. The probability of AQI going over 300 is also very low in all the cities in Summer, except New Delhi, where there is also a 32.19% chance of AQI going even over 500. However, the situation changes during Winter, when all the cities exhibit high level of air pollution. This might primarily be due to the uncontrolled use of firecrackers, while celebrating the festivals like Diwali. Cities like Chennai, Mumbai and Hyderabad, have relatively lower risk of high air pollution, compared to New Delhi and Kolkata. However, Mumbai experiences air pollution throughout the year, as shown by the high probability of the AQI level going above 300, even in Summer. The situation in New Delhi is worse than any other city in India, and maybe even the world. At present, thick gray smog are seen to engulf the city, generated by a mixture of vehicular emissions, industrial pollution, construction dust and crop burning in neighboring states. Table 6.2 and Table 6.3 shows that Diwali is only part of the problem, and not the only cause for the air pollution problem in Delhi. As we can see, the probability of getting high AQI persists in Delhi even in Summer. This might be owing to the huge number of vehicles in the area, $3,172,842$ as reported in 2017. The huge population of 21.75 million, coupled with poor infrastructure, makes it extremely difficult to control the air pollution problem in New Delhi. In Kolkata, however, the festivals in winter seemed to be primary problem behind high levels of AQI, with a steep jump from 0% to 81.2% chance of having AQI level above 500.

## 6.8 Conclusion

Extreme value analysis is an integral aspect of many real -life applications. The conventional approach to extreme value analysis involve complicated testing and estimation procedures. The $m$ out of $n$ non-parametric Bootstrap, first suggested by [13], provides a easy to use alternative to the extreme value problem. This chapter proposes an optimal choice for the Bootstrap sample size $m$, and the corresponding convergence rate to true distribution. The optimal block size is given by $2.8n^{\frac{1}{3}}$ for distributions belonging in the domain of attraction of Type I (Gumbel), Type II (Fréchet) and Type III (only for $\alpha \le 2$). For

Type III cases with $\alpha > 2$, the optimal choice turns out to be $(2^{\alpha}n)^{\frac{1}{1+\alpha}}$. The corresponding optimal convergence rate is of the order $n^{-\frac{1}{3}}$ for Type I, Type II and Type III ($\alpha \leq 2$), and $n^{-\frac{1}{1+\alpha}}$ for Type III with $\alpha > 2$. The result can also be extended to multivariate set-up, where there would be some additional terms corresponding to the convergence rate of the dependence function. We have also established the cases when the non-parametric Bootstrap performs better than approximation by the limit laws. It is to be noted that even though for some cases, the conventional limit law approximation method might be faster than the non-parametric Bootstrap, there exists myriad complications associated with the conventional procedure. An extension of the work in this chapter would be to derive the convergence rate of Bootstrap in dependent random variables. The $m$ out of $n$ Bootstrap might not be an efficient in those cases, moving block Bootstrap being a better choice in general in presence of dependence within the random variables.

## 6.9 Proof of Theorem 6.3.1

First, we will find an expansion for $(1-\frac{x}{m})^m$.

$$\begin{aligned}(1-\frac{x}{m})^m &= e^{m(-\frac{x}{m}-\frac{x^2}{m^2}+R_3(\frac{x}{m}))}\\ &= e^{-x}\left[e^{-\frac{x^2}{m}+R_3(\frac{x}{m})}\right]\end{aligned}$$

$$\Rightarrow (1-\frac{x}{m})^m = e^{-x}e^{g(x)} \qquad (6.9.1)$$

where, $g(x) = -\frac{x^2}{m} + R_3(\frac{x}{m})$.

From the remainder term of Taylor Series Expansion, we can easily see,

$$R_3(x) = -\frac{x^3(1-\theta)^2}{(1-\theta x)^3}$$

where $\theta \in (0,1)$ is a constant.

Upon calculation, we get:

$$R_3'(x) = -\frac{3(1-\theta)^2 x^2}{(1-\theta x)^4} \tag{6.9.2}$$

$$R_3''(x) = -\frac{1+\theta x}{(1-\theta x)^5} 6x(1-\theta)^2 \tag{6.9.3}$$

Now, since, $g'(0) = 0 = g(0)$, a first order Taylor expansion gives:

$$e^{g(x)} = e^{g(0)} + R_2^g(x) \tag{6.9.4}$$

where $R_2^g(x)$ is the error term. Let $f(x) = e^{g(x)}$. Then, $f''(x) = e^{g(x)\left[g'(x)^2+g''(x)\right]}$. We can see:

$$g'(x) = -\frac{2x}{m} + \frac{1}{m}R_3'(\frac{x}{m}) = -\frac{2x}{m} - \frac{1}{m^3}\left(\frac{3(1-\theta)^2x^2}{(1-\theta\frac{x}{m})^4}\right)$$

$$g''(x) = -\frac{2}{m} + \frac{1}{m^2}R_3''(\frac{x}{m}) = -\frac{2}{m} - \frac{1}{m^3}\left(\frac{1+\theta\frac{x}{m}}{(1-\theta\frac{x}{m})^5}6x(1-\theta)^2\right)$$

Thus,

$$\begin{aligned}
R_2^g(x) &= \int_0^x e^{-t^2/m+R_3(t/m)}\Big[g'(t)^2+g''(t)\Big]dt \\
&= e^{-\frac{\tau^2x^2}{m}+R_3(\frac{\tau x}{m})}\Big[g'(\tau x)^2+g''(\tau x)\Big]x, \qquad \text{where } \tau\in(0,1) \\
&= e^{-\frac{\tau^2x^2}{m}+R_3(\frac{\tau x}{m})}\Bigg[\bigg(-\frac{2\tau x}{m}-\frac{1}{m^3}(\frac{3(1-\theta)^2\tau^2x^2}{(1-\theta\frac{\tau x}{m})^4})\bigg)^2+\bigg(-\frac{2}{m}-\frac{1}{m^3}\bigg(\frac{1+\theta\frac{\tau x}{m}}{(1-\theta\frac{\tau x}{m})^5}6\tau x(1-\theta)^2\bigg)\bigg)\Bigg]
\end{aligned} \tag{6.9.5}$$

Thus, using 6.9.1 and 6.3.1, we obtain:

$$\left(1-\frac{x}{m}\right)^m = e^{-x}\{1+R_2^g(x)\} \tag{6.9.6}$$

Hence,

$$\begin{aligned}
F_n^m(a_m+b_mx) &= \left(1-\frac{m(1-F_n(a_m+b_mx))}{m}\right)^m \\
&= \left(1-\frac{z_m^*}{m}\right)^m, \qquad \text{where } z_m^*=m(1-F_n(a_m+b_mx)) \\
&= e^{-z_m^*}+e^{-z_m^*}R_2^g(z_m^*)
\end{aligned} \tag{6.9.7}$$

Finally, we can write:

$$\begin{aligned} F_n^m(a_m + b_m x) - H(x) &= e^{-z_m^*} + e^{-z_m^*} R_2^g(z_m^*) - H(x) \\ &= e^{-z_m^*} R_2^g(z_m^*) + e^{-z_m^*} - e^{-z_m} + e^{-z_n} - H(x), \qquad \text{where } z_m = m(1 - F(a_m + b_m x)) \\ &= \Xi(z_m^*) + e^{-z_m^*} - e^{-z_m} + e^{-z_m} - H(x), \quad \text{where } \Xi(x) = e^{-x} R_2^g(x) \\ &= \underbrace{\Xi(z_m) + \Xi'(z_m)(z_m^* - z_m) + R_2^{\Xi}(z_m)}_{\zeta_1^*(x,m,n)} + \underbrace{e^{-z_m^*} - e^{-z_m}}_{\zeta_2^*(x,m,n)} + \underbrace{e^{-z_m} - H(x)}_{\zeta_3(x,m,n)} \end{aligned} \tag{6.9.8}$$

First let us consider the terms under $\zeta_1^*(x, m, n)$. 6.9.5 shows us that $\Xi(z_m)$ goes to 0 at a rate of $m^{-1}$.

Now for the term $\Xi'(z_n)$. We can see $R_2^g(x) = e^{f(x)}\left[f_1(x)^2 + f_2(x)\right]$, where:

- $f_1(x) = \left( - \frac{2\tau x}{m} - \frac{1}{m^3}\left(\frac{3(1-\theta)^2\tau^2 x^2}{(1-\theta\frac{\tau x}{m})^4}\right)\right)^2 = \frac{1}{m^2}\left( - 2\tau x - \frac{3\kappa_1^2 x^2}{m^2(1-\kappa_2 x)^4}\right)^2$
- $f_2(x) = \left( - \frac{2}{m} - \frac{1}{m^3}\left(\frac{1+\theta\frac{\tau x}{m}}{(1-\theta\frac{\tau x}{m})^5} 6\tau x(1-\theta)^2\right)\right) = \frac{1}{m}\left( - 2 - \frac{6\kappa_3 x(1+\kappa_2 x)}{m^3(1-\kappa_2 x)^5}\right)$
- $f(x) = -\frac{\tau^2 x^2}{m} + R_3(\frac{\tau x}{m}) = -\frac{\tau^2 x^2}{m} - \frac{\kappa_4 x^3}{m^3}$

In the above equations, $\kappa_1, \kappa_2, \kappa_3, \kappa_4 \in (0, 1)$. Thus,

$$\begin{aligned} \frac{d}{dx} R_2^g(x) &= e^{f(x)}\left[2f_1(x)f_1'(x) + f_2'(x)\right] + e^{f(x)} f'(x)[f_1(x)^2 + f_2(x)] \\ &= e^{f(x)}\Big( f'(x)f_1(x)^2 + f'(x)f_2(x) + 2f_1(x)f_1'(x) + f_2'(x)\Big) \end{aligned} \tag{6.9.9}$$

It can be easily observed that this is of the order $m^{-2}$. Thus, $\Xi'(z_m) = e^{-z_m}\frac{d}{dx}R_2^g(x)\Big|_{x=z_m} - e^{-z_m}R_2^g(z_m)$ is of the order $m^{-1}$.

We cam also see: $e^{-z_m^*} - e^{-z_m} \sim e^{-z_m}(z_m^* - z_m)$ and $e^{-z_m} - H(x) = H(x)\Bigg(e^{-\rho_n(x)} - 1\Bigg)$.

Finally, we have to look into the term $R_2^{\Xi}(z_m)$. We can write, for some $\kappa \in (0,1)$:

$$R_2^{\Xi}(z_m) = \Xi''(\kappa(z_m^* - z_n))(z_m^* - z_m)^2$$

Now,

$$\begin{aligned}\frac{d^2}{dx^2}R_2^g(x) = e^{f(x)}\Bigg(&2f_1(x)f_1'(x)f'(x) + f''(x)f_1(x)^2 + f''(x)f_2(x)\\ &+ f'(x)f_2'(x) + 2f_1'(x)^2 + 2f_1(x)f_1''(x) + f_2''(x)\\ &+ f'(x)^2f_1(x)^2 + f'(x)^2f_2(x) + 2f'(x)f_1(x)f_1'(x) + f'(x)f_2'(x)\Bigg)\end{aligned}$$

We can see this is of the order $m^{-2}$. However, $\Xi''(z_m) = e^{-z_m}\frac{\partial}{\partial^2 x^2}R_2^g(x)\Big|_{x=z_m} -2e^{-z_m}\frac{\partial}{\partial x}R_2^g(x)\Big|_{x=z_m} + e^{-z_m}R_2^g(z_m)$ will again be of order $m^{-1}$. Hence:

$$\zeta_1^*(x,m,n) = \Xi(z_m) + \Xi'(z_m)(z_m^* - z_m) + R_2^{\Xi}(z_m)$$

where:

- $\Xi(z_m)$ is of the order $m^{-1}$.

- $\Xi'(z_m)$ is of the order $m^{-1}$

- $R_2^{\Xi}(z_m) = \Xi''(\kappa(z_m^* - z_m))(z_m^* - z_m)^2$ where $\Xi''(z_m)$ is of the order $m^{-1}$

To get an order of $\zeta_1^*(x,m,n)$, we would also need the rate of $z_m^* - z_m$, which will also give the rate of convergence for $\zeta_2(x,m,n)$.

As we saw earlier, $T_{m,n}(x) = n\left(1 - F_n(a_m + b_m x)\right)$. Thus, $\frac{m}{n}T_{m,n}(x) = m\left(1 - F_n(a_m + b_m x)\right) = z_m^*$.

$$\begin{aligned}
P(|T_{m,n} - np_m| > t) &\leq \frac{V(T_{m,n})}{t^2} = \frac{np_m(1-p_m)}{t^2} \leq \frac{np_m}{t^2} \\
\Rightarrow P(|\frac{m}{n}T_{m,n} - mp_m| > \frac{m}{n}t) &\leq \frac{np_m}{t^2} \\
\Rightarrow P(|\frac{m}{n}T_{m,n} - mp_m| > t) &\leq \frac{m^2 p_m}{nt^2} \sim \frac{mc(x)}{nt^2} \qquad (6.9.10) \\
\Rightarrow |\frac{m}{n}T_{m,n} - mp_m| &= O_p(\sqrt{\frac{m}{n}}) \qquad (6.9.11) \\
\Rightarrow |z_m^* - z_m| &= O_p(\sqrt{\frac{m}{n}}) \qquad (6.9.12)
\end{aligned}$$

We can also get another inequality using the above procedure:

$$\begin{aligned}
P(|T_{m,n} - np_m| > t) &\leq \frac{E(T_{m,n} - np_m)^4}{t^4} = \frac{n\mu_4 + 3n(n-1)\sigma^4}{t^2} \\
&= \frac{np_m(1-p_m)^4 + np_m^4(1-p_m) + 3n(n-1)p_m^2(1-p_m)^2}{t^4} \qquad (6.9.13) \\
\Rightarrow P(|\frac{m}{n}T_{m,n} - mp_m| > \frac{m}{n}t) &\leq \frac{np_m(1-p_m)^4 + np_m^4(1-p_m) + 3n^2p_m^2(1-p_m)^2}{t^4} \\
\Rightarrow P(|\frac{m}{n}T_{m,n} - mp_m| > t) &\leq \frac{m^4p_m(1-p_m)^4 + m^4p_m^4(1-p_m) + 3nm^4p_m^2(1-p_m)^2}{n^3t^4} \\
&\sim \frac{m^3c(x) + c(x) + 3nm^2c(x)^2}{4n^3t^4} \qquad (6.9.14) \\
\Rightarrow P\left(|\frac{m}{n}T_{m,n} - mp_m| > t\right) &\leq \frac{1}{t^4}\left(c_1(x)\frac{m^2}{n^2} + c_2(x)\frac{m^3}{n^3} + c_3(x)\frac{1}{n^3}\right) = \frac{g_{m,n}(x)}{t^4} \text{ (say)} \qquad (6.9.15)
\end{aligned}$$

Now, let $Z = z_m^* - z_m$.

Then,

$$\begin{aligned}
EZ^2 &= \int_0^\infty P(Z^2 > t)dt \\
&= \int_0^z P(Z^2 > t)dt + \int_z^\infty P(Z^2 > t)dt \\
&\leq z + \int_z^\infty \frac{g_{m,n}(x)}{t^2}dt \\
&= z + \frac{g_{m,n}(x)}{z} \leq 2\sqrt{g_{m,n}(x)} \\
\Rightarrow E|z_m^* - z_m|^2 &\leq 2\sqrt{\left(c_1(x)\frac{m^2}{n^2} + c_2(x)\frac{m^3}{n^3} + c_3(x)\frac{1}{n^3}\right)}
\end{aligned}$$

Hence, $E(\hat{z}_n - z_n)^2$ is of the order $O(\frac{m}{n})$.

Additionally, if $X_1, \ldots, X_n$ are independent Bernoulli random variables with probability $p$, then for $p \leq \frac{1}{4}$, we know (using Chernoff's tightness bounds):

$$P\left(\left(\sum_{i=1}^{n} X_1 - \mu\right) > t\right) \geq \frac{1}{4} e^{-\frac{2t^2}{\mu}}$$

where $\mu = np$ (The proof involves use of Slud's inequality). Using this, We can show that the lower bound also depends on the rate of $\sqrt{\frac{m}{n}}$.

$$P\left((T_{m,n} - np_m) > t\right) \geq \frac{1}{4} e^{-\frac{2t^2}{np_m}}$$
$$\Rightarrow P\left(\left(\frac{m}{n} T_{m,n} - mp_m\right) > t\right) \geq \frac{1}{4} e^{-\frac{2nt^2}{m^2 p_m}} \sim \frac{1}{4} exp\left(-\left(\sqrt{\frac{2}{c(x)} \frac{n}{m}} t\right)^2\right)$$

Now, we can declare the rate of convergence:

- $\zeta_1^*(x, m, n)$: This term contains three parts:
  - $\Xi(z_m)$ is of the order $m^{-1}$.
  - $\Xi'(z_m)(z_m^* - z_m)$ is of the order $\frac{1}{m}\sqrt{\frac{m}{n}} = \sqrt{\frac{1}{mn}}$
  - $R_2'(z_m) = \Xi''(\kappa(z_m^* - z_m))(z_m^* - z_m)^2$ is of the order $\frac{1}{m}\sqrt{\frac{m}{n}}\frac{m}{n} = \sqrt{\frac{m}{n^3}}$
- $\zeta_2^*(x, m, n)$ is of the order $O_p(\sqrt{\frac{m}{n}})$
- $\zeta_3^*(x, m, n)$ is of the order A(m).

Hence, finally, the convergence rate of non-parametric Bootstrap for extremes of random variables is given by $O_p(\max(\frac{1}{m}, \sqrt{\frac{m}{n}}, A(m)))$.

## 6.10 Appendix section

### 6.10.1 Pareto Distribution

The distribution function of Pareto($\alpha$):

$$F(x) = 1 - x^{-\alpha}, \qquad \text{for } x > 1$$

We can get $f = (\frac{1}{-logF})^{\leftarrow}$ from this. The function f can be easily obtained as follows:

$$\begin{aligned} f(x) &= \inf\left\{ y \middle| F(y) \geq e^{-\frac{1}{x}} \right\} \\ &= \inf\left\{ y \middle| y^{\alpha} \geq \frac{1}{1 - e^{-\frac{1}{x}}} \right\} \\ &= (1 - e^{-\frac{1}{x}})^{-\frac{1}{\alpha}} \end{aligned}$$

Hence,

$$\begin{aligned} f'(x) &= \frac{e^{-\frac{1}{x}}}{1 - e^{-\frac{1}{x}}} \frac{1}{\alpha x^2} f(x) \\ f''(x) &= f'(x) \left( \frac{e^{-\frac{1}{x}}}{1 - e^{-\frac{1}{x}}} \frac{1}{\alpha x^2} - \frac{2}{x} + \frac{1}{x^2(1 - e^{-\frac{1}{x}})} \right) \\ \Rightarrow A(t) = \frac{tf''(t)}{f'(t)} - \gamma + 1 &= \frac{1}{t(1 - e^{-\frac{1}{t}})} \left( 1 + \frac{e^{-\frac{1}{t}}}{\alpha} \right) - 1 - \frac{1}{\alpha} \end{aligned}$$

Expanding $e^{-\frac{1}{x}}$, we get,

$$x(1 - e^{-\frac{1}{x}}) = x\left( \frac{1}{x} - \frac{1}{2x^2} + \frac{1}{6x^3} + \cdots \right) = 1 - \frac{1}{2x} + \frac{1}{6x^2} + \cdots$$

Thus, the rate of convergence of the extreme for Pareto distribution is $O(A(n)) = O(n^{-1})$.

### 6.10.2 Beta Distribution

Here, we are considering the distribution $Beta(1, \beta)$, with density $f(x) = \beta(1-x)^{\beta-1}$ for $x \in (0,1)$. The distribution function can be written as:

$$F(x) = 1 - (1-x)^{\beta}$$

The function $f$ then takes the form:

$$\begin{aligned} f(x) &= \inf\left\{ y \middle| F(y) \geq e^{-\frac{1}{x}} \right\} \\ &= 1 - \left(1 - e^{-\frac{1}{x}}\right)^{\frac{1}{\beta}} \\ \Rightarrow f'(x) &= \frac{e^{-\frac{1}{x}}}{1-e^{-\frac{1}{x}}} \frac{1}{\beta x^2} \left(1 - f(x)\right) \\ f''(x) &= f'(x)\left( -\frac{e^{-\frac{1}{x}}}{1-e^{-\frac{1}{x}}} \frac{1}{\beta x^2} - \frac{2}{x} + \frac{1}{x^2(1-e^{-\frac{1}{x}})} \right) \end{aligned}$$

This gives a similar $A(t)$ as for Pareto, which would give the rate of convergence to be $O(n^{-1})$.

### 6.10.3 Skewed Normal

[80] calculated the rate of convergence of skewed normal distribution. The density of $SN(\lambda)$ is:

$$f_\lambda(x) = 2\phi(x)\Phi(\lambda x)$$

It was shown that $\frac{1-F_\lambda(x)}{f_\lambda(x)} \sim \frac{1}{x}$ for $\lambda > 0$ and $\sim \frac{1}{(1+\lambda^2)x}$ for $\lambda < 0$. The normalizing constants for $\lambda > 0$ and $< 0$ are as follows:

- $\lambda > 0$:
  - $a_n = (2\log n)^{-\frac{1}{2}}$
  - $b_n = (2\log n)^{\frac{1}{2}} - \frac{\log\log n + \log\pi}{2(2\log n)^{\frac{1}{2}}}$
- $\lambda < 0$:
  - $a_n = (1+\lambda^2)^{-\frac{1}{2}} (2\log n)^{-\frac{1}{2}}$
  - $b_n = \left(\frac{2\log n}{1+\lambda^2}\right)^{\frac{1}{2}} - \frac{\log\log n + \log\pi}{(1+\lambda^2)^{\frac{1}{2}} (2\log n)^{\frac{1}{2}}}$

For the above normalizing constants, [80] showed that:

- For $\lambda > 0$,
$$F_\lambda^n(a_n + b_n x) - \Lambda(x) \sim \frac{\Lambda(x)e^{-x}}{16} \frac{(\log\log n)^2}{\log n}$$
- For $\lambda < 0$,
$$F_\lambda^n(a_n + b_n x) - \Lambda(x) \sim \frac{\Lambda(x)e^{-x}}{4} \frac{(\log\log n)^2}{\log n}$$

Hence the rate of convergence for extreme of skewed normal distribution is $O(\frac{1}{\log n})$, just as for standard normal distribution.

# References

[1] HI Abdel-Rahman and BA Marzouk. "Statistical method to predict the sunspots number." In: *NRIAG Journal of Astronomy and Geophysics* 7.2 (2018), pp. 175–179.

[2] Abhay Abhyankar, Laurence S Copeland, and Woon Wong. "Uncovering nonlinear structure in real-time stock-market indexes: the S&P 500, the DAX, the Nikkei 225, and the FTSE-100." In: *Journal of Business & Economic Statistics* 15.1 (1997), pp. 1–14.

[3] L.V. Ahlfors. *Complex Analysis.* McGraw-Hill, New York, 1979.

[4] Saud Alashri, Srinivasa Srivatsav Kandala, Vikash Bajaj, Roopek Ravi, Kendra L Smith, and Kevin C Desouza. "An analysis of sentiments on facebook during the 2016 US presidential election." In: *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).* 2016, pp. 795–802.

[5] Saud Alashri, Srinivasa Srivatsav Kandala, Vikash Bajaj, Emily Parriott, Yukika Awazu, and Kevin C Desouza. "The 2016 US Presidential Election on Facebook: an exploratory analysis of sentiments." In: *Proceedings of the 51st Hawaii International Conference on System Sciences.* 2018, pp. 1771–1780.

[6] John E Angus. "Asymptotic theory for bootstrapping the extremes." In: *Communications in Statistics-Theory and Methods* 22.1 (1992), pp. 15–30.

[7] KB Athreya et al. "Bootstrap of the mean in the infinite variance case." In: *The Annals of Statistics* 15.2 (1987), pp. 724–731.

[8] Adrian G Barnett and Rodney C Wolff. "A time-domain test for some types of nonlinearity." In: *IEEE Transactions on Signal Processing* 53.1 (2004), pp. 26–33.

[9] William Bell. "Signal extraction for nonstationary time series." In: *The Annals of Statistics* 12.2 (1984), pp. 646–664.

[10] Arthur Berg. "Multivariate lag-windows and group representations." In: *Journal of Multivariate Analysis* 99.10 (2008), pp. 2479–2496.

[11] Arthur Berg, Efstathios Paparoditis, and Dimitris N Politis. "A bootstrap test for time series linearity." In: *Journal of Statistical Planning and Inference* 140.12 (2010), pp. 3841–3857.

[12] Arthur Berg and Dimitris N Politis. “Higher-order accurate polyspectral estimation with flat-top lag-windows.” In: *Annals of the Institute of Statistical Mathematics* 61.2 (2009), pp. 477–498.

[13] Peter J Bickel, David A Freedman, et al. “Some asymptotic theory for the bootstrap.” In: *The annals of statistics* 9.6 (1981), pp. 1196–1217.

[14] Peter J Bickel, Friedrich Götze, and Willem R van Zwet. “Resampling fewer than n observations: gains, losses, and remedies for losses.” In: *Selected works of Willem van Zwet*. Springer, 2012, pp. 267–297.

[15] Peter J Bickel and Anat Sakov. “On the choice of m in the m out of n bootstrap and confidence bounds for extrema.” In: *Statistica Sinica* 18.3 (2008), pp. 967–985.

[16] T. Bollerslev. “Generalized autoregressive conditional heteroskedasticity.” In: *Journal of Econometrics* 31.3 (1986), pp. 307–327.

[17] David R Brillinger. “Asymptotic properties of spectral estimates of second order.” In: *Selected Works of David Brillinger*. Springer, 2012, pp. 179–194.

[18] David R Brillinger. *Time series: data analysis and theory.* Vol. 36. Siam, 1981.

[19] David R Brillinger. *Time series: data analysis and theory.* SIAM, 2001.

[20] David R Brillinger and Murray Rosenblatt. “Asymptotic theory of estimates of kth-order spectra.” In: *Proceedings of the National Academy of Sciences of the United States of America* 57.2 (1967), p. 206.

[21] David R. Brillinger. “An introduction to polyspectra.” In: *The Annals of mathematical statistics* (1965), pp. 1351–1374.

[22] David R. Brillinger. “The identification of polynomial systems by means of higher order spectra.” In: *Journal of Sound and Vibration* 12.3 (1970), pp. 301–313.

[23] Patrick L. Brockett, Melvin J. Hinich, and Douglas Patterson. “Bispectral-based tests for the detection of Gaussianity and linearity in time series.” In: *Journal of the American Statistical Association* 83.403 (1988), pp. 657–664.

[24] Peter J Brockwell and Richard A Davis. *Introduction to time series and forecasting.* springer, 2016.

[25] Peter J. Brockwell, Richard A. Davis, and Stephen E Fienberg. *Time Series: Theory and Methods: Theory and Methods*. Springer Science & Business Media, 1991.

[26] Widodo Budiharto and Meiliana Meiliana. “Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis.” In: *Journal of Big data* 5.51 (2018), pp. 1–10.

[27] Pete Burnap, Rachel Gibson, Luke Sloan, Rosalynd Southern, and Matthew Williams. “140 characters to victory?: Using Twitter to predict the UK 2015 General Election.” In: *Electoral Studies* 41 (2016), pp. 230–233.

[28] Jennifer L. Castle and David F. Hendry. “A low-dimension portmanteau test for non-linearity.” In: *Journal of Econometrics* 158.2 (2010), pp. 231–245.

[29] WS Chan and Howell Tong. “On tests for non-linearity in time series analysis.” In: *Journal of forecasting* 5.4 (1986), pp. 217–228.

[30] Norbert Christopeit. “Estimating parameters of an extreme value distribution by the method of moments.” In: *Journal of Statistical Planning and Inference* 41.2 (1994), pp. 173–186.

[31] Kuang Chua Chua, Vinod Chandran, U Rajendra Acharya, and Choo Min Lim. “Application of higher order statistics/spectra in biomedical signals—A review.” In: *Medical engineering & physics* 32.7 (2010), pp. 679–689.

[32] Mooi Choo Chuah and Fen Fu. “ECG anomaly detection via time series analysis.” In: *International Symposium on Parallel and Distributed Processing and Applications*. Springer. 2007, pp. 123–135.

[33] Michael Clements and David Hendry. *Forecasting economic time series*. Cambridge University Press, 1998.

[34] Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[35] Michael Connell and Sarah Vogler. *Russia’s Approach to Cyber Warfare (1Rev)*. Tech. rep. Center for Naval Analyses Arlington United States, 2017.

[36] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. “Fame for sale: Efficient detection of fake Twitter followers.” In: *Decision Support Systems* 80 (2015), pp. 56–71.

[37] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. “Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling.” In: *IEEE Transactions on Dependable and Secure Computing* 15.4 (2017), pp. 561–576.

[38] Rainer Dahlhaus. “Asymptotic normality of spectral estimates.” In: *Journal of Multivariate Analysis* 16.3 (1985), pp. 412–431.

[39] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. “Botornot: A system to evaluate social bots.” In: *Proceedings of the 25th international conference companion on world wide web*. 2016, pp. 273–274.

[40] Laurens De Haan and Ana Ferreira. *Extreme value theory: an introduction*. Springer Science & Business Media, 2007.

[41] Laurens De Haan, Sidney Resnick, et al. “Second-order regular variation and rates of convergence in extreme-value theory.” In: *The Annals of Probability* 24.1 (1996), pp. 97–124.

[42] Laurentius De Haan and Maria Franciscus. *On regular variation and its application to the weak convergence of sample extremes*. 1970.

[43] Chirag Deb, Fan Zhang, Junjing Yang, Siew Eang Lee, and Kwok Wei Shah. “A review on time series forecasting techniques for building energy consumption.” In: *Renewable and Sustainable Energy Reviews* 74 (2017), pp. 902–924.

[44] Shlomo Dubnov. *Polyspectral analysis of musical timbre.* Citeseer, 1996.

[45] B. Efron. “Bootstrap Methods: Another Look at the Jackknife.” In: *Ann. Statist.* 7.1 (Jan. 1979), pp. 1–26. DOI: 10.1214/aos/1176344552.

[46] Jianqing Fan and Qiwei Yao. *Nonlinear time series: nonparametric and parametric methods.* Springer Science & Business Media, 2008.

[47] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. “The rise of social bots.” In: *Communications of the ACM* 59.7 (2016), pp. 96–104.

[48] Linton C Freeman. “Centrality in social networks conceptual clarification.” In: *Social networks* 1.3 (1978), pp. 215–239.

[49] Atle Fretheim, Stephen B Soumerai, Fang Zhang, Andrew D Oxman, and Dennis Ross-Degnan. “Interrupted time-series analysis yielded an effect estimate concordant with the cluster-randomized controlled trial result.” In: *Journal of clinical epidemiology* 66.8 (2013), pp. 883–887.

[50] Jun-ichiro Fukuchi. “Bootstrapping extremes of random variables.” In: *Digital Repository@ Iowa State University, http://lib. dr. iastate. edu/* (1994).

[51] M.M. Gabr. “On the third-order moment structure and bispectral analysis of some bilinear time series.” In: *Journal of Time Series Analysis* 9.1 (1988), pp. 11–20.

[52] Janos Galambos. “Extreme value theory for applications.” In: *Extreme value theory and applications.* Springer, 1994, pp. 1–14.

[53] Daniel Gayo-Avello, Panagiotis Takis Metaxas, and Eni Mustafaraj. “Limits of electoral predictions using twitter.” In: *Fifth International AAAI Conference on Weblogs and Social Media.* 2011.

[54] Manfred Gilli et al. “An application of extreme value theory for measuring financial risk.” In: *Computational Economics* 27.2-3 (2006), pp. 207–228.

[55] Boris Gnedenko. “Sur la distribution limite du terme maximum d’une serie aleatoire.” In: *Annals of mathematics* (1943), pp. 423–453.

[56] Clive William John Granger and Paul Newbold. *Forecasting economic time series.* Academic press, 2014.

[57] Purva Grover, Arpan Kumar Kar, Yogesh K Dwivedi, and Marijn Janssen. “Polarization and acculturation in US Election 2016 outcomes–Can twitter analytics predict changes in voting preferences.” In: *Technological Forecasting and Social Change* 145 (2019), pp. 438–460.

[58] Robert Harris. “An application of extreme value theory to reliability theory.” In: *The Annals of Mathematical Statistics* 41.5 (1970), pp. 1456–1465.

[59] Campbell R. Harvey and Akhtar Siddique. “Conditional skewness in asset pricing tests.” In: *The Journal of Finance* 55.3 (2000), pp. 1263–1295.

[60] Michael V Hayden. *Playing to the edge: American intelligence in the age of terror*. Penguin, 2017.

[61] Todd C Helmus, Elizabeth Bodine-Baron, Andrew Radin, Madeline Magnuson, Joshua Mendelsohn, William Marcellino, Andriy Bega, and Zev Winkelman. *Russian social media influence: Understanding Russian propaganda in Eastern Europe*. Rand Corporation, 2018.

[62] Bruce M Hill. "A simple general approach to inference about the tail of a distribution." In: *The annals of statistics* (1975), pp. 1163–1174.

[63] Melvin J Hinich and Gary R Wilson. "Detection of non-Gaussian signals in non-Gaussian noise using the bispectrum." In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 38.7 (1990), pp. 1126–1131.

[64] Melvin J. Hinich and Douglas M. Patterson. "Identification of the coefficients in a non-linear: time series of the quadratic type." In: *Journal of Econometrics* 30.1-2 (1985), pp. 269–288.

[65] Vinay K Jain and Shishir Kumar. "Towards prediction of election outcomes using social media." In: *International Journal of Intelligent Systems and Applications* 20 (2017), pp. 20–28.

[66] Phillip A Jang and David S Matteson. "Spatial correlation in weather forecast accuracy: A functional time series approach." In: *arXiv preprint arXiv:2111.11381* (2021).

[67] Ryan Janicki and Tucker S. McElroy. "Hermite expansion and estimation of monotonic transformations of Gaussian data." In: *Journal of Nonparametric Statistics* 28.1 (2016), pp. 207–234.

[68] Kevin Judd and Alistair Mees. "On selecting models for nonlinear time series." In: *Physica D: Nonlinear Phenomena* 82.4 (1995), pp. 426–444.

[69] Holger Kantz and Thomas Schreiber. *Nonlinear time series analysis*. Vol. 7. Cambridge university press, 2004.

[70] Alec N. Kercheval and Yang Liu. "Risk forecasting with garch, skewed t distributions, and multiple timescales." In: *Handbook of Modeling High-Frequency Data in Finance* 4 (2011), p. 163.

[71] Robert R Kinnison. *Applied extreme value statistics*. Battelle Press Columbus, OH, 1985.

[72] L Knopoff and Y Kagan. "Analysis of the theory of extremes as applied to earthquake problems." In: *Journal of Geophysical Research* 82.36 (1977), pp. 5647–5657.

[73] Anders B. Kock and Timo Teräsvirta. "Forecasting with nonlinear time series models." In: *Oxford handbook of economic forecasting* (2011), pp. 61–87.

[74] S. Kotz, T. J. Kozubowski, and K. Podgórski. "Asymmetric multivariate Laplace distribution." In: *The Laplace distribution and generalizations*. Springer, 2001, pp. 239–272.

[75] Martin Kragh and Sebastian Åsberg. "Russia's strategy for influence through public diplomacy and active measures: the Swedish case." In: *Journal of Strategic Studies* 40.6 (2017), pp. 773–816.

[76] Hans-Martin Krolzig and David F Hendry. "Computer automation of general-to-specific model selection procedures." In: *Journal of Economic Dynamics and Control* 25.6-7 (2001), pp. 831–866.

[77] George Kyrkilis, Arhontoula Chaloulakou, and Pavlos A Kassomenos. "Development of an aggregate Air Quality Index for an urban Mediterranean agglomeration: Relation to potential health effects." In: *Environment International* 33.5 (2007), pp. 670–676.

[78] Blake LeBaron, W Brian Arthur, and Richard Palmer. "Time series properties of an artificial stock market." In: *Journal of Economic Dynamics and control* 23.9-10 (1999), pp. 1487–1516.

[79] Kyumin Lee, Brian Eoff, and James Caverlee. "Seven months with the devils: A long-term study of content polluters on twitter." In: *Proceedings of the international AAAI conference on web and social media*. Vol. 5. 1. 2011, pp. 185–192.

[80] Xin Liao, Zuoxiang Peng, Saralees Nadarajah, and Xiaoqian Wang. "Rates of convergence of extremes from skew-normal samples." In: *Statistics & Probability Letters* 84 (2014), pp. 40–47.

[81] K.S. Lii and M. Rosenblatt. "Asymptotic normality of cumulant spectral estimates." In: *Journal of Theoretical Probability* 3.2 (1990), pp. 367–385.

[82] Jian Liu. "On stationarity and asymptotic inference of bilinear time series models." In: *Statistica Sinica* (1992), pp. 479–494.

[83] Jian Liu and Peter J. Brockwell. "On the general bilinear time series model." In: *Journal of Applied Probability* 25.3 (1988), pp. 553–564.

[84] Elizabeth Ann Maharaj and Andrés M Alonso. "Discriminant analysis of multivariate time series: Application to diagnosis based on ECG signals." In: *Computational Statistics & Data Analysis* 70 (2014), pp. 67–87.

[85] Agustin Maravall. "An application of nonlinear time series forecasting." In: *Journal of Business & Economic Statistics* 1.1 (1983), pp. 66–74.

[86] Massimo Marchiori and Vito Latora. "Harmony in the small-world." In: *Physica A: Statistical Mechanics and its Applications* 285.3-4 (2000), pp. 539–546.

[87] Michele Mazza, Stefano Cresci, Marco Avvenuti, Walter Quattrociocchi, and Maurizio Tesconi. "Rtbust: Exploiting temporal patterns for botnet detection on twitter." In: *Proceedings of the 10th ACM conference on web science*. 2019, pp. 183–192.

[88] T.S. McElroy, D. Ghosh, and S. Lahiri. "Quadratic prediction of time series via autocumulants." In: *Mimeo* (2022).

[89] Tucker McElroy. “Recursive Computation for Block-Nested Covariance Matrices.” In: *Journal of Time Series Analysis* 39.3 (2018), pp. 299–312.

[90] Tucker S McElroy and Dimitris N Politis. *Time Series: A First Course with Bootstrap Starter.* CRC Press, 2020.

[91] Jerry M Mendel. “Tutorial on higher-order statistics (spectra) in signal processing and system theory: Theoretical results and some applications.” In: *Proceedings of the IEEE* 79.3 (1991), pp. 278–305.

[92] Willi Mutschler. “Higher-order statistics for DSGE models.” In: *Econometrics and statistics* 6 (2018), pp. 44–56.

[93] Diana C Mutz. “Status threat, not economic hardship, explains the 2016 presidential vote.” In: *Proceedings of the National Academy of Sciences* 115.19 (2018), E4330–E4339.

[94] Vidya Narayanan, Philip N Howard, Bence Kollanyi, and Mona Elswah. “Russian involvement and junk news during Brexit.” In: *The computational propaganda project. Algorithms, automation and digital politics. https://comprop. oii. ox. ac. uk/research/working-papers/russia-and-brexit* (2017).

[95] Chrysostomos L. Nikias and Jerry M. Mendel. “Signal processing with higher-order spectra.” In: *IEEE Signal processing magazine* 10.3 (1993), pp. 10–37.

[96] Chrysostomos L. Nikias and Mysore R. Raghuveer. “Bispectrum estimation: A digital signal processing framework.” In: *Proceedings of the IEEE* 75.7 (1987), pp. 869–891.

[97] E Omey and ST Rachev. “Rates of convergence in multivariate extreme value theory.” In: *Journal of multivariate analysis* 38.1 (1991), pp. 36–50.

[98] Alan V. Oppenheim and Jae S. Lim. “The importance of phase in signals.” In: *Proceedings of the IEEE* 69.5 (1981), pp. 529–541.

[99] Bernard Picinbono. “Polyspectra of ordered signals.” In: *IEEE Transactions on Information Theory* 45.7 (1999), pp. 2239–2252.

[100] James Pickands III et al. “Statistical inference using extreme order statistics.” In: *the Annals of Statistics* 3.1 (1975), pp. 119–131.

[101] Simon Potter. “Nonlinear time series modelling: An introduction.” In: *Journal of Economic Surveys* 13.5 (1999), pp. 505–528.

[102] BG Quinn. “Stationarity and invertibility of simple bilinear models.” In: *Stochastic Processes and their Applications* 12.2 (1982), pp. 225–230.

[103] Mysore R Raghuveer and Chrysostomos L Nikias. “Bispectrum estimation via AR modeling.” In: *Signal Processing* 10.1 (1986), pp. 35–48.

[104] James B. Ramsey and Philip Rothman. “Time irreversibility and business cycle asymmetry.” In: *Journal of Money, Credit and Banking* 28.1 (1996), pp. 1–21.

[105] M. Bhaskara Rao, T. Subba Rao, and A.M. Walker. "On the existence of strictly stationary solutions to bilinear equations." In: *J. Time Series Anal* 4 (1983), pp. 95–l.

[106] T Subba Rao and MM Gabr. "A test for linearity of stationary time series." In: *Journal of time series analysis* 1.2 (1980), pp. 145–158.

[107] T. Subba Rao. "On the theory of bilinear time series models." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 43.2 (1981), pp. 244–255.

[108] T. Subba Rao and M.M. Gabr. *An introduction to bispectral analysis and bilinear time series models*. Vol. 24. Springer Science & Business Media, 2012.

[109] Benjamin Renard and Michel Lang. "Use of a Gaussian copula for multivariate extreme value analysis: some case studies in hydrology." In: *Advances in Water Resources* 30.4 (2007), pp. 897–912.

[110] Marco Rocco. "Extreme value theory in finance: A survey." In: *Journal of Economic Surveys* 28.1 (2014), pp. 82–108.

[111] Murray Rosenblatt and John W. Van Ness. "Estimation of the bispectrum." In: *The Annals of Mathematical Statistics* 36.4 (1965), pp. 1120–1136.

[112] Jim Rutenberg. "RT, Sputnik and Russia's new theory of war." In: *New York Times* 13 (2017).

[113] Ali Saeb. "General extreme value modeling and application of bootstrap on rainfall data-A case study." In: *arXiv preprint arXiv:1402.0944* (2014).

[114] Vidhya Vasantrao Sambare and Arvind Jain. "The Application of Weather Forecast using Time Series Analysis." In: (2020).

[115] Mohsen Sayyadiharikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. "Detection of novel social bots by ensembles of specialized classifiers." In: *Proceedings of the 29th ACM international conference on information & knowledge management*. 2020, pp. 2725–2732.

[116] Dan Schill and John Allen Hendricks. "THE PRESIDENCY AND SOCIAL MEDIA." In: (2017).

[117] Johan Segers et al. "Asymptotics of empirical copula processes under non-restrictive smoothness assumptions." In: *Bernoulli* 18.3 (2012), pp. 764–782.

[118] A. N. Shiryaev. "Some problems in the spectral theory of higher-order moments. I." In: *Theory of Probability & Its Applications* 5.3 (1960), pp. 265–284.

[119] Robert H Shumway and David S Stoffer. *Time series analysis and its applications (Springer texts in statistics)*. 2005.

[120] SILSO World Data Center. "The International Sunspot Number." In: *International Sunspot Number Monthly Bulletin and online catalogue* (Jan. 1749-2021).

[121] CM Spooner and WA Gardner. “Estimation of cyclic polyspectra.” In: *[1991] Conference Record of the Twenty-Fifth Asilomar Conference on Signals, Systems & Computers.* IEEE. 1991, pp. 370–376.

[122] Venkatramanan S Subrahmanian et al. “The DARPA Twitter bot challenge.” In: *Computer* 49.6 (2016), pp. 38–46.

[123] Wiparat Suwanwaiphatthana, Kannika Ruangdej, and Anne Turner-Henson. “Outdoor air pollution and children’s health.” In: *Pediatric Nursing* 36.1 (2010), p. 25.

[124] Robert CR Swanson. “Two Case Studies of Russian Propaganda in Romania and Hungary.” In: *International Affairs Review* 1 (2018), pp. 1–23.

[125] Sumei Tang, Eliyathamby Antony Selvanathan, and Saroja Selvanathan. “Foreign direct investment, domestic investment and economic growth in China: A time series analysis.” In: *World Economy* 31.10 (2008), pp. 1292–1309.

[126] A. Murat Tekalp and A. Tanju Erdem. “Higher-order spectrum factorization in one and two dimensions with applications in signal modeling and nonminimum phase system identification.” In: *IEEE Transactions on Acoustics Speech and Signal Processing* 37.10 (1989), pp. 1537–1549.

[127] Timo Teräsvirta, Dag Tjøstheim, Clive William John Granger, et al. *Modelling nonlinear economic time series.* Oxford University Press Oxford, 2010.

[128] György Terdik. *Bilinear stochastic models and related problems of nonlinear time series analysis: a frequency domain approach.* Vol. 142. Springer Science & Business Media, 2012.

[129] Maud Thomas, Magali Lemaitre, Mark L Wilson, Cécile Viboud, Youri Yordanov, Hans Wackernagel, and Fabrice Carrat. “Applications of extreme value theory in public health.” In: *PloS one* 11.7 (2016), e0159312.

[130] Dag Tjøstheim. “Non-linear time series: a selective review.” In: *Scandinavian Journal of Statistics* (1994), pp. 97–130.

[131] Gwladys Toulemonde, Pierre Ribereau, and Philippe Naveau. “Applications of Extreme Value Theory to Environmental Data Analysis.” In: *Observations, Modeling, and Economics* (2016), p. 9.

[132] A. Alexandre Trindade, Yun Zhu, and Beth Andrews. “Time series models with asymmetric Laplace innovations.” In: *Journal of Statistical Computation and Simulation* 80.12 (2010), pp. 1317–1333.

[133] Adam Tsakalidis, Symeon Papadopoulos, Alexandra I Cristea, and Yiannis Kompatsiaris. “Predicting elections for multiple countries using Twitter and polls.” In: *IEEE Intelligent Systems* 30.2 (2015), pp. 10–17.

[134] Ruey S. Tsay. “Nonlinearity tests for time series.” In: *Biometrika* 73.2 (1986), pp. 461–466.

[135] Hideatsu Tsukahara. “Empirical copulas and some applications.” In: *Seijo University: Tokyo, Japan* (2000).

[136] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. “Predicting elections with twitter: What 140 characters reveal about political sentiment.” In: *Fourth international AAAI conference on weblogs and social media.* 2010.

[137] John W Van Ness. “Asymptotic normality of bispectral estimates.” In: *The Annals of Mathematical Statistics* (1966), pp. 1257–1272.

[138] Vito Volterra. *Theory of functionals and of integral and integro-differential equations.* Courier Corporation, 2005.

[139] Valeriu Vrabie, Pierre Granjon, and Christine Serviere. “Spectral kurtosis: from definition to application.” In: *6th IEEE international workshop on Nonlinear Signal and Image Processing (NSIP 2003).* 2003, p. xx.

[140] Clint Watts. *Messing with the enemy: Surviving in a social media world of hackers, terrorists, Russians, and fake news.* Harper Business, 2018.

[141] Peter Whittle. *Hypothesis testing in time series analysis.* Vol. 4. Almqvist & Wiksells boktr., 1951.

[142] Stefan Wojcik and Adam Hughes. “Sizing up Twitter users.” In: *Pew Research Center* 24 (2019), pp. 1–23.

[143] Laura Wrubel, Justin Littman, and Dan Kerchner. “2018 U.S. Congressional Election Tweet Ids.” Version V1. In: *Harvard Dataverse.* (2019), DOI : 10.7910/DVN/AEZPLU.

[144] Hsuan-Chia Yang, Shu-Hao Chang, Richard Lu, and Der-Ming Liou. “The effect of particulate matter size on cardiovascular health in Taipei Basin, Taiwan.” In: *Computer methods and programs in biomedicine* 137 (2016), pp. 261–268.