DISSERTATION

TAIL DEPENDENCE: APPLICATION, EXPLORATION, AND DEVELOPMENT OF NOVEL METHODS

Submitted by

Troy P. Wixson

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2025

Doctoral Committee:

Advisor: Daniel Cooley
Co-Advisor: Benjamin Shaby

Dongzhou Huang
Tianying Wang
Elizabeth Barnes

ABSTRACT

TAIL DEPENDENCE: APPLICATION, EXPLORATION, AND DEVELOPMENT OF NOVEL METHODS

The study of multivariate extreme events is largely concerned with modeling the dependence in the tail of the joint distribution. The understanding of extremal dependence and methodology for modeling that dependence has been an active research field over the past few decades and we contribute to that literature with three projects that are detailed in this dissertation.

In the first project we consider the challenge of assessing the changing risk of wildfires. Wildfire risk is greatest during high winds after sustained periods of dry and hot conditions. This chapter is a statistical extreme event risk attribution study which aims to answer whether extreme wildfire seasons are more likely now than under past climate. This requires modeling temporal dependence at extreme levels. We propose the use of transformed-linear time series models which are constructed similarly to traditional ARMA models while having a dependence structure that is tied to a widely used framework for extremes (regular variation). We fit the models to the extreme values of the seasonally adjusted Fire Weather Index (FWI) time series to capture the dependence in the upper tail for past and present climate. Ten-thousand fire seasons are simulated from each fitted model and we compare the proportion of simulated high-risk fire seasons to quantify the increase in risk. Our method suggests that the risk of experiencing an extreme wildfire season in Grand Lake, Colorado under current climate has increased dramatically compared to the risk under the climate of the mid-20^{th} century. Our method also finds some evidence of increased risk of extreme wildfire seasons in Quincy, California, but large uncertainties do not allow us to reject a null hypothesis of no change.

In the second project we explore a fundamental characterization of tail dependence and develop a method to classify data into the two regimes. Classifying a data set as asymptotically dependent

(AD) or asymptotically independent (AI) is a necessary early choice in the modeling of multivariate extremes. These two dependence regimes are defined asymptotically which complicates inference as practitioners have finite samples. We perform a series of experiments to determine whether a finite sample has enough information for a convolutional neural network to reliably distinguish between these regimes in the bivariate case. Along the way we develop a new classification tool for practitioners which we call `nnadic` as it is a **N**eural **N**etwork for **A**symptotic **D**ependence/ **I**ndependence **C**lassification. This tool accurately classifies 95% of test datasets and is robust to a wide range of sample sizes. The datasets which we are unable to correctly classify tend to either be nearly exactly independent or exhibit near perfect dependence, which are boundary cases for both the AD and AI models used for training.

In the third project we consider the challenge of using likelihood methods for models developed for the tail of the distribution. Many multivariate extremes models have intractable likelihoods thus practitioners must use alternative fitting methods and likelihood-based methods for uncertainty quantification and model selection are unavailable. We develop a proxy-likelihood estimator for multivariate extremes models. Our method is based on the tail pairwise dependence (TPD) which is a summary measure of the dependence in the tail of any multivariate extremes model. The TPD parameter has a one-to-one relationship with the dependence parameter of the HR distribution. We use the HR distribution as a proxy for the likelihood in a composite likelihood approach. The method is demonstrated using the transformed linear extremes time series (TLETS) models of Mhatre & Cooley (2024).

ACKNOWLEDGEMENTS

I thank my advisors Dan Cooley and Ben Shaby for their guidance and support. Both advisors trusted me throughout the research process providing resources, insightful questions (and necessary corrections), and freedom to explore. The work in this dissertation is a direct result of my work with Dan. Ben provided me with the opportunity to work as a GRA developing hierarchical Bayesian methodology which is not included here. Many thanks to both for allowing me to pursue both branches of research. Any expertise I have gained is a small part of what you shared.

I thank my Dissertation Committee, Dongzhou Huang, Tianying Wang, and Elizabeth Barnes for their time, comments, questions, and encouragement.

I thank the many teachers I have had along the way. The foundation which you helped me develop has led me here! A special thanks to Wen Zhou for frank encouragement in my first year in the PhD program.

I thank my fellow PhD students at CSU for the conversations we have had around homework, research, and other areas of life. Thank you for improving my understanding of statistics and the world around us.

I thank my parents for their perpetual support. Thank you for the plethora of opportunities you have opened up for me, for encouraging me along my many turns, for your advice, and for reminding me to have fun.

I thank my parents-in-law for their encouragement, involvement, and support throughout this journey. I am lucky to have such a phenomenal expanded family!

Finally, I cannot thank enough my incredible wife Maddi and son Cyrus. Thank you for giving me the opportunity to continue seeking and for exploring this wonderful life with me! Thank you for giving me something to always smile about, for believing in me, and for listening to me ramble about things I barely understand.

Funding

Thanks to Dan and Ben for providing funding through the following sources. The work in this dissertation was partially supported by NSF grants on Extremes Models and Methods from Transformed Linear Operations (DMS-1811657); A new parametric model, likelihood methods, and other advancements for multivariate extremes (DMS-2311164); and Collaborative Research: Combining Heterogeneous Data Sources to Identify Genetic Modifiers of Diseases (DMS-2309825). Thanks to taxpayers for supporting the advancement of knowledge.

DEDICATION

This dissertation is dedicated to

my wonderful family.

It is for, and because of, you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Many natural disasters can be thought of as events coming from the tail of some distribution. These events are rare but have an outsized impact on livelihoods, infrastructure, and ecosystems. Studying these multivariate events often involves statistical modeling which captures the dependence in the tail of the distribution and thus most non-extreme models are ill-suited. Central to many standard mulitvariate, time series, and spatial methods is the notion of covariance; PCA, LDA, linear time series analysis, factor analysis, and kriging are all based on covariance. However, covariance is not designed to capture dependence in the region of concern: the tail. The statistical field of extreme value theory has garnered much attention over the past few decades as it focuses on these tail events (see, e.g., Davison & Huser, 2015 for a review of some of the key ideas and Coles (2001) for an accessible introduction to the field). A fundamental aspect of any extreme value analysis is that it uses only a small fraction of the data, retaining only data which inform about extreme behavior.

Much of the development in extreme value theory is driven by the desire to estimate the probability of an event which is more extreme than any observed in the data. The well-developed study of univariate extremes began with the asymptotic study of block maxima which converge to the max-stable distributions. This field was expanded to consider threshold exceedances which converge to generalized Pareto distributions. Justifiable extrapolation for multivariate events rests on models which capture the dependence in the tail. Early work in multivariate extremes extended the univariate framework of block maxima into the study of componentwise maxima, which converge in distribution to the class of multivariate max-stable (equivalently, multivariate extreme value) distributions. These distributions have univariate max-stable margins and thus the advancement is in the understanding of the dependence structure. This dependence structure does not have a finite parametrization and the sparsity of information exacerbates the modeling challenges. The understanding of extremal dependence and methodology for modeling that dependence has been

an active research field over the past few decades and we contribute to that literature with three projects that are detailed in this document.

The dependence in multivariate max-stable distributions has been characterized by the so-called exponent function V (seen in Chapter 4), Pickands dependence function (Pickands, 1981; Marcon et al., 2017), and angular measure H (seen in Chapters 2 and 4). While the dependence space cannot be covered by a finite dimensional parametric family, several parametric models have been developed. These models have angular measures that smoothly transition between perfect dependence and independence by adjusting the dependence parameters. In other words, scalar summary measures of the dependence from these models can take on any possible value and thus, in some sense, the models can capture any "level" of dependence despite merely covering a finite dimensional subspace of the full space of models (Figure 3.8).

Geffroy (1958) and Sibuya (1960) demonstrated that properly re-scaled block maxima from a bivariate Gaussian distribution with any $\rho < 1$ converge to a separable (i.e., independent) bivariate extreme value distribution. This observation, and the further observation that max-stable models are either asymptotically dependent or exactly independent, led to the recognition of two different tail dependence regimes which are termed asymptotic dependence (AD) and asymptotic independence (AI). This distinction, which has been the subject of much research and discussion in the field, is the subject of Chapter 3.

The mathematical framework of regularly varying functions (see, e.g., Resnick, 2008a) is often used to study tail events. This framework is naturally tied to max-stable models because multivariate regular varying distributions comprise the domain of attraction for the multivariate max-stable distributions with heavy-tailed margins. In addition, the exponent measure and angular measure are concepts that arise through the framework of regular variation. Regular variation is a broader class of models than max-stable models as it is also found in heavy-tailed multivariate representations of threshold exceedances and thus it is commonly used for modeling large values. However, these regularly varying models are also AD and thus are viewed as insufficient in some cases which led to the development of hidden regular variation (Resnick, 2008b) and of so-called sub-asymptotic

or AI models (Ledford & Tawn, 1996, 1997). Determining when these AD models are insufficient is a challenging problem that is the focus of Chapter 3.

When the underlying margins are heavy-tailed, the dependence structure of multivariate regular variation naturally describes the dependence in the tail. We can use this dependence framework to describe the dependence in the tails even when the margins are not heavy-tailed. In that case, a marginal transformation to heavy tails is performed before modeling. This is similar to a copula approach but instead of transforming to uniform margins we transform to heavy-tailed margins. We discuss regular variation in Section 2.2 and rely on it in Chapters 2 and 4.

Much of the recent work on extremes, including the work in this volume, is motivated by environmental applications and thus models that describe random processes have been a focal point for the community over the past few decades. That work is largely spatial. See Davison et al. (2012); Cooley et al. (2012); Huser & Wadsworth (2022) for summaries of some of the most important advances in spatial extremes and extensive additional references.

The work in Chapters 2 and 4 focuses on times series models which capture dependence in the upper tail. The literature in this sub-field is different from the spatial literature. Two recent volumes on time series for extremes, (Kulik & Soulier, 2020; Mikosch & Wintenberger, 2024), demonstrate that this literature is often heavier on theory. This literature seems to be more common in financial applications than environmental applications.

The projects in this document are presented in chronological order of completion. However, when considering a new analysis one could start with applying the method detailed in Chapter 3 to classify data as either AD or AI. We see this as a necessary early step in a principled analysis which focuses on the tail of the distribution as it narrows the library of models available to the practitioner and thus is part of checking model assumptions. Chapters 2 and 4 are strongly related. In chapter 2 we detail an application of extremal time series models to real data which includes transforming the margins, handling seasonality, and performing inference. The model selection in chapter 2 is subjective which motivates development of more principled techniques in Chapter 4. That final project develops a likelihood-like model-selection and fitting technique which we

develop with respect to the same time series models used in Chapter 2, even though the method can be used with any regularly varying model.

In addition to the projects in this dissertation, I have spent considerable effort under the advisement of Dr. Ben Shaby on developing and applying methods to probabilistically identify genes associated with Parkinson's disease during my time at CSU. A primary challenge in this work is that each experiment, taken individually, may contain too little information to distinguish some important genes from incidental ones. Our method is built on a hierarchical three-groups mixture of distributions which describes the gene labels (beneficial, deleterious, or null). A Dirichlet distribution apportions prior probability of gene label assignment which leads to natural multiplicity correction. Modeling each data type as conditionally independent given the gene labels allows for modular inclusion of heterogeneous data types in a single coherent probability model. Our method results in parsimonious inference with enhanced power to detect signals. Simulation studies show that our method performs at least as well as commonly used tools for GWAS and RNA-seq, and in some cases it performs better. We applied our method to publicly-available GWAS and RNA-seq datasets and discovered novel genes as potential therapeutic targets. A manuscript is available on arXiv (Wixson et al., 2024) and has been submitted for review.

1.1 Outline

This dissertation details three advancements that target challenges in the modeling of extremal dependence with an eye towards principled application of previously developed models. We briefly introduce the projects and the major challenges they address in this outline.

1.1.1 Modeling season-long extreme events for risk assessment

The Colorado Division of Fire Prevention and Control reported that "20 of 20 largest wildfires have occurred in the last 20 years (since 2001)" and the three largest fires, which burned more than 500,000 acres, occurred in 2020 (State of Colorado, 2023). We seek to quantify how much more likely a season as extreme as 2020 is under current climate than past climate. Wildfires are not a

point-in-time event and the risk of extreme wildfires rises and falls with differing weather conditions. The most extreme risk is built over time as high temperatures and low moisture dry out fuels and high winds allow for rapid spread. The time-dependent nature of wildfire risk requires a more nuanced study than previous extreme attribution studies which typically estimate and compare very high quantiles of marginal distributions under different climate regimes.

In Chapter 2, we treat weather-related wildfire risk as a seasonal quantity and develop a method to model the upper tail of an entire wildfire season. This requires using time series models built to capture the dependence in the upper tail. Before we can fit these models, we need the data to be plausibly tail-stationary, which is achieved through handling the seasonality in the wildfire risk. To apply our method we develop a model selection method for these time series models. We fit models of several different orders with a method of moments type estimator and model selection is performed subjectively through diagnostic plots and assessing whether the fitted models can reproduce summary statistics. This work has been published in the Journal of Applied Meteorology and Climatology (Wixson & Cooley, 2023).

1.1.2 Classifying finite samples into asymptotically defined regimes

The study of extremes relies on intermediate asymptotics. In the threshold exceedances setting, this means that the sample size $n \to \infty$ and the number of retained order statistics $k \to \infty$ but the ratio $k/n \to 0$. These mathematically defensible results are hard to apply, as practitioners have finite sample sizes and convergence is often slow. This challenge is inherent to the difficulty in determining whether a dataset is AD or AI. These regimes are defined asymptotically and the best classification methods suffer from the slow convergence and the paucity of information in the tail. Many developed models describe the dependence of one regime and thus classification is an important exploratory step in the analysis of multivariate extremes.

In Chapter 3 we perform a series of experiments to determine whether there is enough information in finite sets of data to reliably classify them into their respective tail dependence regime. We use a supervised learning approach to develop a classifier that is trained on finite sets of data

that are labeled AD or AI and test whether the classifier can reliably distinguish between these asymptotically-defined regimes. Our classifier is a convolutional neural network (CNN) that is *a priori* agnostic about which regime is more likely and was developed as a finite sample classifier by which we mean that error rates are not expected to decrease with sample size. Our CNN is reasonably robust to a variety of sample sizes and correctly classifies at least 95% of testing datasets. This work has been submitted for review.

1.1.3 Using a proxy for the likelihood to fit models without densities

Likelihood methods are prized in statistics for their many nice properties including efficient use of data, natural uncertainty quantification, and model selection methods. Many multivariate extremes models are hard to fit with likelihood methods. Sometimes this is due to the dimension of the problem, which is an issue because the number of terms in the likelihood of any max-stable model grows combinatorially with the dimension. Other times it is because the density does not exist or does not have a closed form. The time series models used in Chapter 2 have both of these challenges.

In Chapter 4 we develop a proxy-likelihood method for multivariate extremes models that are built on the framework of regular variation. The method uses a regularly varying model with a likelihood as a proxy for the likelihood in models with intractable densities. The link between the two different models is a summary measure of the dependence known as the tail pairwise dependence (TPD, Cooley & Thibaud, 2019). We fully develop the method for use with the time series models from Chapter 2 but comment along the way how it can be adapted for use with other models. This method is the first to link the Hüsler-Reiss distribution (and related Brown-Resnick process) to linear methods for extremes.

1.1.4 Conclusion

We close the dissertation with a summary of the contributions and a brief discussion of areas for future work.

Chapter 2

Attribution of Seasonal Wildfire Risk to Changes in Climate: A Statistical Extremes Approach

2.1 Introduction

The Sixth Assessment Report of the IPCC suggests, with high confidence, that climate change has led to warmer and drier conditions which have increased wildfire risk in North America (Hicke et al., 2022). These worsening conditions have led to “increased burned area in recent decades in western North America" and thus in “the USA, annual costs of federal wildland fire suppression have increased by a factor of 4 since 1985" (Hicke et al., 2022, pg 1948). These national and regional trends have been echoed at state levels. The Colorado Division of Fire Prevention and Control reported that “20 of 20 largest wildfires have occurred in the last 20 years (since 2001)" and the three largest fires, which burned more than 500,000 acres, occurred in 2020 (State of Colorado, 2023). Data from the California Department of Forestry and Fire Protection (State of California, 2023) show that 18 of the 20 largest fires, 19 of the 20 most destructive fires, and 11 of the 20 most deadly fires in CA history occurred between 2003 and 2021. Abatzoglou & Williams (2016) state that the “increased forest fire activity across the western continental United States (US) in recent decades has likely been enabled by a number of factors, including the legacy of fire suppression and human settlement, natural climate variability, and human-caused climate change." The focus of this study is the climate signal; we do not consider questions of forest management, fuel availability, or the impact of more people in the wildland-urban interface. This chapter is an extreme event risk attribution study which aims to quantify how much more likely these extreme fire seasons are now than they were previously, due to observed changes in climate.

To narrow our focus to the climate signal we use the well-recognized Canadian Forest fire Weather Index (FWI) as the object of this study (Van Wagner, 1987). The FWI system is a series

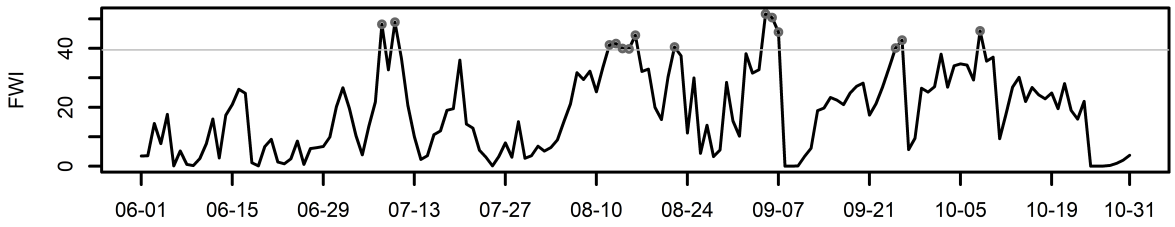


Figure 2.1: Daily FWI values from 2020, ERA5 data near Grand Lake, CO.

of equations which takes month, latitude, 24-hour precipitation, noon wind-speed, noon relative humidity, and noon temperature as inputs. These location and weather variables are used daily to compute the fine fuel moisture code (FFMC), duff moisture code (DMC), and drought code (DC) which represent dryness, and thus fuel availability, at differing levels of depth and time scales. The FFMC is combined with wind-speed to compute the initial spread index (ISI). The build up index (BUI), which represents fuel availability, is computed from the two longer range moisture codes (DMC and DC). Daily FWI values are computed from the ISI and BUI. The FWI was designed to represent the energy which would be released at the edge of a fire but can be interpreted in several ways including fire intensity risk. FWI is based on weather variables and thus changes discovered in the analysis are due to changes in observed climate.

The 153 day time series of FWI values for the 2020 fire season near Grand Lake Colorado is shown in Figure 2.1. Highlighted are the 14 days with FWI values above the 0.975 empirical quantile of present climate. Each highlighted, high-risk day was the result of processes like the drying of fuels through extended periods of low precipitation coupled with hot days and high winds. The time-dependent nature of wildfire risk makes it clear that simple estimation of some high marginal quantile would not capture the phenomenon we are concerned with. Wildfire risk, though quantifiable on smaller time scales (e.g. daily), is more sensibly thought of as a seasonal quantity and thus our aim is to model entire wildfire seasons.

We perform statistical attribution; a statistical model for entire wildfire seasons is built and fit to each climate period (past and present). The fitted models are used to simulate wildfire seasons. The proportions of simulated high-risk seasons from each period are compared to make our attribution statement. We consider a season to be high-risk if it has at least as many days above a high quantile

as were observed in the most extreme year in the region. These years, 2020 in Colorado and 2021 in California, are used to define high-risk seasons so that communication of results is simple and relatable.

This study uses a reanalysis product and weather station data. Use of these data sources allows us to directly make within-product comparisons as we can use the 2020 (2021) season from each source to define high-risk. However, these data sources could have confounding factors. Reanalysis products are a data-assimilation of model output and global observational data from ground sensors and satellites but satellite data are only available in the present period. Weather station data collection methods transitioned from manual to automated measurements during the study period and there may have been changes in the location of the measurements. Additionally, weather station data has the usual challenges inherent in true observational data; missing observations, short historical records for some variables, and the noisiness expected with truly local measurements.

An alternative and often used attribution approach employs climate models which were run under factual and counterfactual worlds (i.e., without anthropogenic forcings). Our year-specific definition of high-risk makes it difficult to use climate model output in the estimation of the proportion of high-risk seasons. Comparing climate model output to observed conditions in 2020 (2021) would require accounting for model bias but model bias is not well understood for FWI. A recent attribution approach studies specific events using custom climate model runs which were set up to adequately reproduce the event (e.g., Patricola & Wehner (2018) study fifteen tropical cyclones with an emphasis on hurricane Katrina). These studies require extensive computing resources and, to our knowledge, none have studied season-long events. An advantage of our method is its simplicity; we use available data. A disadvantage is that our attribution statement can only reference differences between past and present climate and cannot speak causally about an anthropogenic effect.

This study compares the proportion of high-risk seasons expected under past climate to the proportion expected under present climate. We define past climate as the 20 earliest available years (1959 to 1978) and present climate as the years from 2002 to 2021. Our definitions of past

and present climate do not consider the potential effects of cyclical climate patterns (ENSO, etc.). Additionally, each climate period only has 20 observed seasons and thus uncertainty (computed through bootstrapping) is large. Within each climate period we consider a fire season to be June 1 through October 31 (153 days). We first consider an area north of Grand Lake, CO which was burned in the East Troublesome fire in 2020 and then repeat the analysis on a region just outside the burn area of the 2020 North Complex and 2018 Camp fires in California.

We use time series models to capture dependence throughout the fire season. Classical linear time series models (like $ARMA$ models) are based on the autocovariance function (ACVF) which averages the linear relationship around the mean and thus may not accurately capture the dependence at extreme levels. We characterize tail dependence with the tail pairwise dependence function (TPDF) and employ the transformed-linear extremes time series models (TLETS) of Mhatre & Cooley (2024). These models resemble the familiar $ARMA$ models from classical time series but are tied to regular variation (a common framework in extremes) and thus are a natural choice for this study.

This attribution study fits into the EEA framework of Jézéquel et.al. (2018) as a "risk-based" study because we focus on computing an increase in risk rather than explaining the link between climate change and the physical processes which led to the event. We compute how much more likely a high-risk season is now than in a past climate and thus are interested in a "class of events". This differs from studies which are focused on computing the probability that a single event (i.e., the 2020 fire season in Colorado) was caused by climate change. Finally, while some studies condition on sea-surface temperature, greenhouse gas concentrations, etc., our study uses the observational record to compute the changes in risk based on any detectable changes in climate and thus is considered "unconditional".

This chapter is organized as follows. We review regular variation which is the framework for the models that we use in Section 2.2 and introduce the models in Section 2.3. Our method is explained in detail as it is applied to ERA5 data from one location inside the East Troublesome burn area in Colorado in Section 2.4. We discuss the data and pre-processsing in Section 3.1, estimation

of the pairwise dependence and model fitting is Section 3.2, and simulation from the models and attribution is Section 3.3. In Section 2.5 we apply our method to weather station observations (RAWS data) from the same location as in Section 3. In Section 2.6 we apply our method to ERA5 data from a location near the 2020 North Complex Fire in California. We conclude the paper with discussion of the benefits and some limitations of our method.

2.2 Regular Variation and Dependence

To model dependence in the upper tail of our time series, we rely on the framework of regular variation. One way to understand the tail of the distribution function F of a random variable X is to consider the rate of decay of the survival function $\bar{F}(x) = 1 - F(x)$ as $x \to \infty$. The framework of regular variation naturally considers this decay.

Regular variation is a property of some functions which intuitively says that the function decays like a power function in the neighborhood of some value. We say that f is a regularly varying function *at infinity* if there exists an $\alpha \in \mathbb{R}$ such that $\lim_{t\to\infty} \frac{f(tx)}{f(t)} = x^{-\alpha}$. We denote this $f \in RV_\alpha$. Random variable X is regularly varying if $\bar{F} \in RV_\alpha, \alpha > 0$. Resnick (2007, Theorem 3.6) showed that $\bar{F} \in RV_\alpha$ if and only if there exists a sequence $\{b_n\}$ with $b_n \to \infty$ such that $n\mathbb{P}(\frac{X}{b_n} \in \cdot) \xrightarrow{v} \nu_\alpha(\cdot)$ where $\xrightarrow{v}$ denotes vague convergence in the space of nonnegative radon measures $M_+(0, \infty]$ and $\nu_\alpha(x, \infty] = x^{-\alpha}$.

This second definition of univariate regular variation generalizes well to multiple dimensions. Resnick (2007, Section 6) demonstrates that random vector $\mathbf{X} \in \mathbb{R}^d$ is multivariate regularly varying if there exists a sequence $\{b_n\}$ with $b_n \to \infty$ and a Radon measure $\boldsymbol{\nu}$ on the space $\mathbb{E} = [0, \infty]^d \setminus \{\mathbf{0}\}$ such that in $M_+(\mathbb{E})$

$$n\mathbb{P}\left(\frac{\mathbf{X}}{b_n} \in \cdot\right) \xrightarrow{v} \boldsymbol{\nu}(\cdot). \tag{2.1}$$

The limiting measure $\boldsymbol{\nu}$ has scaling property $\boldsymbol{\nu}(cB) = c^{-\alpha}\boldsymbol{\nu}(B)$ for any $c > 0$ and set $B \subset \mathbb{R}^d$. This scaling property highlights the independent decomposition of $\boldsymbol{\nu}$ which shows that these

random vectors have an independent pseudo-polar decomposition at infinity. Let $||\cdot||$ be a norm and define the unit ball $\mathbb{S}_{d-1} = \{\mathbf{x} \in \mathbb{R}^d : ||\mathbf{x}|| = 1\}$. Let $C(r, B) = \left\{\mathbf{x} \in \mathbb{R}^d : ||\mathbf{x}|| > r, \frac{\mathbf{x}}{||\mathbf{x}||}\right\}$ for some radial value $r > 0$ and Borel set $B \subset \mathbb{S}_{d-1}$. Then $\boldsymbol{\nu}[C(r, B)] = r^{-\alpha} H_{\mathbf{X}}(B)$ where $H_{\mathbf{X}}$ is an angular measure on $\mathbb{S}_{d-1}$. This angular measure contains all of the dependence information. Finally, it should be noted that b_n, $\boldsymbol{\nu}$, and $H_{\mathbf{X}}$ are not uniquely determined by (2.1) as b_n can be scaled by any positive constant and this will be absorbed into the limiting measure.

The dependence structure of regular variation is found in characterizations of multivariate extreme value distributions which makes it a natural and common mathematical framework for the study of extremes. The definition only describes the joint tail behavior and thus it is only useful to characterize the joint tail of our data. Additionally, regular variation is a useful framework for capturing asymptotic dependence (see Chapter 3 or Coles, 2001, Section 8.4 for further discussion), which loosely implies that variables can be at their most extreme levels at the same time. Our FWI time series appears to exhibit asymptotic dependence at short lags.

Intuitively, a d-dimensional random vector must have heavy tails in each dimension to be a multivariate regularly varying random vector. The decay rate of those tails is described by the tail index α. The distribution of a regularly varying random vector at infinity can be decomposed into independent radial (distance from the origin) and an angular (the point where the vector intersects the unit ball) components (Figure 2.2). We use this near-independent (in the large but not limiting case) decomposition to describe the pairwise tail dependence between the dimensions of the random vector. Consider a two dimensional random vector $\mathbf{X} = (X_1, X_2)$ which has positive components with probability 1 (i.e., the random vector takes values in the first quadrant of the xy-plane). Further assume that $\mathbf{X}$ is large in at least one component. Information about the tail dependence between components of the vector is contained in the angular measure $H_{\mathbf{X}}$ which describes the distribution of angles (between 0° and 90°). If there is strong tail dependence we would expect that when X_1 is extreme X_2 will also be extreme. In this case $H_{\mathbf{X}}$ will have mass near 45° and we would expect to see points like A in Figure 2.2. When the tail dependence is weak (or zero) knowing X_1 is extreme tells us little about X_2 . This suggests X_2 will likely be in the bulk

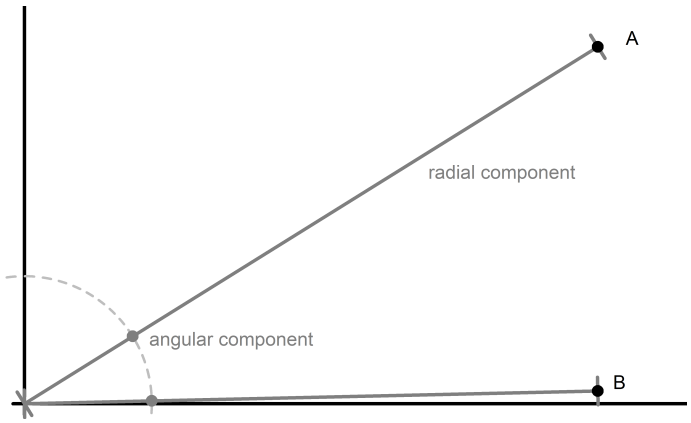


Figure 2.2: Polar decomposition of points A and B. The radial component is the norm of the point and the angular component places the point on the unit ball. We expect points like A (large together) under strong tail dependence and points like B (large separately) under weak tail dependence.

of its distribution (which is near the axis due to the distribution's heavy tail) and we would expect extreme points to be like point B in Figure 2.2. In this case the mass of the angular distribution $H_{\mathbf{X}}$ is concentrated near the axes.

Tail dependence of a d-dimensional $\mathbf{X}$ continues to be described by its angular measure $H_{\mathbf{X}}$, but estimating or modeling this measure which lies on the d-dimensional unit ball becomes increasingly difficult as d grows. Multivariate models tend to be employed in, at most, moderate ($d \approx 5$) dimensions.

When analyzing time series we treat the vector of observations as a single (partial) realization of an infinite dimensional random vector. In our analysis we consider each season as an observation from the climate of that period, and thus we have 20 (partial) realizations of length 153 (the number of days in our definition of a fire season). The advantage of taking a time series approach over simply viewing the data as realizations of a 153-dimensional random vector is that we can characterize and model dependence as a function of lag (time difference between variables). Classical (non-extreme) time series analysis generally assumes a time series is weakly stationary and restricts focus to characterizing only pairwise dependencies through the autocovariance function.

Here, we summarize the angular measure using a pairwise metric analogous to the autocovariance function, but which characterizes tail behavior. Our pairwise metric is the TPDF (2.2) which is the time series analogue to the tail pairwise dependence matrix (TPDM) introduced

by Cooley & Thibaud (2019) and employed by Jiang et al. (2020) to perform extremal principal component analysis for US precipitation. Let the time series $\{X_t\}$ be regularly varying with tail index $\alpha = 2$ for all $t = 1, 2, ...$ and tail stationary (Mhatre & Cooley, 2024); that is, the TPDF is a function of lag only. Define the two dimensional unit ball in the positive orthant $\mathbb{S}_1^+ = \{\mathbf{X} \in \mathbb{R}^2 : x_1, x_2 \geq 0 \text{ and } ||\mathbf{X}||_2 = 1\}$ (where $||.||_2$ is the euclidean norm) and for each lag h, let the radial component $r_t = ||(x_t, x_{t+h})||_2$. The TPDF is

$$\sigma(h) = \sigma(X_t, X_{t+h}) = \int_{\mathbb{S}_1^+} s_1 s_2 dH_{X_t, X_{t+h}}(s). \tag{2.2}$$

where $\mathbf{s}$ is the angular component of (X_t, X_{t+h}): $s_1 = \frac{x_t}{r_t}$ and $s_2 = \frac{x_{t+h}}{r_t}$.

Although regular variation assumes the data are heavy tailed, it can be used as a dependence model for data which are not heavy tailed. We will transform data such as the FWI data so that they can be assumed to come from a tail-stationary regularly varying time series with tail index $\alpha = 2$. This idea is not uncommon in classical extreme value analysis where characterizations of the extreme value distributions are typically made assuming a particular marginal distribution (de Haan & Ferreira, 2006, Section 6.1.2), and is similar in spirit to transforming time series data (such as applying a square root or logarithmic transformation) to make it appear more Gaussian in order to fit traditional time series models. Our assumption that $\alpha = 2$ is made for convenience, Kiriliouk & Zhou (2022) recently extended the definition of the TPDM for a general tail index, but its definition includes α in the integrand.

2.3 Transformed Linear Extremes Time Series Models

We want models specifically designed to fit the upper tail as this is when fire risk becomes a concern. The TLETS models of Mhatre & Cooley (2024) are analogous to classical ARMA models (see, e.g., Brockwell & Davis, 2002 for an introduction) but are built to capture the dependence in the upper tail. Classical ARMA models are built out of linear combinations of white noise and it is standard, but not necessary, to consider Gaussian noise. The TLETS models use transformed-linear combinations of regularly varying noise noise. Additionally, classical ARMA models are

investigated based on the second-order property of covariance. TLETS models also consider a second-order property but in this case that property is the TPDF.

2.3.1 Transformed-linear Operations

Standard estimators of tail dependence average across the unit ball which results in biased estimates if the dependence is not symmetric. Often the extremes that we care about occur in one tail and thus TLETS models were specifically designed to fit the upper tail (other tails can be easily considered through rotations). TLETS models only fit the upper tail because the random processes only exist in the positive orthant as they are constructed with the transformed-linear operations of Cooley & Thibaud (2019). Transformed-linear operations are defined component-wise and involve a map, f, from the real line to the positive half line. For any two vectors in the positive orthant, $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}_+^d$, transformed-linear addition, denoted $\oplus$, is performed by mapping the components of $\mathbf{X}_1$ and $\mathbf{X}_2$ back to the real line, adding the two vectors, and then transforming the components back to the half-line: $\mathbf{X}_1 \oplus \mathbf{X}_2 = f\{f^{-1}(\mathbf{X}_1) + f^{-1}(\mathbf{X}_2)\}$. Transformed-linear multiplication, denoted $\circ$, works similarly: $a \circ \mathbf{X} = f\{af^{-1}(\mathbf{X})\}$. We use the function $f(x) = \log\{1 + \exp(x)\}$ as it has a negligible effect on the upper tail (i.e., $\lim_{x\to\infty} \frac{f(x)}{x} = 1$) and thus regular variation is preserved under these operations (Cooley & Thibaud, 2019). Transformed-linear time series are constructed using transformed-linear operations in the place of classic arithmetic operations on a noise sequence Z_t of independent, tail stationary, regularly varying $\alpha = 2$ random variables.

2.3.2 TLETS MA(q) Models

Transformed linear time series are constructed using transformed linear operations in the place of classic arithmetic operations on a noise sequence Z_t of independent, tail stationary, regularly varying $\alpha = 2$ random variables. We say that $\{X_t\}$ is a transformed-linear moving average process of order q (denoted $TL - MA(q)$) if, for all t,

$$X_t = \bigoplus_{j=0}^{q} \theta_j \circ Z_{t-j} \tag{2.3}$$

for $\theta_j \in \mathbb{R}$, $\theta_0 = 1$, and $\theta_q > 0$. The TPD at lag-h from a $TL - MA(q)$ is 0 whenever $h > q$. For $h \leq q$, When the marginal distribution of X_t has scale 1, the TPD is

$$\sigma(h, \theta_1, \ldots, \theta_q) = \frac{\sum_{l=0}^{q} \theta_l^{(0)} \theta_{l+h}^{(0)}}{\sum_{l=0}^{q} \theta_l^2} \tag{2.4}$$

where $a^{(0)} = \max(a, 0)$, and the last h terms in the sum in the numerator are 0.

For any transformed linear $ARMA(p, q)$ process:

$$X_t \oplus (-\psi_1) \circ X_{t-1} \oplus \cdots \oplus (-\psi_p) \circ X_{t-p} = Z_t \oplus \theta_1 \circ Z_{t-1} \oplus \cdots \oplus \theta_q \circ Z_{t-q}, \tag{2.5}$$

there exists an equivalent $TL - MA(\infty)$ process. Due to this equivalence, and the ease of fitting $TL-MA(q)$ models for arbitrarily large q, we restrict our attention to the $TL-MA(q)$ processes. Other models are considered in more detail in Chapter 4.

In classical time series the innovations algorithm can be used to recursively compute the one step predictors and then uses the associated prediction errors (innovations) to estimate that steps' MA coefficients. The transformed-linear extremes innovations algorithm (Mhatre, 2022) is similar and is used to fit the $TL - MA(q)$. Mhatre (2022) showed that even if the underlying model is not a transformed-linear model, use of the extremal innovations algorithm will result in a transformed-linear model with a TPDF that closely matches the TPDF of the underlying model and thus the pairwise dependence will match even if the model is not the correct one.

2.4 Application to ERA5 Data in Colorado

2.4.1 Data and Pre-processing

We first apply our method to a reanalysis data product (ERA5 from the Copernicus Climate Change Service) (European Centre for Medium-Range Weather Forecasts, 2023) on one grid-box in Grand Lake, Colorado. ERA5 data are produced on a globally complete, hourly, 30 km grid from 1959 to about two months prior to access. The grid-box we considered includes latitude 40.27

and longitude -105.84 which is the location of a Remote Automatic Weather Station (RAWS) and is in the burn area of the East Troublesome Fire. Proximity to a weather station will allow us to repeat analyses performed on ERA5 data using observed weather for comparison (Section 2.5). The weather variables downloaded were the hourly ten meter wind component in the eastern and northern directions (used to calculate wind speed), two meter dew point temperature (used to calculate relative humidity), two meter air temperature, and total precipitation. The time series of noontime measurements for wind speed, relative humidity, and temperature were combined with the time series of 24-hour precipitation to compute daily FWI values for the season.

An early task in any time series analysis is the assessment of seasonality. It is reasonable to expect that wildfire risk changes as each wildfire season progresses (Figure 2.3) and thus we must address the seasonality before assessing the dependence. We consider each day of the wildfire season to have its own distribution under two simplifying assumptions. First, we assume there is no meaningful trend within each defined climate period. Exploratory analysis does not provide evidence against this assumption. Second, we assume that the marginal (daily) distribution of FWI changes throughout the year in a smooth manner. This assumption, while untested, seems reasonable as we do not expect the distribution of fire risk on July 15$^{\text{th}}$ to be meaningfully different than the distribution on July 16$^{\text{th}}$. This smoothness assumption allows us to borrow strength from nearby days in the estimation of each day's marginal distribution without introducing much bias.

Seasonal behavior was explored with daily high quantile plots (0.975 in Figure 2.3) which were computed and smoothed using a $(2k+1)$-day moving window. Exploratory analysis led us to choose $k = 14$ resulting in a 29-day window (e.g., FWI values for July 1$^{\text{st}}$ through July 29$^{\text{th}}$ are used to estimate quantiles for July 15$^{\text{th}}$). As anticipated, seasonal behavior was evident and will need to be accounted for via marginal transformation. However, the bimodal sub-seasonality seen in high quantiles of the FWI in Figure 2.3 was unexpected, and was apparent across a wide range of window sizes (smaller windows exhibited the predictable increase in noise). Investigating this bimodality further, we found that the high quantiles of the DMC exhibit similar bimodality. This high quantile bimodality is not as evident in the FFMC and is not in the DC. The DMC captures

dryness of mid-depth fuels, retains moisture information for around two weeks, and precipitation less than 1.5mm is considered too little to reach these fuels. Estimated precipitation quantiles (between 0.05 and 0.9) are largest between mid-July and early-August which suggests that precipitation large enough to dampen the DMC is more common at this time than during the rest of the season.

The marginal shift in high quantiles shown in Figure 2.3 suggests an increase in fire risk; values at the past climate 0.975 quantile are now observed more frequently. We sub-sampled years for both periods and computed the 0.975 quantile for each day to test whether the quantiles were significantly different. Subsampling seasons allows the bootstrap to account for temporal dependence in the data. With each of these 500 bootstrap samples we computed the differences (for each day) of the quantiles and computed 95% bootstrap intervals for the difference. Only the intervals for June 1st through 3rd, June 17th, July 8th through 12th, and October 13th through 31st contained zero. To illustrate uncertainty in Figure 2.3, day-wise confidence bands plotted were computed using the method of Goldstein & Healy (1995), and overlap of these confidence bands closely mimics the dates found by our hypothesis test. We will see that the tail dependence is similar between climate periods (Figure 2.4) which suggests that most of the increase in risk that we will find is due to this marginal shift of extreme levels.

Quantile plots of the underlying sub-indices, moisture codes, and weather variables were explored to better understand the meteorological drivers contributing to this observed change in FWI. This shift towards more extreme values under current climate than under past climate is evident across BUI quantiles. ISI quantiles have also increased, but that increase fades in mid-October. These sub-index increases can be traced back through the FFMC, DMC, and DC to an increase in temperatures and a decrease in moisture. The median change between study periods in daily 0.975 quantile of temperature is $2^{\circ}C$. A similar increase in temperatures is evident across quantiles. Decreases in the distribution of precipitation and relative humidity are also apparent. For example, the 0.5 quantile of daily precipitation under past climate has a similar seasonal pattern and level to

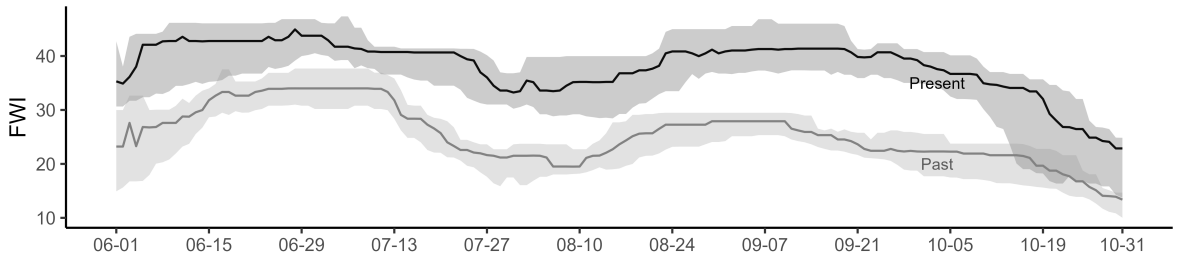


Figure 2.3: Daily 0.975 quantile of the ERA5 FWI time series from Grand Lake, CO for both past and present periods with day-wise confidence bounds computed through bootstrapping. The confidence bounds correspond to 95% intervals for the differences between periods. Quantiles were computed by borrowing strength from surrounding days using a 29-day moving window. The marginal distribution at high quantiles seems to have shifted up in a nearly uniform fashion.

the 0.6 quantile under present climate. Windspeed quantiles appear relatively unchanged between periods.

In order to fit the models of Mhatre and Cooley (2022), we transform so the data exhibits weak tail stationarity and the marginal distribution is regularly varying with tail index $\alpha = 2$. We first transform the marginal distribution to be uniform and, in doing so, handle the observed seasonal behavior.

We do this daily with the same 29-day moving window as we used to plot the daily FWI high quantile. Each day's marginal distribution, F_t^p, is estimated with the empirical cumulative distribution function (ECDF) below the 0.975 quantile, μ_t, and a fitted generalized Pareto distribution (GPD) above. The GPD is a common model for exceedances above a threshold, μ, and can be defined with the conditional survival function $P(Y > y|Y > \mu) = \{1 + \xi(y - \mu)/\psi\}^{-1/\xi}$ where $\xi \in \mathbb{R}$ is the shape parameter and $\psi > 0$ is the scale parameter (Coles, 2001). Our estimated daily semi-parametric marginal distribution is

$$\hat{F}_t^p(y) = \begin{cases} \{n_p(2k+1)+1\}^{-1} \sum_{i=1}^{n_p} \sum_{j=t-k}^{t+k} \mathbb{I}(y_{ij}^p < y) & y \leq \hat{\mu}_t^p \\ 1 - 0.025 \left\{1 + \hat{\xi}^p (y - \mu_t^p)/\hat{\psi}^p\right\}^{-1/\hat{\xi}^p} & y > \hat{\mu}_t^p, \end{cases} \tag{2.6}$$

where y_{ij}^p is the observed FWI from day j of year i in period p (past or present), n_p is the number of years in that period, and $2k + 1$ accounts for the 29-day moving window. The denominator of the

ECDF is increased by one so that the observed maximum does not have ECDF value one (Coles, 2001, Definition 2.4) and the scaling by 0.025 in the GPD arises from modeling exceedances over the 0.975 quantile. Likelihood ratio tests do not provide evidence against the null hypothesis that the scale and shape parameters of the daily GPD were the same for each day in the year. Mean residual life plots confirm the use of the daily 0.975 quantile as the threshold parameter of the GPD. The inverse Frechet CDF, $x_t = \log(1/u_t)^{-1/2}$, is applied to the time series with uniform marginals to make it regularly varying with tail index $\alpha = 2$.

2.4.2 Dependence Estimation and Model Fitting

To estimate the dependence we use the natural TPDF estimator of Cooley and Thibaud (2019). We reparameterize the lag-h pairs of points $(x_t, x_{t+h}),\ t = 1, ..., n-h$ with polar coordinates (Figure 2.2). The radial component is the L_2-norm: $r_t = ||(x_t, x_{t+h})||_2$. The angular component $\mathbf{s} = (s_t, s_{t+h}) = (x_t, x_{t+h})/r_t$ places the point on the unit ball. The TPDF estimator under this parameterization is

$$\hat{\sigma}(h) = \frac{2}{\sum_{t=1}^{n-h} \mathbb{I}(r_t > r_0)} \sum_{t=1}^{n-h} s_t s_{t+h} \mathbb{I}(r_t > r_0), \tag{2.7}$$

where the two arises from the known (after transformation) marginal distribution. This estimator replaces the angular measure by its empirical estimate and considers only points above some high threshold r_0 (we used the 0.975 quantile).

The estimated TPDF (Figure 2.4) is similar between the two climate periods. It is well known that extremal dependence measures are biased when tail dependence is weak (see e.g. Huser et al. (2016)) and this bias was noticed and explored in Mhatre (2022). We follow Mhatre's suggestion and subtract off the mean of the time series before estimating the dependence. This correction reduces, but does not eliminate, bias. Our tail-dependence estimates decrease with lag, but they appear to level off around 0.1 near lag 15 presumably because of remaining bias. Simulations of $TL-MA(q)$ time series ($q = 3, 5, 10$) had TPDF values, after bias adjustment, beyond lag q close to 0.1.

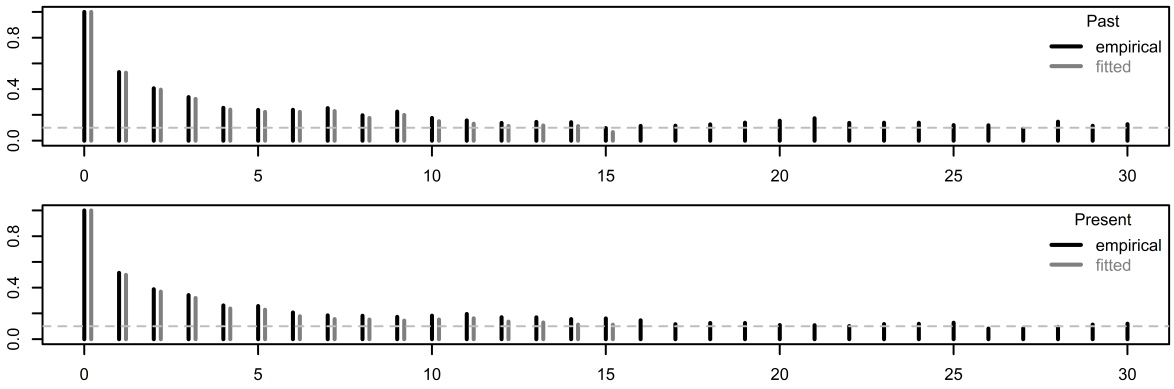


Figure 2.4: Empirical and fitted TPDF for past (top) and present (bottom) periods of transformed CO ERA5 FWI time series. Our fitted model captures the empirical dependence well. The dependence appears similar between periods. We note that the plots appear nearly constant around 0.1 after 10-15 lags.

The order q of our $TL - MA(q)$ is chosen by fitting and comparing several orders. The extremal innovations algorithm is used to estimate the parameters of $TL - MA(q)$ models for $q = 1, 2, ..., 30$. Comparing the estimated TPDF plots to the theoretical TPDF from each iteration of the innovations algorithm demonstrates that our fitted model has very similar pairwise tail dependence as our data and is our first step in model choice (Figure 2.4 compares the empirical TPDF to the fitted $TL - MA(15)$). The empirical TPDF appears to level off between lag 10 and 15 under both past and present climate which suggests that the unbiased TPDF would likely be zero somewhere between lag 10 and 20. We compare the ability of these models to reproduce summary statistics (mean run length above high quantiles and high quantile of sum of 5 consecutive terms) and note, once again, that models of order 10 through 20 are reasonable. Due to the leveling off of the empirical TPDF after lag 15 we continue the analysis with the $TL - MA(15)$ for both periods.

2.4.3 Simulation of Seasons and Attribution

Once we have fit our chosen models we simulate 10,000 seasons from each period. To simulate seasons we first generate a noise sequence of 168 (length of the season 153 plus model order 15) independent Frechet random variables with scale one and shape two. This independent noise sequence is iteratively put through (2.3) with $q = 15$, $\theta_0 = 1$, and θ_j ($j = 1, 2, ..., 15$) equal to the coefficients that were estimated using the extremal innovations algorithm for the given period.

This ensures the pairwise tail dependence of the simulated season matches what was observed. To reintroduce the observed seasonality into these 10,000 stationary simulated seasons we perform a two step back transformation. We first use the ECDF to transform to a uniform marginal distribution. We then use the inverse of (2.6) to transform the uniform marginal time series into seasons on the original FWI scale.

Attribution is done by comparing probabilities of observing high-risk seasons under past and present climates. We consider a season to be high-risk if it had at least as many days above a high threshold as were observed in 2020. Our thresholds were the 0.95, 0.975, and 0.99 quantiles of the observed present climate FWI time series (34.41, 39.47, and 44.94 respectively) (Table 2.1). In 2020 there were 22, 14, and 6 days over the respective thresholds. Under past climate 31 of the 10,000 seasons had at least 22 days with FWI values above 34.41 compared to 322 simulated present climate seasons. The ratio of the point estimates (10.39) suggests that a season as extreme as 2020 is more than ten times as likely under present climate as it was under past climate. The ratios for the 0.975 and 0.99 quantile thresholds were 4.33 and 8.77 respectively.

Uncertainty in our estimates is reported with intervals computed through bootstrapping. Each bootstrapped estimate was computed by sampling 20 years, with replacement, from each climate period. We use the observed data from these 20 years to estimate the marginal distribution (2.6), transform to be regularly varying, estimate the TPDF, fit the $TL - MA(15)$, simulate 10,000 seasons, and compute proportion of high-risk seasons.

The uncertainty bounds computed from these bootstrapped results are large, due to only having 20 observed seasons with which to estimate all aspects of our method. Despite the width of the confidence intervals, the ratios are entirely above one, and the change in attributed seasonal risk is significant. The bootstrapping is computationally intensive, but simple to implement. Five hundred full analyses (minus model order selection) were completed with each using less than 0.5 GB of RAM running on a single thread and thus the analysis could be completed on a laptop computer. Access to a computer with 64 2.7GHz cores and 128 GB of RAM, combined with the

Table 2.1: Results from Grand Lake, CO ERA5 data using a transformed-linear $TL - MA(15)$ for past and present climate. Columns 1 and 2 report three high thresholds of the present climate and the number of days in 2020 which exceeded those high quantiles is in column 3 (definition of high-risk season). The proportion of simulated high-risk seasons are in columns 4 and 5 for the respective periods. Column 5 is the multiplicative change from past to present (ratio of columns 3 and 4). Bootstrapped 95% CI's are reported in parenthesis.

Quantile	Threshhold	# in 2020	Past	Present	Ratio
0.95	34.41	22	0.003	0.032	10.39
			(0.000, 0.007)	(0.001, 0.097)	(5.96, ∞)
0.975	39.47	14	0.012	0.050	4.33
			(0.001, 0.035)	(0.009, 0.174)	(2.46, 51.56)
0.99	44.94	6	0.023	0.201	8.77
			(0.000, 0.092)	(0.074, 0.388)	(2.33, ∞)

embarrassingly parallel nature of the bootstrapping, allowed for rapid results from all 500 analyses (each bootstrap analysis took around 12 minutes to complete on this machine).

Sensitivity to model order was explored by completing the analysis with $TL - MA$ orders 10 and 20. The ratios of proportion of extreme fire seasons (present to past) were 24, 20.68, and 9.88 respectively when using the fitted $TL - MA(10)$ to simulate seasons. There were fewer simulated high-risk seasons with the $TL - MA(10)$ than either of the other models but the ratio is increased as there were proportionally fewer high-risk seasons in the past as in the present. When using the fitted $TL - MA(20)$ to simulate seasons the ratios were 8.02, 4.94, and 9.79 respectively which are similar to those obtained with our chosen model. We expect the ratio estimates to be relatively stable between models of orders that accurately capture the tail dependence.

We summarize our findings with the following attribution statement. Applied to FWI data generated from ERA5 output, our method estimates that a wildfire season like the one observed in 2020 near Grand Lake Colorado is 4 to 10 times more likely under recently observed climate than under the climate of roughly 50 years ago. Our method rejects the null hypothesis that the risk of observing a season like 2020 is unchanged between these two periods. Figures 2.3 and 2.4 indicate that this increase in risk is due to a shift in the marginal distribution of risk, not the tail dependence.

2.5 Application to RAWS data in Colorado

2.5.1 Data and Pre-processing

We apply the same method to RAWS weather station data (National Interagency Fire Center, 2023) from Harbison Meadow (NWS ID 050402) which is located within the grid-box used for the ERA5 analysis in Section 3. The analysis is repeated on weather station data to allow for comparison with reanalysis products and is easily explainable to a broader audience. The RAWS data for Harbison Meadow were downloaded from National Wildfire Coordinating Group (NWCG) Cognos portal (https://famprod.nwcg.gov/cognos11).

The present period had a nearly complete record but the past period had only 12 years of usable data. Harbison Meadow has data recorded from 1964 to the present but the years from 1972 through 1974 were deemed unusable for two reasons: approximately half of the data in each year was missing and the maximum FWI value for those years was 0.47 which is well below the next lowest maximum value of the past period, 16.04, and is the 0.56 quantile of the past period.

Seasonal behavior of the RAWS FWI time series was explored in the same manner as was done with the ERA5 data. The FWI daily high-quantile (Figure 2.5) again shows a clear shift in the marginal distribution at the highest quantiles which is often larger in magnitude than it was in the ERA5 data. The maximum difference between present and past estimated 0.975 quantiles in the ERA5 data was 18 units. There were 62 days in RAWS data (40% of the season) which had differences that were greater than 18 units. The present period indicates bi-modality in the highest quantiles but the second peak is missing in the past period. The 0.975 quantile does not remain at its highest values for as much of the season in the RAWS data as it did in the ERA5 data but these highest values are greater than what was observed in the ERA5 data. We expect that some of this difference is due to the different spatial scales and model bias. Exploration of high quantiles of FWI components indicated extreme temperatures are hotter, extreme wind-speeds are faster, and humidity is lower in the current period than in the past period. Additionally, high quantiles of all three fuel moisture codes (FFMC, DMC, and DC) and of the fire behavior sub-indices (ISI and BUI) show an increase in risk.

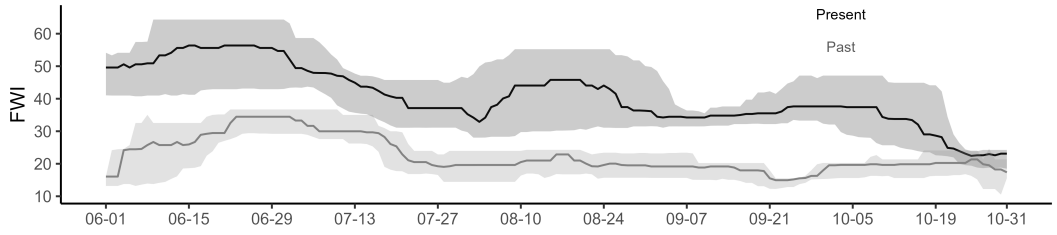


Figure 2.5: Daily 0.975 quantile of the FWI time series computed from the RAWS Harbison Meadow data as in Figure 2.3.

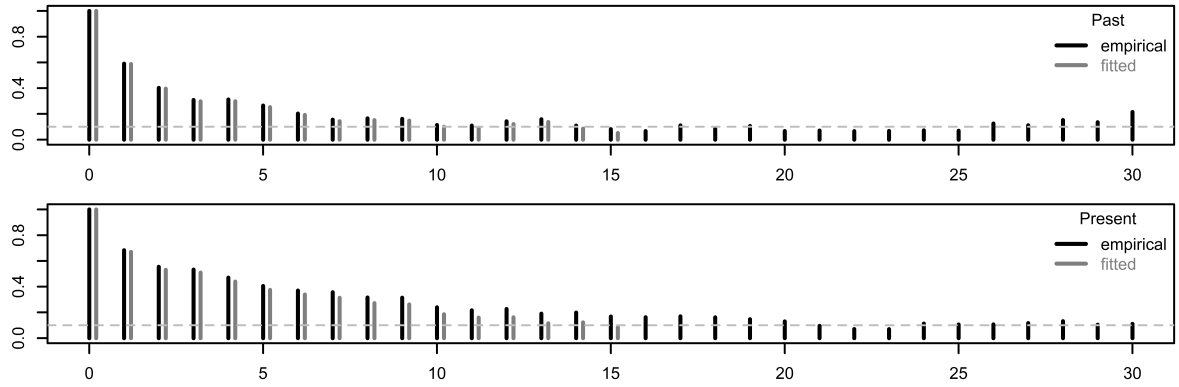


Figure 2.6: Empirical and fitted TPDF for RAWS Harbison Meadow FWI time series as in Figure 2.4.

The high quantiles for past and present climate are significantly different. This was determined using joint daily 95% bootstrap intervals which only overlap at the end of the season (Figure 2.5).

The TPDF for the past period appears to differ from the present period TPDF (Figure 2.6) which was not apparent in the ERA5 data (Figure 2.4). This suggests that there may be a change in the dependence as well as a change in the marginal distribution. We explore the effects of this by fitting models of orders 10, 15, and 20 to both periods. The results were similar in all cases including when we compared different orders for the two periods. The following results are from the $TL - MA(15)$ fitted to both periods because that matches the previous analysis.

2.5.2 Attribution

Table 2.2 summarizes results from the analysis of the RAWS data. The 0.95, 0.975, and 0.99 quantiles of the present climate RAWS data were 34.96, 42.97, and 51.54 respectively. In 2020 there were 32, 24, and 11 days with FWI values above those respective high thresholds. Out of

Table 2.2: As in table 2.1, results from Harbison Meadow, CO RAWS data using a transformed-linear $TL - MA(15)$ model.

Quantile	Threshhold	# in 2020	Past	Present	Ratio
0.95	34.96	32	0	0.0078	∞
			(0, 0)	(0.001, 0.032)	(∞, ∞)
0.975	42.97	24	0	0.0104	∞
			(0, 0.0004)	(0.001, 0.118)	(101.40, ∞)
0.99	51.54	11	0	0.0951	∞
			(0, 0.007)	(0.020, 0.398)	(22.63, ∞)

10,000 simulated present climate seasons there were 78, 104, and 951 that were high-risk. Zero of the past climate seasons were high-risk. The GPD used to transform the upper tail of the simulated past climate seasons has a bounded tail that is below each high quantile for the majority of the season. This suggests that the FWI values observed in 2020 were not plausible under past climate.

An additional analysis was performed which ensured that the extreme FWI values observed in 2020 were possible under past climate. We repeated the analysis but enforced Gumbel (unbounded but light) tails on the GPD. This has the added benefit of assessing sensitivity to the GPD parameters in the back-transformation. When the GPD is restricted in this manner there were 0, 0, and 8 simulated past climate seasons out of 10,000 that were high-risk.

As before, we summarize with the following attribution statement. Applied to RAWS data, our method estimates that the extreme fire weather observed in the 2020 was not possible under the observed climate of roughly 50 years ago. Our method rejects the null hypothesis that seasonal wildfire risk is unchanged between the two studied periods. When reanalyzed using an approach that forces a possibility of observing high-risk seasons under past climate, the method estimates that the risk of observing a season like 2020 is at least 138 times greater under the recently observed climate. The increased risk appears to be due to a shift in the marginal distribution at high quantiles (Figure 2.5) and an increase in the tail dependence (Figure 2.6).

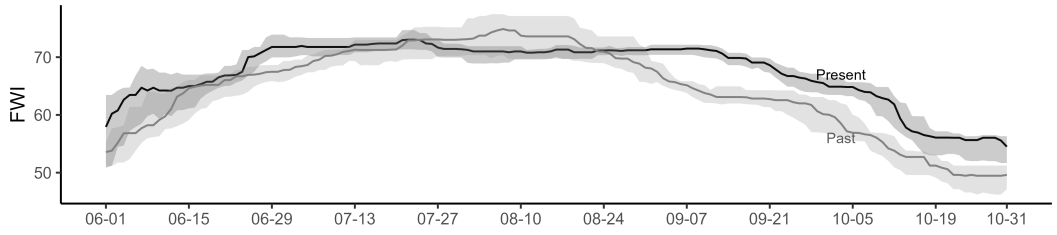


Figure 2.7: Daily 0.975 quantile of the FWI time series computed from the ERA5 data from Quincy California as in Figure 2.3.

2.6 Application to ERA5 data in California

2.6.1 Data and Pre-processing

We apply our method to an ERA5 grid-box containing Quincy California. This grid-box was burned in the 2020 North Complex fire (over 300,000 acres burned) and is adjacent to a grid-box burned by the 2018 Camp Fire, the deadliest fire in California history (Reyes-Velarde, 2019). There is a RAWS weather station in the grid box (Quincy Rd. station, NWS ID 040910) which could be used for comparison (this comparison was omitted for brevity).

The high quantile seasonality observed in Quincy, CA (Figure 2.7) is unimodal and does not show the large marginal shift between periods which we noticed in Colorado. The statistically significant change in the marginal distribution is that the season is longer; that is, the FWI time series are at the highest levels for a longer period of time in the present period than in the past period. Exploratory analyses show that the high quantiles of temperature are at the highest levels longer and precipitation and relative humidity quantiles are at their lowest levels for longer periods under present climate than they were under past climate.

We note that the tail dependence in the California region (Figure 2.8) seems to level off at shorter lags than in Colorado (Figures 2.4 and 2.6). This is likely because the dependence estimated by the TPDF is in the 'residuals' after accounting for seasonality. Compared to Colorado, there is less season-to-season variation in the FWI in California, and TPDF values seem to be driven more by variables (like ISI) with shorter time scales. TPDF values also appear slightly stronger and longer lasting under past climate than present climate. This may be attributable to the observed

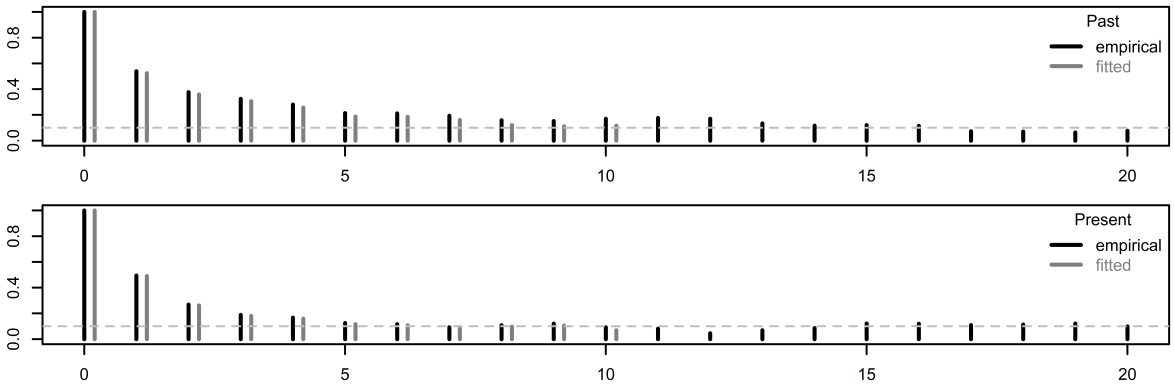


Figure 2.8: Empirical and fitted TPDF for Quincy, California ERA5 FWI time series as in Figure 2.4.

Table 2.3: As in table 2.1, results from Quincy, CA ERA5 data using a transformed-linear $TL - MA(10)$ model.

Quantile	Threshhold	# in 2021	Past	Present	Ratio
0.95	67.51	22	0.0081	0.0148	1.82
			(0.001, 0.022)	(0.002, 0.034)	(0.34, 9.45)
0.975	70.12	12	0.045	0.055	1.22
			(0.005, 0.092)	(0.003, 0.101)	(0.43, 3.41)
0.99	72.33	5	0.163	0.283	1.73
			(0.067, 0.302)	(0.077, 0.436)	(0.79, 2.68)

change in the fire behavior sub-index related to dryness (BUI) which has stronger and longer lasting tail dependence under current climate and thus may be included while accounting for seasonality. To assess sensitivity to model order, we have completed the analysis on models with order 5 and 10 and found similar results. The reported estimates are from using model order 10 for both periods.

2.6.2 Attribution

Table 2.3 shows results from our attribution study. The 0.95, 0.975, and 0.99 quantiles of the present climate ERA5 data in California were 67.5, 70.1, and 72.3 respectively. In 2021 there were 22, 12, and 5 days with FWI values above those respective high thresholds. Forty-five of the simulated past climate seasons were classified as high-risk at the 0.975 quantile compared to 55 of the present climate seasons.

Applied to FWI data generated from ERA5 output for the grid cell containing Quincy, California, our method estimates that a wildfire season like the one observed in 2021 is 1.22 to 1.89

times more likely under recently observed climate than under the observed climate of roughly 50 years ago. However, bootstrap confidence intervals contain one and thus our method fails to reject a null hypothesis of no change. It appears that under current climate, high quantiles of FWI are at higher values for a longer portion of the season, and bootstrap-based hypothesis tests support this conclusion.

2.7 Discussion

This chapter develops a relatively simple, computationally inexpensive, theoretically justifiable method to quantify how much more likely an extreme fire season is now than it was 50 years ago. We quantify the increase in risk in reference to a well-recognized high-risk season which enables easy communication of results; that is, we perform extreme event attribution of *seasonal* risk. Our method relies on time series models that specifically focus on extreme behavior. Because these models capture the pairwise tail dependence well, and because the method treats the two time periods in the same manner, our approach is useful for making meaningful comparisons.

Our method was applied to FWI data from two different locations and two different data sources. In the Grand Lake area of Colorado, our method estimates a dramatic increase in the risk of observing a fire season like the one observed in 2020, and this was seen in both reanalysis and weather station data. This increased risk is mostly attributable to an upward shift of the quantiles of the marginal FWI distribution. In California, our method's point estimates suggest an increase in the risk, but uncertainty associated with these estimates does not allow one to reject a null hypothesis of no change in seasonal risk. It is possible that analyzing a different fire weather index could result in different results. The FWI was developed in Canada for use with forests that are primarily Jack and Lodgepole pine, and it may not be the best index to summarize weather-driven fire risk in California. Fortunately, our same analysis could readily be repeated with any other indicator of fire risk. As this chapter only considers the fire risk associated with the meteorologically-derived FWI, it focuses specifically on risk due to changes in observed climate

and does not directly consider other factors such as past forest management which could contribute to overall fire risk.

Like any analysis, our method requires several modeling choices. Perhaps chief among these is the selection of the $TL - MA(q)$'s order. Our order selection method was based on subjectively interpreting TPDF plots; fortunately, sensitivity analysis showed that so long as q was chosen to be sufficiently large, conclusions were not affected. Another choice was to fit an $TL - MA(q)$ rather than a $TL - ARMA(p, q)$, which possibly could capture the dependence seen in the TPDF with a smaller number of parameters. We chose an $TL - MA(q)$ because we employed the innovations algorithm to fit our model, and model fitting of general $TL - ARMA$ models was an area for future investigation (see Chapter 4 for more). Our approach also employs a two-step estimation method, first fitting the marginal and transforming, and then assessing the dependence. This two step approach propagates any error from the first step forward; however our bootstrap method for assessing uncertainty accounts for this error propagation. One observed shortcoming of our model is that in Colorado, we observed seasons where the FWI remained moderate for the entire season, and simulated seasons all were extreme at least once during a season. This concern is consistent between climate periods and thus past-to-current comparison should still be relevant. As model fitting uses only large observations, our approach should only be used to assess extreme behavior and should not be used to assess quantities associated with the bulk of the distribution.

This chapter uses ERA5 data (European Centre for Medium-Range Weather Forecasts, 2023) which were accessed from the Climate Data Store (https://cds.climate.copernicus.eu/) and RAWS data (National Interagency Fire Center, 2023) which were downloaded using the "Weather Data" link on https://www.wildfire.gov/application/fire-and-weather-data-extract. Raw data and formatted files can be accessed on https://github.com/twixson/seasonal_wildfire_risk_attribution.

Chapter 3

Neural Network for Asymptotic Dependence/Independence Classification: A Series of Experiments

3.1 Introduction

Asymptotic dependence/independence is a property which summarizes the behavior in the joint tail of pairs of random variables. Consider the continuous bivariate random vector (X, Y) with marginal distributions F_X and F_Y and define

$$\chi(u) = P(F_X(X) > u \mid F_Y(Y) > u). \tag{3.1}$$

The pair of variables are asymptotically dependent (AD) if $\chi = \lim_{u \to 1} \chi(u) > 0$ and asymptotically independent (AI) if $\chi = 0$. Intuitively, this fundamental characterization of tail dependence describes whether the elements of a random vector can be at their most extreme at the same time. Deciding whether a dataset is AD or AI is an interesting classification problem as the distinction occurs in the limit, but the classification must be based on a finite sample. Classification as AD or AI is also important as extremes studies often aim to extrapolate into the tail beyond the range of the data; for example, to assess risk of the combined effect of extreme precipitation and storm surge. Modeling an AD random vector with an AI model will result in smaller magnitude predicted extreme events than an AD model, which could result in infrastructure design which is inadequate to withstand the combined effect of the variables. Conversely, using an AD model when the data are in fact AI would overestimate the magnitude of an extreme event and could result in overbuilding infrastructure: an unnecessary expense.

Characterizing tail dependence into these two regimes arises naturally from classical bivariate extreme value distributions (BEVDs) which can capture any level of AD but are only AI in the degenerate case of exact independence (i.e., a separable joint distribution function) (Coles, 2001). Consequently, classical extremes models such as the bivariate logistic (Gumbel, 1960), or max-stable process models (Kabluchko et al., 2009; Brown & Resnick, 1977) exhibit AD. More recent work has sought to develop models suitable for the AI setting, like the inverted max-stable models of Wadsworth & Tawn, 2012. Classification is a necessary precursor when choosing between models that are either AD or AI. Motivated by the difficulty of making an AD/AI decision, very recent work has led to more complex models intended to encompass both regimes (e.g., Wadsworth et al., 2017; Huser et al., 2017; Huser & Wadsworth, 2019; Bopp et al., 2021). Even with the availability of these regime-crossing models, we believe that AD/AI classification is a useful exploratory step due to the complexity of these models and their fitting procedures, and a clear answer to the AD/AI question can allow practitioners to use the more common and simple models.

A common approach to infer whether a data set comes from an AD (or AI) model is to plot empirical estimates of $\chi(u)$ at increasing levels u as u approaches 1 (Coles et al., 1999). Unfortunately, as $u \to 1$, the number of points used to estimate $\chi(u)$ goes to zero and thus the estimates that we are most interested in are the noisiest. Recognition that χ, and similar extremal dependence measures, were unable to describe the strength of dependence in the AI case led to the development of more nuanced measures of dependence in the AI case (e.g., η in Ledford & Tawn, 1996 and $\overline{\chi}$ in Coles et al., 1999). Like $\chi(u)$, an estimate of $\overline{\chi}$ can be plotted for increasing quantities of u, but each of these plots can only give evidence for one regime and thus practitioners may be left with seemingly conflicting or inconclusive evidence due to extrapolation from the noisiest quantiles (see, e.g., the wave-surge example in Coles et al., 1999). Another principled approach is hypothesis testing, which requires setting one regime as the null. This approach goes back as far as Gumbel & Goldstein (1964), and Dey & Yan (2016, ch. 17, 18) gives a review of several different tests (the null is AD in Draisma et al., 2004 and Einmahl et al., 2006 and is AI in Tawn, 1988, Ramos & Ledford, 2005, and Zhang, 2008). Hypothesis testing may provide evidence that

the null regime is implausible enough to choose the other regime but it cannot provide evidence for the null regime. These tools are useful but the challenge of distinguishing between asymptotically defined dependence regimes is inherently hard as information about the tail is sparse.

Our novel approach attempts to classify data sets as AD/AI using a supervised machine learning method. This machine learning approach differs from diagnostic plots and hypothesis testing in that we have not structured the classifier using the definition of AD but instead allow the classifier to learn the distinction from the training data. We want to see if machine learning methods can be a useful addition to the practitioner's toolkit as they try to determine which of these asymptotically defined regimes to use in the modeling of their finite sample.

In similarly motivated work, Ahmed et al. (2022) use a machine learning method to classify spatial processes as either AD or AI. There are a couple notable differences between these two investigations. First, we investigate bivariate data, whereas the classifier of Ahmed et al. (2022) learns from dependence behavior at multiple spatial distances. The information input into the classifier also differs: Ahmed et al. (2022) input tensors of $\hat{\chi}(0.975)$ and $\hat{\overline{\chi}}(0.975)$ computed for all pairs of locations across realizations of the generating process, whereas we input the original data. Both investigations however choose to use a convolutional neural network (CNN) as CNN's can take advantage of known structure in the data: the tensor structure in Ahmed et al. (2022) and the bivariate structure of our data.

Another recent area of work that combines extremes and machine learning uses neural networks as parameter estimators in a so-called amortized learning framework (see, e.g., Sainsbury-Dale et al., 2024 and references therein). Neural network-based estimation is particularly useful for extremes as many models have intractable likelihoods and these estimators are amortized in the sense that training is computationally expensive but, once trained, estimation is extremely fast and thus the per-use training cost can be minimal. While both our work and neural-network based estimation combines extremes and machine learning, the work differs in several important ways. First, the space that we explore is much larger than a single parametric family and, as such, we know that our training cannot cover the entire space. Instead, we use flexible parametric models

that cover dependence within regimes from independence to complete dependence and we assess perfomance with experiments on an expanding training and testing space. Second, the root of our challenge is not computational. We want to learn about the tail of the process underlying the data and use neural networks as universal function approximators to define a map between our finite sample and these asymptotically defined regimes.

In Section 3.2 we review the models used in the training data (Section 3.2.1) and the structure of the CNN that we used (Section 3.2.2). Section 3.3 details our first experiment in which we consider whether a CNN can distinguish between familiar models in each regime. Section 3.4 tests the generalizability of our trained CNN from Section 3.3 with additional models in both out-of-sample and in-sample settings. Section 3.5 considers generalizing our CNN to account for different sample sizes. In Section 3.6 we consider whether our CNN agrees with expert opinion. Finally we discuss limitations and our `R` package, `nnadic` in Section 3.7.

3.2 Preliminaries

3.2.1 Models Used in Training

In an effort to keep the current study at a reasonable length we have limited ourselves to the four following models: the Gaussian, Logistic, Inverted Logistic, and Asymmetric Logistic. The bivariate Students t distribution is included in this section as it is referenced in the paper but it is not used in the training of the CNN. The models should be understood as determining only the dependence structure (i.e. copula), as a transformation will standardize the marginals. Model/copula information is not included in the training of the CNN; the dependence regime (AI or AD) is the sole response.

The familiar bivariate Gaussian distribution is parameterized by the correlation parameter $\rho \in [0, 1]$. This family is perfectly dependent when $\rho = 1$, AI for all $\rho < 1$, and independent when $\rho = 0$. We do not consider negative dependence. The Gaussian model has $\chi = 0$ and $\overline{\chi} = \rho$ unless

it is perfectly dependent and can be defined using the pdf

$$f_{Gaussian}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right].$$

Perhaps the most common AD model is the logistic model of Gumbel (1960). The upper tail of this flexible one-parameter model can be perfectly dependent ($\alpha = 0$) or independent ($\alpha = 1$) and is AD as long as $\alpha < 1$. Notice that increasing α results in decreased dependence. The logistic model has $\chi = 2 - 2^\alpha$ and $\overline{\chi} = 1$ except under independence. On Frechet margins this model has CDF

$$G_{Logistic}(x,y) = \exp\left\{-\left(x^{-1/\alpha} + y^{-1/\alpha}\right)^\alpha\right\}, \quad x > 0, \quad y > 0, \quad \alpha \in (0,1].$$

The inverted logistic model of Wadsworth & Tawn (2012) is essentially the lower tail of the Logistic model. To generate samples from the inverted logistic model with exponential margins we first draw bivariate logistic random vectors and then invert them [i.e., for $(X_1, X_2) \sim$ Logistic(α), $(1/X_1, 1/X_2) \sim$ Inverted Logistic(α)]. The inverted logistic model has $\chi = 0$ and $\overline{\chi} = 2^\alpha - 1$ unless it is perfectly dependent and thus it is AI for all $\alpha > 0$. We can define the inverted logistic model by its survival function

$$\overline{G}_{Inverted}(x,y) = \exp\left\{-\left(x^{1/\alpha} + y^{1/\alpha}\right)^\alpha\right\}, \quad x > 0, \quad y > 0, \quad \alpha \in (0,1].$$

A natural expansion of the Logistic model is to relax the exchangeability of the two variables which is done in the asymmetric logistic model of Tawn (1988). It has CDF

$$G_{Asymmetric}(x,y) = \exp\left\{-(1-t_1)x^{-1} - (1-t_2)y^{-1} - \left((x/t_1)^{-1/\alpha} + (y/t_2)^{-1/\alpha}\right)^\alpha\right\},$$
$$x > 0, \quad y > 0, \quad \alpha, t_1, t_2 \in (0,1].$$

The model simplifies to the logistic model when the asymmetry parameters $t_1 = t_2 = 1$. It is AD unless $\alpha = 1$, $t_1 = 0$, or $t_2 = 0$. The asymmetric logistic model has $\chi = t_1 + t_2 - (t_1^{1/\alpha} + t_2^{1/\alpha})^\alpha$ and $\overline{\chi} = 1$ unless it is independent.

The generalization of the Student t distribution to the bivariate case is a distribution which is AD for all $\rho > -1$ and finite degrees of freedom ν but which converges to the bivariate Gaussian as $\nu \to \infty$. With finite ν and $\rho > -1$, $\chi = 2 * t_{\nu+1}\left(-\sqrt{(\nu+1)\frac{1-\rho}{1+\rho}}\right)$ where t_m is the lower tail of a t distribution with m degrees of freedom (Embrechts et al., 2002). We define the bivarate t distribution with the pdf

$$f_\nu(x,y) = \frac{\Gamma[(\nu+2)/2]}{\Gamma(\nu/2)\nu\pi\sqrt{1-\rho^2}}\left[1+\frac{x^2-2\rho xy+y^2}{\nu(1-\rho^2)}\right]^{-(\nu+2)/2}.$$

The preceding models do not have parameters that are directly comparable. In order to present results we use analytical χ and $\overline{\chi}$ (Coles et al., 1999) values to group models by their strength of dependence. We note that in all cases (except the Student t distribution) the boundaries of the dependence parameters (and thus of χ or $\overline{\chi}$) indicate independence and exact dependence: as the dependence parameters approach either boundary AD and AI models become more similar.

3.2.2 Structure of our Convolutional Neural Network

Neural networks are commonly used to approximate unknown non-linear functions. A deep neural network f is a composition of J simpler functions $f^{(j)}$ which are termed layers. These simpler functions typically pass a linear combination of the inputs to a non-linear activation function [e.g., $f^{(j)}(\mathbf{z}) = \max\left(\mathbf{0}, \mathbf{W}^{(j)T}\mathbf{z} + \mathbf{b}^{(j)}\right)$ which has weight matrix $\mathbf{W}^{(j)}$, biases $\mathbf{b}^{(j)}$, and a ReLU activation fuction (Jarrett et al., 2009, referred to as "positive part")]. The resulting composition of functions can be represented as a network of so-called neurons and edges. Possible introductory sources include Bishop (2006) and Goodfellow et al. (2016). Our neural network is designed to approximate the mapping from a set of m observations of the joint tail of a bivariate random vector to the asymptotically defined dependence regime of the generating model (AD or AI). We train the neural network on K sets of m two-dimensional points where the ith observation from the kth data set is denoted (x_{ki}, y_{ki}) and each data set is preclassified as AD or AI based on the kth generating model.

Because the classification of a set of data into either the AI or AD regime is a problem where the bivariate structure of the data is fundamental, we choose to use a CNN. A fully connected neural network would not know the bivariate nature of the data and thus would have unstructured vectors of values which include all of the x_{ki}'s and y_{ki}'s as inputs. A fully connected neural network would first have to learn the bivariate structure in order to perform the classification task in accordance with our statistical understanding of the problem. We use a CNN to ensure that our network exploits the relationship between the dimensions of each point by restricting the learning so that it first considers each (x_{ki}, y_{ki}) as a point.

Choices regarding the structure of our CNN were made in an *ad hoc* manner, which were tweaked until good results were obtained. Our CNN includes four convolutional layers and seven fully connected layers. Convolutional layers output 32, 32, 16, and 8 features respectively. Each convolutional layer uses one dimensional filters, and thus the first two layers output vectors of linear combinations of each point's dimensions. These linear combinations highlight different features of the distribution of (X_k, Y_k). Layers two, three, and four include a max-pooling step which, for layers three and four, allows for cross-point linear combinations. Outputs from the final convolutional layer are stacked and input into the first fully connected layer. Fully connected layers output dimensions are 32, 32, 32, 16, 16, 8, and 1 respectively.

All layers except the final layer are activated with the leaky relu function (Maas et al., 2013). The last layer is activated with the sigmoid (logistic) function which ensures output values are between zero and one. The middle five fully connected layers include 50% dropout to avoid "memorizing" the inputs. Parameter weights for our CNN are initialized with a uniform distribution that has limits set according to He et al. (2015) and initial bias terms are all 0.01. These weights and biases are optimized with the RMSprop algorithm (Tieleman & Hinton, 2012) with gradients computed using binary cross-entropy loss (i.e., Bernoulli log-liklihood). The learning rate is halved after 50 training epochs without improvement in the validation loss. We regularize (i.e., inhibit the learning to increase generalizability) the CNN by allowing for early stopping based on validation loss. We call our CNN `nnadic` as it is a **N**eural **N**etwork for **A**symptotic **D**ependence/

Independence **C**lassification. The final anatomy can be seen in detail in the code available on the `nnadic` github page.

In an attempt to improve results beyond our *ad hoc* approach, we performed an iterative grid search on number of layers, layer sizes, inclusion of bias on initial parameter weights, inclusion and proportion of dropout, learning rate, batch size, and optimizer. This iterative search gave similar results across most iterations (including networks with fewer layers); however, our *ad hoc* network outperformed the final iterative grid search network.

To generate training, validation, and testing data for the CNN we draw each model's dependence parameters uniformly from their support, simulate $n = 10000$ data points from each distribution on the natural margins for each dependence parameter, and then transform the margins to be unit-exponential. The l_∞-norm is used to select the most extreme $m = 500$ points (5% of generated points) from each set of simulated data. This process is demonstrated for one dataset from the Gaussian and Logistic models with dependence parameter set to 0.5 in Figure 3.1 and with another in Figure 3.2.

All of the following results reflect output from our final chosen CNN anatomy fitted to the datasets available in that experiment. These results are similar to results initially obtained using differing anatomies and allow for easier replication. This study used the `keras R` package for all neural network fitting and computing was done on a high performance computer and on Google's Colaboratory servers. Training took less than one hour to complete for each CNN.

3.3 Experiment 1: Can a CNN distinguish between bivariate Gaussian and Logistic data?

In our first experiment, we use bivariate Gaussian and logistic data and test whether a CNN can distinguish between these two models. Figure 3.2 demonstrates that this task may not be straightforward as column three has a datasets from each model and the dependence in the upper tail is not immediately distinguishable. We simulate 10000 Gaussian and 10000 logistic datasets as explained in Section 3.2.2. The datasets are randomly split into three groups: 80% were used for

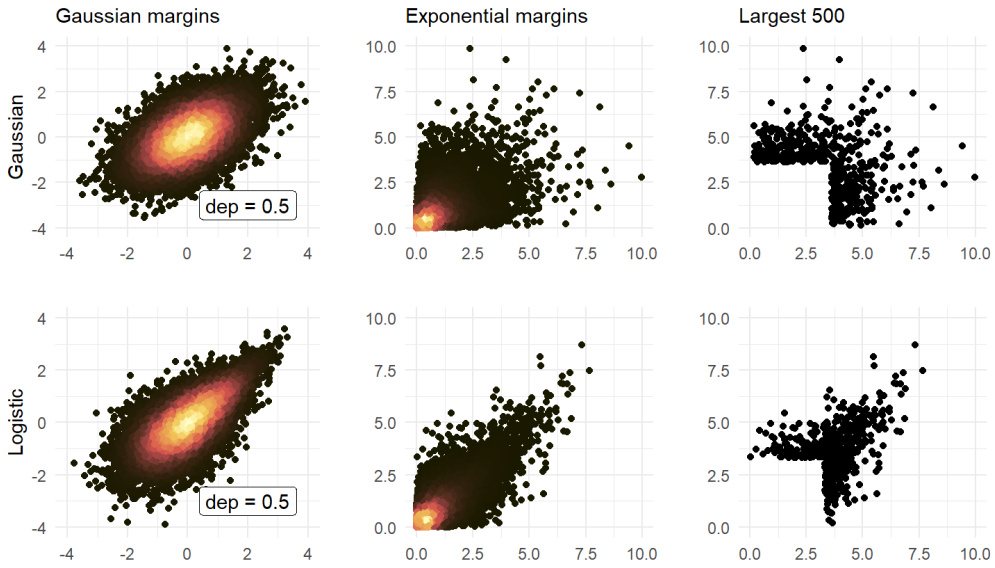


Figure 3.1: Scatterplots from Gaussian (top row) and Logistic (bottom row) copulas with dependence parameter set to 0.5 demonstrating the data generation and pre-processing steps. Column one shows 10000 points generated on the natural margins and column two shows the data after marginal transformation to unit-exponential. Column three indicates the large points that will be kept for training, validation, or testing with our CNN.

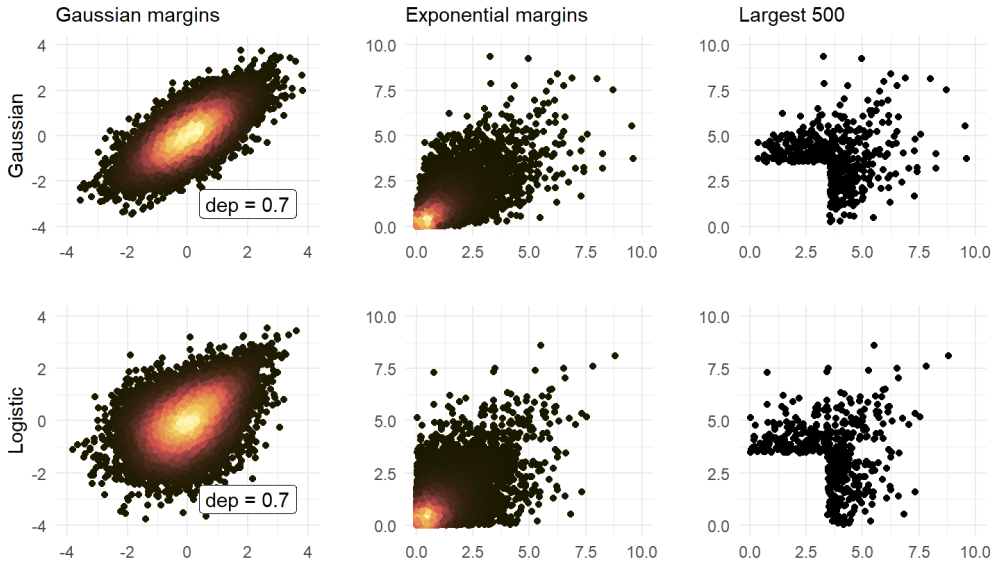


Figure 3.2: Scatterplots as in Figure 3.1. Here we set the dependence parameters so that the difference in the large points is less obvious to the human eye.

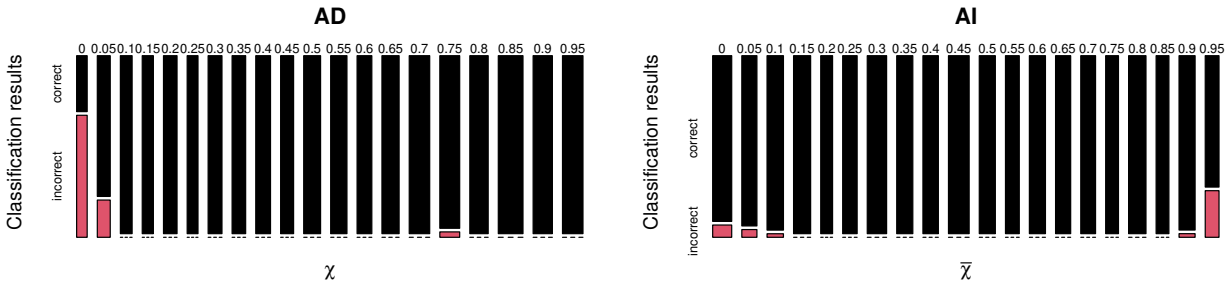


Figure 3.3: Proportion of test datasets accurately classified in experiment 1 (Section 3.3); black indicates correct classification and red indicates incorrect classification. The first plot includes logistic (AD) datasets split by analytic χ-values. The second plot includes Gaussian datasets (AI) split by analytic $\overline{\chi}$-values. Each vertical bar represents an interval (i.e., the first bar in the left plot includes results from test datasets generated from the logistic model with $\chi \in [0, 0.05)$).

training (thus $K = 0.8 * 10000 * 2$ models = $16,000$ datasets), 10% ($2,000$ datasets) for validation, and 10% for testing. A quick check suggested that approximately half of each group was Gaussian and histograms of dependence parameters for each group (not shown) appeared approximately uniformly distributed.

The CNN returns a value between zero and one for each dataset in the testing group. This value can be thought of as the CNN's predicted probability that the dataset was generated from a model which is AI. We use a cutoff of 0.5 to turn these outputs into a prediction of AD or AI, but our results are not sensitive to this cutoff as the CNN is very confident in its predictions: less than 10% of the output values are between 0.1 and 0.9.

Trained on these data, our CNN is able to accurately classify over 97% of test datasets (Figure 3.3). There are 19 Gaussian datasets which are classified as logistic and 35 which are misclassified in the other direction. The majority of datasets which are not accurately classified have parameter values close to the boundaries and thus are very close to being exactly dependent or independent. These cases are clearly the most difficult to distinguish as the two copulas become identical when the dependence parameter reaches either extreme as indicated by Figure 3.4. Output values from the CNN reflect this uncertainty: 29 of the 54 incorrect classifications have output values between 0.2 and 0.8.

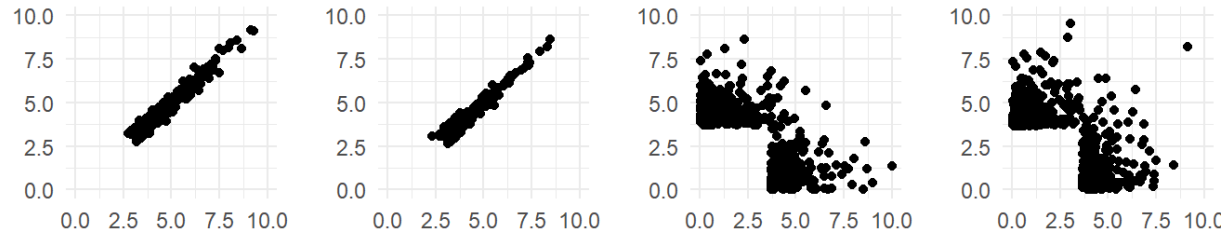


Figure 3.4: When dependence is very strong or very weak the copulas become identical. Scatterplots as in column three of Figure 3.1. Plots one and three are from the Gaussian model with dependence parameters 0.995 and 0.05 respectively. Plots two and fourare from the Logistic model with dependence parameters 0.1 and 0.94 respectively.

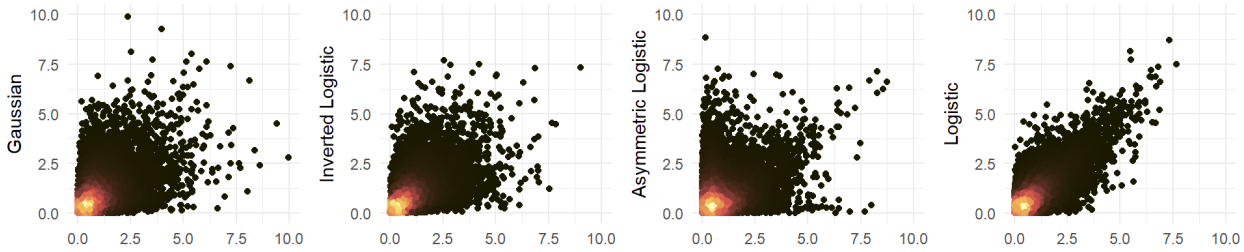


Figure 3.5: Scatterplots of the four models used in Experiment 2 on unit-exponential margins.

3.4 Experiment 2: How does the CNN perform on other models?

3.4.1 Can the CNN predict AD/AI for models outside its training set?

We begin this three-part experiment with an assessment of the generalizability of the CNN that was fit in Section 3.3. We use the CNN trained only on Gaussian and Logistic data to classify asymmetric and inverted logistic datasets. This out-of-sample test mimics real-world use as practitioners data are not generated in the same manner as training data. One example of a dataset from each model can be seen in Figure 3.5.

The CNN accurately classifies 86% of these 2,000 test datasets (Figure 3.6). The misclassified datasets are primarily asymmetric logistic datasets (213 of the 288). This model has three parameters and each has a boundary which, when reached, results in independence. All hyperparameters

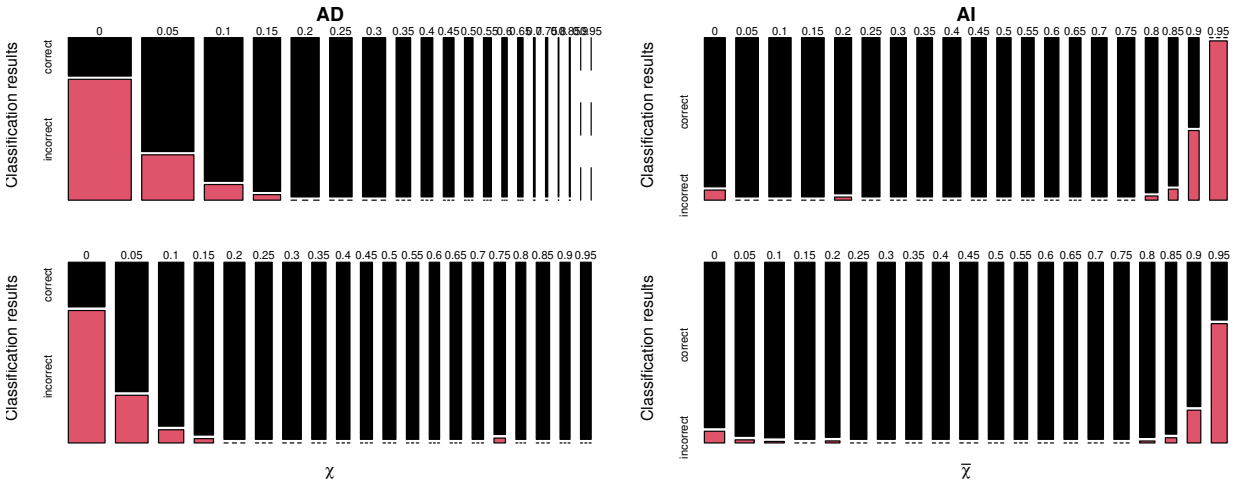


Figure 3.6: Proportion of test datasets accurately classified in experiment 2.1 (Section 3.4.1). Top: Out-of-sample classification results of asymmetric and inverted logistic test datasets. Bottom: Classification of all test datasets (i.e., combining Top with 3.3) for comparison with 3.7. Each vertical bar represents an interval (i.e., the first bar includes results from asymmetric logistic test datasets with $\chi \in [0, 0.05)$).

are independently generated from a uniform distribution and thus the χ-values are skewed toward the smaller values (10% are smaller than 0.05). In addition to these very small χ-values we note that a larger proportion of inverted logistic datasets with large $\overline{\chi}$-values are misclassified. Compared to experiment 1, the CNN is much more confident in these incorrect predictions, as over half of the misclassified datasets have output values smaller than 0.1 or greater than 0.9.

3.4.2 Does training with more models improve the results?

While we hope to learn something about the two asymptotically-defined regimes, the results from Section 3.3 and Section 3.4.1 may tell us more about the Gaussian and Logistic copulas than they do about the dependence regimes. We continue the learning by training CNN's on all four models in two ways. First we perform a two-stage training wherein we take the parameter values from the fitted CNN in experiment 1 as the initial values in the training of a CNN with all four models. Second we train a new CNN from scratch using the same anatomy as in experiment 1. The difference in results between the two training methods is negligible and thus we only report results from the two-stage CNN which correctly classified 5 more datasets than the all-at-once approach.

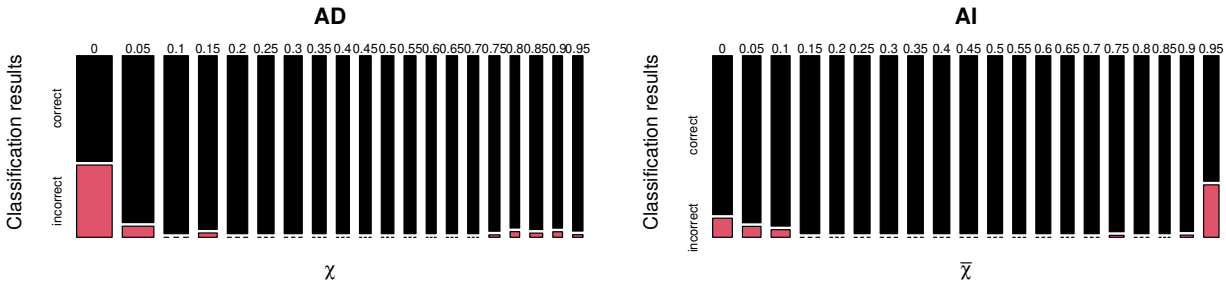


Figure 3.7: Proportion of test datasets accurately classified in experiment 2.2 (Section 3.4.2). Two-stage training with all four test sets by χ (left) and $\overline{\chi}$ (right). Each vertical bar represents an interval (i.e., the first bar includes results from asymmetric logistic test datasets with $\chi \in [0, 0.05)$).

The CNN is able to accurately classify more than 95% of the test datasets (Figure 3.7) after inclusion of the asymmetric and inverted logistic datasets in the training. Here, the CNN is less confident about the misclassified datasets than before; over 85% of the misclassified output values are between 0.2 and 0.8. Just under a third of the misclassified outputs are AI datasets which is due, in part, to the disproportionate number of AD datasets which are nearly exactly independent (i.e., very small χ values). Of the 174 misclassified datasets 93 were asymmetric logistic, 34 were inverted logistic, 26 were logistic, and 21 were Gaussian.

3.4.3 How does the CNN perform with the bivariate *t* distribution?

We continue this exploration of other models with an experiment using data generated from a bivariate t distribution. This distribution does not have the independence copula on the boundary of the ρ parameter space but converges to the Gaussian distribution as $\nu \to \infty$. We fix $\rho = 0.7$ and generate 2000 test datasets at each of three different degrees of freedom ($\nu = 2, 10,$ and 30). These parameter combinations have analytic χ values of 0.52, 0.19, and 0.03 respectively. The CNN that is trained on the four models from in Section 3.4.2 correctly classified all 2000 of the datasets that had two degrees of freedom, about a third (627 datasets) of the datasets with ten degrees of freedom, and only 21 of the datasets with 30 degrees of freedom. We believe the CNN is unable to distinguish data arising from a t distribution with 30 degrees of freedom from the asymptotically independent data generated from a Gaussian distribution.

3.5 Experiment 3: Can we generalize to different sample sizes?

Experiments 1 and 2 were run with training, validation, and testing datasets which contain the most extreme (by l_∞-norm) $m = 500$ points from a sample of $n = 10000$ points. This helps answer whether a machine can accurately distinguish between AD and AI in one specific finite-sample case but it does not reflect practitioners' use. In practice one does not have the luxury of pre-selecting the sample size, and varying the input size of a CNN is challenging: we do not want to re-train the CNN for every sample size. For this third experiment we consider whether the CNN from Section 3.4.2 (`nnadic`) can perform well on datasets with fewer/more than 10000 points. In addition to assessing the generalizability across sample sizes, this experiment assesses whether increasing the sample size increases classification accuracy.

Central to this experiment is an investigation of how best to go from different size datasets to the 500 points needed to input into the trained CNN. We investigate when both n is larger or smaller than 10000. For each sample size n, we generate 2000 test datasets from each of the four models used in training. These datasets are then subset with several methods to assess how to automatically provide `nnadic` with the most useful 500 points for classification. The methods we investigate are zero-padding, largest 500, and variations on sub/resampling methods.

It is common practice with CNN's to pad inputs with zeros when the test data dimensions are smaller than the training data. We investigate a zero-padding method which takes the largest $0.05 * n$ points and fills in the other $500 - 0.05 * n$ points with zeros. Padding with zeros performs poorly, likely because adding points at the origin introduces input points in the test data sets which are unlike any points appearing in the training data. The zero-padding method resulted in AD predictions for almost all datasets, and we do not report further results from this method.

Another straightforward method uses the largest 500 points as `nnadic` inputs, regardless of whether n is larger or smaller than 10000. Results appear in Table 3.1. Classification performance for `nnadic` declines as n decreases from 10000. One aspect of this is likely due to the situation, common to extremes, that it becomes increasingly difficult to characterize a distribution's (joint) tail as sample size decreases. However, another affect arises in that retaining the largest 500 points

Table 3.1: Results from experiment 3. Columns are the total number of points generated in each dataset (n); the number of points above the 0.95 quantile (m); the accuracy across the 8000 test datasets (2000 from each of four models) when the largest 500 points from each dataset are used as the test points; and the accuracy when using the Re/Sub-sampling method of getting 500 test points.

n	m	Largest 500	Re/Sub-sample
500	25	0.50	0.79
1000	50	0.50	0.85
2500	125	0.56	0.91
5000	250	0.87	0.93
7500	375	0.94	0.94
*10000	500	0.95	.
12500	625	0.94	0.94
15000	750	0.93	0.94
20000	1000	0.91	0.94
30000	1500	0.90	0.94
50000	2500	0.88	0.95

* This row is the original testing data which is the correct dimensions so that no re/sub sampling is needed.

as n decreases means that points are retained whose joint values differ considerably from the training data which all had l_∞-norms greater than the empirical 0.95 quantile. `nnadic` is is barely better than random chance when using the method which retains the largest 500 points from datasets of size 2500. This marginal mismatch also appears to affect performance as n increases away from 10000. When $n = 50000$, the only 88% of the datasets were correctly classified, compared to 95% when $n = 10000$ and matched the training sets' size. This is interesting not only because statistical methods' performance typically improves with sample size, but also because retaining the largest 500 points from a larger data set implies that this data should be less affected by higher order effects that interfere with estimating asymptotic quantities like χ or $\bar{\chi}$. This illustrates how `nnadic` is a non-asymptotic classifier in contrast to traditional methods like plotting $\hat{\chi}(u)$ or $\hat{\bar{\chi}}(u)$.

The method with best classification performance obtains 500 points in the tail by resampling the largest $0.05 * n$ points when $n < 10000$ and subsampling the points above the estimated 0.95 quantile when $n > 10000$. We explored three different methods of resampling which results in repeated points: (1) randomly drawing 500 points with replacement from the largest $0.05 * n$ points, (2) including the first $0.05 * n$ points and then resampling those same points to fill in the rest,

and (3) taking the largest multiple of $0.05 * n$ which is not larger than 500 (i.e., $\lfloor 500/(0.05 * n) \rfloor$) and including that number of copies of all large points, then randomly filling any remaining slots. Method 3 is the most balanced in the sense that all large points (those with l_∞-norms greater than the empirical 0.95 quantile) are included the greatest number of times possible. The largest difference in accuracy between the three resampling methods is less than three percentage points and thus results in Table 3.1 are from the best (and simplest) method (1) which re-samples the largest $0.05 * n$ points 500 times with replacement.

For data sets with smaller sample sizes (n < 10000), `nnadic` is able to accurately classify data from reasonably small sample sizes using the resampling method (Table 3.1). The accuracy drops off as the sample size decreases but is still nearly 80% when the top 25 points from datasets with sample size 500 are re-sampled. When $n > 10000$, the subsampling method results in datasets which are similar to our training data (but with a better estimate of the 0.95 quantile), and the correct classification rate seems to be about 95% regardless of sample size. It is again interesting to contrast this behavior with a more typical statistical procedure whose performance should improve as a larger sample provides improved evidence for tail characterization. We note that the `nnadic` package (see Section 3.7) allows for practitioners to test whether different re/sub-samples give different results on their data.

3.6 Experiment 4: Does `nnadic` agree with expert opinion?

Our final experiment applies the re/sub-sampling method from Experiment 4 to several data sets which have appeared in the literature and in which the authors chose between AD and AI regimes. The aim of this experiment is to see if `nnadic` chooses the same regime as the author. We consider six datasets from four papers: the rainfall and wave-surge data from Coles et al. (1999) (accessible in the `ismev` `R` package), the Santa Ana wind speed and negated relative humidity data from Cooley et al. (2019), the Danube river basin data in Asadi et al. (2015, and others), and the summer and winter air pollution data from Heffernan & Tawn (2004) (accessible in the `texmex` `R` package). Table 3.2 summarizes the results which are explained in detail below.

Table 3.2: Comparison of expert opinion and `nnadic` output on five different datasets. The Danube river basin numbers represent station numbers.

	Rainfall	Wave-surge	Santa Ana	Danube 1 - 2	Danube 11 - 12	Summer O_3 - NO	Summer NO - PM_{10}	Summer 8 others
Expert	AI	AD*	AD	AD	AD	AI	AI	AI
`nnadic` prediction	AI	AD	AI*	AI	AD	AD	AD	AI
Number AI	100	0	71	100	0	0	0	800

* Indicates evidence leans toward this regime.

Coles et al. (1999) uses $\chi(u)$ and $\overline{\chi}(u)$ plots to classify a dataset of daily rainfall amounts at a location in England. These data are a time series so we consider lag-1 pairs of points of which there are 17530. The authors conclude that these data are AI. Points are sub-sampled 100 different times and input into `nnadic`, and all 100 datasets are classified as AI with our method (Table 3.2).

Coles et al. (1999) also investigates a dataset of 3-hourly measurements of wave height and storm surge from Newlyn, England. The authors consider the $\chi(u)$ and $\overline{\chi}(u)$ values as inconclusive, and choose an AD model. These data consist of 2894 bivariate observations and we re-sample the points above the empirical 0.95 quantile 100 times to get 100 datasets with 500 large points each. This allows us to assess the sensitivity to which points are duplicated. All 100 datasets are classified as AD (Table 3.2).

The Santa Ana wind dataset consist of 3902 daily observations of windspeed and negated relative humidity at a weather station in southern California. These data are classified as AD in Cooley et al. (2019) using a $\chi(u)$-plot which is truncated at a large quantile beyond which they determine there is too much noise in the empirical estimates. We re-sample 100 different datasets which each have 500 points that are above the empirical 0.95 quantile. `nnadic` classifies 71 of the 100 datasets as AI.

The Danube river basin data are modeled in Asadi et al. (2015) using a max-stable process which assumes AD. The dataset contains 4692 daily observations at 32 stations and include both spatial and temporal dependence as the different variables are stations which are in the same basin and lie up/downstream of other stations in the dataset. We look at daily observations from pairs of stations. Stations 1 and 2 which lie directly downriver from one another are classified by `nnadic`

to be AI in all 100 datasets. However, stations 11 and 12 which also lie directly downriver are classified as AD in all re-sampled datasets.

The air pollution data in Heffernan & Tawn (2004) consist of daily observations of five pollutants from Leeds city centre, UK: O_3, NO_2, NO, SO_2, and PM_{10}. The data are split into summer and winter datasets which have 578 and 532 observations respectively. The authors conclude that all pairs of pollutants in the summer season are AI, whereas `nnadic` classifies two of the ten summer pairs (O_3 - NO and NO - PM_{10}) as AD in all 100 re-sampled datasets. In the winter season (not shown in Table 3.2) the authors state that there is only evidence of AD in the NO_2 - NO, NO_2 - PM_{10}, and NO - PM_{10} pairings. Our method classifies five of the ten winter pairs as AD, and only one of these five AD pairs (NO_2 - PM_{10}) agrees with the opinion of the experts. These data provide the most disagreements between `nnadic` classifications and expert opinion, which probably demonstrates that determining the dependence regime of datasets of less than 600 points is quite challenging.

3.7 Discussion and `R` package

In this chapter we have discussed a series of experiments which were aimed at determining whether a CNN can accurately classify data as AD or AI. We find that our CNN performs very well on data generated from the models used in its training. We note that `nnadic` struggles under very high or very low dependence and these cases (easily checked with a scatterplot) demand caution. The classification accuracy, across dependence parameters, of our CNN is robust to reasonable changes of sample size.

Our classifier differs from hypothesis testing in that it is *a priori* agnostic about regime. The definition of these asymptotically defined regimes was not used in the training of `nnadic` and, in turn, increasing the sample size does not decrease the expected error rate (Table 3.1). This differs from both diagnostic plots where increasing the sample size will decrease noise at high quantiles and from hypothesis testing where larger sample size leads to increased power. Because our approach does not rely on asymptotic definitions or benefit from increased sample sizes, our

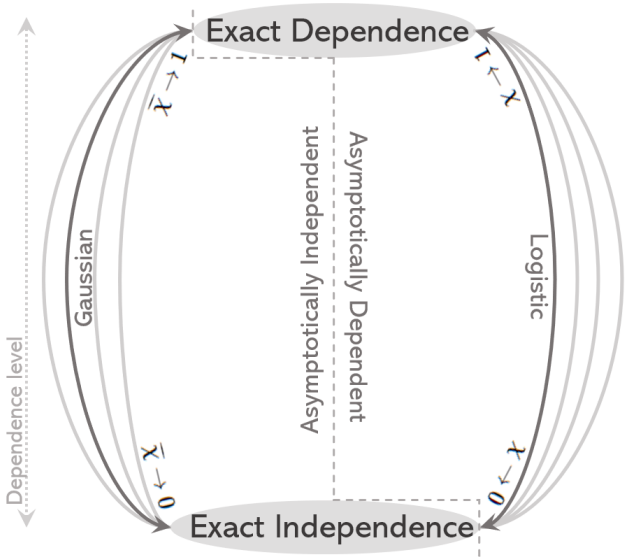


Figure 3.8: Heuristic map of dependence paths between exact independence and exact dependence.

CNN is a finite-sample classifier. While `nnadic`'s accuracy is not tied to sample size the use of this tool requires the practitioner to assume that the tail dependence structure of their data is sufficiently close to something in the training data to give useful results. In addition to any evidence provided by `nnadic` or any other tool, the practitioner should consider the consequences of choosing the wrong modeling regime.

It is tempting to think of AI and AD as existing on a spectrum since, as $\chi \to 0$ one approaches asymptotic independence, and as $\overline{\chi} \to 1$ one approaches asymptotic dependence. However for many AD models, as the model's dependence parameters drive the value of χ toward zero, the model approaches *exact* independence. Likewise, for many AI models as $\overline{\chi}$ is driven toward 1, the model approaches *perfect* dependence. Our CNN results highlight the fact that different models exist on different continuous paths from exact dependence to independence, as we try to illustrate in Figure 3.8.

On the other hand, the bivariate t distribution does not fit nicely into Figure 3.8. For fixed $\rho > -1$, the bivariate t distribution is AD for any finite ν and it approaches asymptotic independence (but not exact independence) as $\nu \to \infty$. As seen previously `nnadic` cannot reliably distinguish between the two regimes with vanishing χ and the bivariate t distribution is no exception (Section

3.4.3). We note that an important distinction between the bivariate t distribution and the other AD models used here is that it is not max-stable.

Our CNN is an addition to the array of tools that practitioners can use to inform their decision regarding the asymptotic dependence regime of their data. As such, we have developed a package which seamlessly transforms the marginal distribution, re/sub-samples the largest points, and classifies the data. The `get_nnadic_input()` function prepares data according to the users preferences and the `nnadic()` function classifies the dataset(s) and plots a histogram of the output values. The package, download instructions, and a brief tutorial are available at `https://github.com/twixson/nnadic`.

Data are available on `https://github.com/twixson/nnadicTestData` and everything is available at `https://zenodo.org/records/10870040`. References for data used in Section 3.6 are included in that section.

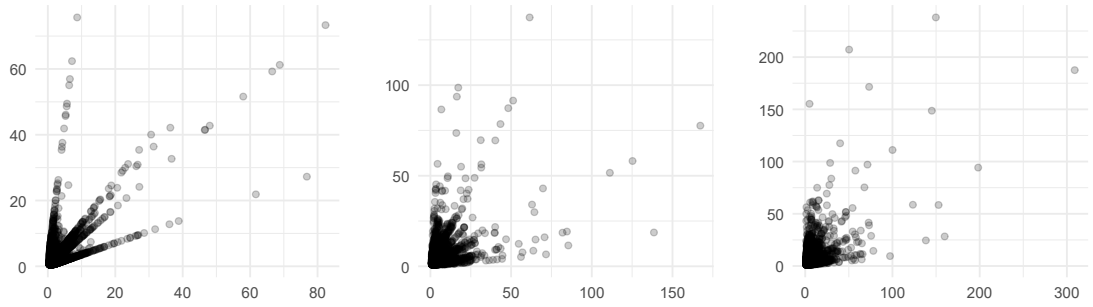
Chapter 4

A Proxy-likelihood Estimator for Multivariate Extremes Models with Intractable Likelihoods

4.1 Introduction

Some models for extremes cannot be fit with likelihood methods because the density does not exist or is unknown. Some of these models have broad appeal due to the simplicity of construction, ease of simulation, or interpretability. One such example is the family of max-linear models. These models have discrete angular measures with mass on k points (Figure 4.1). Fougères et al. (2013) showed that any angular measure can be approximated arbitrarily well by the angular measure of a max-linear model if the number of point masses is large enough and thus any regularly varying model can be approximated by a max-linear model. Max-linear models show up in causal extremes (see, e.g., Gissibl & Klüppelberg, 2018), k-means clustering of extremes (Janßen & Wan, 2020), and others. The discrete angular measure of this popular family shows up as a maximum operator in the exponent measure of the distribution and thus the derivative (and therefore density) does not exist. Without a likelihood, practitioners must use alternative fitting methods, like least-squares estimators (see, e.g., Einmahl et al., 2018).

The transformed linear extremes time series (TLETS) models of Mhatre & Cooley (2024) (employed in Chapter 2) are a second class of models which cannot be fit with likelihood methods. TLETS models are ARMA-like but are combinations of regularly varying noise terms (Section 2.3). In particular, these noise terms are regularly varying with tail index 2 as this simplifies the tail dependence metric and allows for some covariance-like properties (Cooley & Thibaud, 2019; Mhatre & Cooley, 2024). The challenge in this case is because the distribution of a transformed-linear sum of regularly varying random variables with tail index $\alpha = 2$ is unknown. In general, it is hard to convolve regularly varying random variables with tail index $\alpha = 2$ as this is a boundary

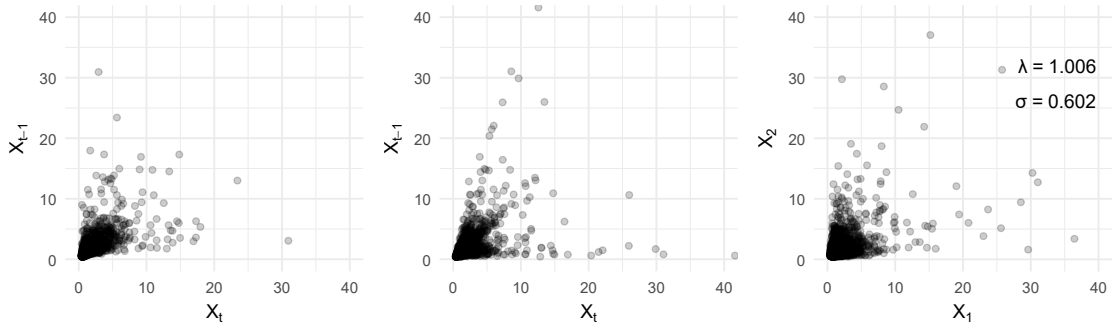


(a) $k = 5$ **(b)** $k = 15$ **(c)** $k = 25$

Figure 4.1: Max-linear models have discrete angular measures. Point clouds were generated with 10000 points from max-linear modes with $k = 5, 15, 25$ point massess respectively.

point where sums converge to be normal and thus there is no α-stable distribution with tail index 2 (see, e.g., Nolan, 2020). The transformed-linear operations (Section 2.3.1) complicate the picture beyond the challenge which comes from this boundary. These operations are employed in the TLETS models as they enhance flexibility and ensure that the time series is non-negative (Mhatre & Cooley, 2024). These models have been fit with method-of-moments (Chapter 2) and least-squares-type methods (see, e.g., Mhatre & Cooley, 2023) which lack some of the nice properties of likelihood methods including natural model selection methods.

We seek a likelihood-like model-fitting method that can be used with extremes models with intractable likelihoods. Our method involves obtaining an objective function that can be treated like a likelihood for fitting and model selection. We use the likelihood from a second parametric family of models as a proxy for the likelihood of the model(s) that we want to fit. This family of proxies must have the same marginal tail behavior as our desired models (and our transformed data). The proxy model is linked to our desired model through the tail pairwise dependence (TPD, Cooley & Thibaud, 2019) which is a second-order summary of the dependence in any regularly varying model and is defined in (2.2). The proxy model must have a likelihood and, because we use the TPD as a link, both models must be regularly varying. Fitting in this way can be thought of as minimizing the KL-divergence between the misspecified model and the data. We will demonstrate



(a) ERA5 lag-1 FWI **(b)** ARMA$(1, 1)$ generated lag-1 **(c)** HR generated points

Figure 4.2

our method on the TLETS models from Chapter 2 but our method can be readily applied to any regularly varying model.

Both max-linear and TLETS models are useful models with nice properties but the angular measures often do not match angular measures in data. Figure (4.2a) shows the scatter plot from the lag-1 Fire Weather Index time series in Wixson & Cooley (2023). Figure (4.2b) shows lag-1 data generated from the TLETS ARMA$(1, 1)$ model fitted to the data in Figure (4.2a). The angular measure for TLETS models are concentrated on point masses though this concentration is fuzzy in observations of the series because of the transformed-linear operations (in Figure (4.2b) mass is concentrated near the horizontal axis, near the identity ray, and near the ray with slope four).

Many models for multivariate extremes, like the bivariate Hüsler-Reiss (HR) distribution which we will introduce in Section 4.3, have continuous angular measures. The angular measure of these models may match the data better than models with discrete angular measures. Figure (4.2c) is a scatter plot from a bivariate HR distribution that has the same TPD as the fitted TLETS ARMA model at lag-1.

In this work we focus on capturing a second-order property of the dependence. The angular measures in Figure (4.2) are all different but the point clouds have the same second-order dependence as measured by the TPD; the model TPD in Figures (4.2b) and (4.2c) is the same as the estimated TPD in Figure (4.2a). The TPD (termed TPDF in Section 2.2 and equation 2.2) is a summary measure of the dependence in the pairwise tail of any multivariate regularly varying random

vector and is analogous to the autocovariance in classical time series models. We want to perform likelihood-like fitting and inference based on this second-order property of the dependence of any multivariate regularly varying model.

Focusing on second-order properties is a common practice in the statistical literature. Least-squares estimators in regression and time series analyses focus on these properties and only assume a model for the error when needed for estimating the uncertainty. In this way the Gaussian likelihood is often thought of as an objective function which provides useful information whether or not we are actually willing to assume the errors are Gaussian. We view our method in the same framework. Recall that knowing the covariance in the Gaussian case means that you know the full dependence structure but knowing covariance in general does not fully describe the dependence. This same idea holds when considering tail dependence as can be seen in Figure (4.2). It will be seen below that the TPD determines the full dependence structure of our chosen proxy-likelihood.

This chapter uses the HR distribution as our proxy model. The primary reason we chose the HR model (besides having a likelihood) is because it is parameterized in a manner that eases the translation between our desired model parameters and the parameters of the proxy model. The HR distribution (and related Brown-Resnick process) is parameterized by a dependence matrix (function) that has a parameter for each pair of variables. In other words, the ij entry of the matrix describes the dependence between the i^{th} and j^{th} components of the model. Furthermore, the model has a closure property by which we mean that lower dimensional margins of a $d > 2$ dimensional HR distribution are HR distributions with dependence parameters equal to the respective entries in the higher-dimensional dependence matrix. This closure property suggested that the TPD information is contained in the bivariate parameter because Mhatre & Cooley (2024, proposition 2.2) showed that the TPD computed from the bivariate marginal distribution of a regularly varying random process is equal to the TPD computed from any higher dimensional marginal distribution. We will see in section 4.3.1 that there is a bijection between the dependence parameter of a bivariate HR distribution and the TPD (figure 4.3).

Using the HR model as a proxy for our desired model allows us to write down a likelihood but it does not overcome the challenge which faces all max-stable processes; the number of terms in the likelihood grows combinatorially with the dimension of the problem. Many authors have devised approximate likelihood methods (see, e.g., Huser et al. 2016 and references therein) that have similar properties but are less efficient than full likelihood methods. One common approach that was popularized by Padoan et al. (2010) is to use a composite likelihood, which considers the product of lower-dimensional (e.g., pairwise) likelihoods as an approximation of the full joint likelihood. We follow suit and rely on the pairwise composite likelihood approach to overcome this challenge.

We emphasize that building an objective function out of bivariate HR distributions does not give us a likelihood as the model is misspecified and the composite likelihood approach overuses the data. Instead we term the objective function a proxy-likelihood as we use it in place of the likelihood for fitting the models and selection of a parsimonious model that fits well without over-fitting. We expect that, despite the misspecification, this method will prove useful for the given tasks. Development of the method suggests a way to construct confidence intervals based on the surface of the loss function which is a common motivation for likelihood methods. Initial simulations indicate that intervals created with standard arguments are heuristic at best.

The rest of this chapter is as follows. In section 4.2 we review the desired models that we will use to demonstrate our method (TLETS models) and define their TPD functions. In section 4.3 we describe our proxy model and the TPD link. Section 4.4 defines our composite likelihood and section 4.5 discusses model selection. In section 4.6 we discuss two methods of ensuring large points inform about the tail dependence. Simulations and application to the wildfire data of Chapter 2 are in section 4.7 and we close the chapter with a discussion of the method.

4.2 Transformed Linear Extremes Time Series

The transformed-linear extremes time series models (TLETS) of Mhatre & Cooley (2024) are the motivating family of models for this work. These models are built on the framework of regular

variation (Section 2.2) and are introduced in Section 2.3. Here we introduce the transformed linear auto regressive and transformed linear auto regressive moving average models.

4.2.1 Auto Regressive Model

A time series $\{X_t\}$ is a transformed-linear auto regressive process of order 1 (denoted $TL - AR(1)$) if, for all t,

$$X_t = \phi \circ X_{t-1} \oplus Z_t \tag{4.1}$$

where $|\phi| < 1$. When the marginal distribution of X_t has scale 1 the TPD at lag-h from a $TL - AR(1)$ is

$$\sigma(h, \phi) = \max(0, \phi^h). \tag{4.2}$$

4.2.2 Auto Regressive Moving Average Models

The transformed-linear auto regressive moving average models of order $p = 1$, $q = 1$ (denoted $TL - ARMA(1, 1)$) is given by

$$X_t \oplus (-\phi) \circ X_{t-1} = Z_t \oplus \theta Z_{t-1} \tag{4.3}$$

where $\phi, \theta \neq 0$ and $\theta + \phi \neq 0$. The TPD function for the ARMA(1,1) model is

$$\sigma(h, \theta, \phi) = \begin{cases} \frac{(\phi+\theta)\phi^{h-1}(1+\phi\theta)}{1+2\theta\phi+\theta^2} & \text{if } \phi > 0, \phi + \theta > 0 \\ 0 & \text{if } \phi > 0, \phi + \theta < 0 \\ \frac{(\phi+\theta)^2\phi^h}{1-\phi^4+(\phi+\theta)^2} & \text{if } \phi < 0, \phi + \theta > 0, h \text{ is even} \\ \frac{(\phi+\theta)\phi^{h-1}(1-\phi^4)}{1-\phi^4+(\phi+\theta)^2} & \text{if } \phi < 0, \phi + \theta > 0, h \text{ is odd} \\ \frac{(\phi+\theta)\phi^{h-1}(1+\theta\phi^3)}{1+\theta^2\phi^2+2\theta\phi^3} & \text{if } \phi < 0, \phi + \theta < 0, h \text{ is even} \\ 0 & \text{if } \phi < 0, \phi + \theta < 0, h \text{ is odd.} \end{cases} \tag{4.4}$$

Here we note a difference in the TPD from what was computed in the supplement to Mhatre & Cooley (2024). In case 1 ($\phi > 0, \phi + \theta > 0$) we have ϕ^{h-1} rather than their ϕ^h. The derivation in appendix A.5 of the version on arXiv (Mhatre & Cooley, 2021) has a mistake in the transition from line two to line three. Since $\phi \neq 0$ we can factor out ϕ^h as they did but this would leave a $1/\phi$ term not a $1/\theta$ term. The result is simpler if instead we factor out ϕ^{h-1} from line two.

4.2.3 TL-ARMA models of other orders

Mhatre & Cooley (2024) showed that a causal $TL - ARMA$ model of any order can be represented by a $TL - MA(\infty)$ model. The TPD function of the $TL - ARMA$ model is given by the coefficients of the $TL - MA(\infty)$ representation. The same is done in the computation of the auto covariance function for a classical $ARMA(p, q)$ model in Brockwell & Davis (2002). In practice if an $ARMA(1, 1)$ is not able to capture the dependence we can instead choose some large order MA model.

4.3 Hüsler-Reiss Distribution

It is well known that block maxima of bivariate normal random variables with any constant covariance less than one converge in distribution to a separable bivariate max-stable model (see e.g., Geffroy, 1958; Sibuya, 1960). Hüsler & Reiss (1989) constructed a non-degenerate max-stable distribution from block maxima of normal random variables by making the covariance ρ increase with the sample size. If the covariance $\rho(n)$ satisfies $[1 - \rho(n)] \log(n) \to \lambda^2 \in [0, \infty]$ as $n \to \infty$ then the distribution of block maxima is given by

$$F_\lambda(x_1, x_2) = \exp\left\{-e^{-x_1}\Phi\left(\lambda + \frac{x_2 - x_1}{2\lambda}\right) - e^{-x_2}\Phi\left(\lambda + \frac{x_1 - x_2}{2\lambda}\right)\right\} \tag{4.5}$$

where $\Phi(\cdot)$ is the standard normal distribution function.

Hüsler & Reiss (1989) extend the result to the multivariate case. As with all multivariate max-stable models, the distribution function can be written as $\exp[-V(\mathbf{x})]$ where $V(\mathbf{x})$ is the exponent measure. In this case $V(\mathbf{x})$ is in the form of a sum of a function of lower-dimensional margins. The

sum is indexed by all possible combinations of indices from lower-dimensional margins. Let Λ be the the symmetric, conditionally negative definite, matrix which parameterizes a d-dimensional HR distribution. Let $\mathbf{m} = m_0, \ldots, m_l$ where $l \in 1, \ldots, (d-1)$ and $0 \leq m_0 < \cdots < m_l \leq d-1$ such that $\mathbf{m}$ indicates which margins to include and l indicates the number of included indices (i.e., the lower dimension). The distribution function of the d-dimensional HR distribution, on Gumbel margins, is given by

$$F_\Lambda(\mathbf{x}) = \exp\left\{ \sum_{l=0}^{d-1} \sum_{\mathbf{m}:0\leq m_0<\cdots<m_l\leq d-1} f_{l,\mathbf{m},\Lambda}(x_{m_1}, \ldots, x_{m_l}) \right\}. \tag{4.6}$$

Here $f_{l,\mathbf{m},\Lambda}$ is the l-dimensional component of the exponent measure that includes margins $\mathbf{m}$. For the full form of the multivariate HR distribution see Hüsler & Reiss (1989) or Engelke et al. (2015). Equation (4.6) makes it clear that the number of terms in the exponent measure grows combinatorially as the dimension of the problem grows. This result, and the multivariate integrals required, demonstrates that the likelihood is onerous even at reasonably small dimensions.

Engelke et al. (2015) showed that, when the data are in the maximum domain of attraction of F_Λ, the distribution of so-called extremal increments is multivariate normal. They use this normal likelihood to perform inference. Their definition of extremal increments requires conditioning on a fixed component being large which does not naturally fit our use case.

We rely on the closure property of the HR distribution; the lower dimensional margins of the HR distribution are HR distributions with dependence parameters which are equal to the respective components of Λ. This allows us to follow Padoan et al. (2010) and use a bivariate composite likelihood approach. As such, we only consider the bivariate HR distribution.

The TLETS models are defined on regularly varying $\alpha = 2$ margins (e.g., Frechet(2)) so we transform the margins of the HR distribution (4.5) to be Frechet(2). For notational convenience let $a_{ij} = a(x_i, x_j, \lambda_{ij}) = \lambda_{ij} - \log(x_i/x_j)/\lambda_{ij}$ and $a_{ji} = a(x_j, x_i, \lambda_{ij}) = \lambda_{ij} - \log(x_j/x_i)/\lambda_{ij}$. Let

$G_{\lambda_{ij}}(x_i, x_j) = H_{\lambda_{ij}}(\log(x_i^2), \log(x_j^2))$ which has Frechet(2) margins and is given by

$$\begin{aligned} G_{\lambda_{ij}}(x_i, x_j) =& \exp\left\{-x_i^{-2}\Phi\left(a_{ij}\right) - x_j^{-2}\Phi\left(a_{ji}\right)\right\} \\ =& \exp\left\{-V(x_i, x_j, \lambda_{ij})\right\}. \end{aligned} \tag{4.7}$$

Hereafter we suppress the arguments for V. The density is

$$\begin{aligned} g_{\lambda_{ij}}(x_i, x_j) =& \frac{\partial^2}{\partial x_i \partial x_j} G_{\lambda_{ij}}(x_i, x_j) \\ =& G_{\lambda_{ij}}(x_i, x_j)\left(\frac{\partial}{\partial x_i}V\frac{\partial}{\partial x_j}V - \frac{\partial^2}{\partial x_i \partial x_j}V\right) \end{aligned} \tag{4.8}$$

where the derivatives in (4.8) are

$$\frac{\partial}{\partial x_i}V = \frac{-2}{x_i^3}\Phi(a_{ij}) - \frac{1}{\lambda_{ij}x_i^3}\phi(a_{ij}) + \frac{1}{\lambda_{ij}x_i x_j^2}\phi(a_{ji}) \tag{4.9}$$

$$\frac{\partial^2}{\partial x_i \partial x_j}V = \frac{-\lambda_{ij}^2 - \log\frac{x_i}{x_j}}{\lambda_{ij}^3 x_i^3 x_j}\phi(a_{ij}) + \frac{-\lambda_{ij}^2 - \log\frac{x_j}{x_i}}{\lambda_{ij}^3 x_i x_j^3}\phi(a_{ji}). \tag{4.10}$$

4.3.1 TPD link

The basis of our method is the link between the ij^{th} parameter of the HR distribution and the ij^{th} TPD parameter (σ_{ij}) which is defined in (2.2). In order to compute the TPD from the HR model we need to first obtain the angular measure H.

Following Theorem 1 in Coles & Tawn (1991) we note that we can obtain the angular density $h^*_{\lambda_{ij}}(\cdot)$ by taking partial derivatives of the exponent measure V and transforming to pseudo-polar coordinates. As mentioned above, it is convenient for the models that we will work with to use the L_2-norm and for the margins to be Frechet(2). This means that we cannot simply apply Theorem 1 from Coles and Tawn. In their case the Jacobian of the transformation is the L_1-norm $(x_i + x_j)$ whereas ours includes an extra term. Let $r = ||(x_i, x_j)||_2 = \sqrt{x_i^2 + x_j^2}$ and $\mathbf{s} = (s_i, s_j) = (x_i, x_j)/r$ so that r is the radial component of each point and $\mathbf{s}$ is the pseudo-angular component. Theorem 1.1 in Song and Gupta (1997) show that the Jacobian is r/s_j. We have already taken the

derivatives in (4.8) which we rewrite in anticipation of the pseudo-polar change of variables:

$$\begin{aligned}
\frac{\partial^2}{\partial x_i \partial x_j} V =& \mu(d\mathbf{x}_{ij}) \\
=& \left\{ \frac{-\lambda_{ij}^2 - \log \frac{x_i}{x_j}}{\lambda_{ij}^3 x_i^3 x_j} \phi(a_{ij}) + \frac{-\lambda_{ij}^2 - \log \frac{x_j}{x_i}}{\lambda_{ij}^3 x_i x_j^3} \phi(a_{ji}) \right\} d\mathbf{x}_{ij} \\
=& \left\{ \left(x_i^2 + x_j^2\right)^{-2} \frac{-\lambda_{ij}^2 - \log \frac{x_i/\sqrt{x_i^2+x_j^2}}{x_j/\sqrt{x_i^2+x_j^2}}}{\lambda_{ij}^3 x_i^3 x_j \left(x_i^2 + x_j^2\right)^{-2}} \phi \left(\lambda_{ij} - \frac{1}{\lambda_{ij}} \log \frac{\frac{x_i}{\sqrt{x_i^2+x_j^2}}}{\frac{x_j}{\sqrt{x_i^2+x_j^2}}} \right) + \right. \\
& \left. \left(x_i^2 + x_j^2\right)^{-2} \frac{-\lambda_{ij}^2 - \log \frac{x_j/\sqrt{x_i^2+x_j^2}}{x_i/\sqrt{x_i^2+x_j^2}}}{\lambda_{ij}^3 x_i x_j^3 \left(x_i^2 + x_j^2\right)^{-2}} \phi \left(\lambda_{ij} - \frac{1}{\lambda_{ij}} \log \frac{\frac{x_j}{\sqrt{x_i^2+x_j^2}}}{\frac{x_i}{\sqrt{x_i^2+x_j^2}}} \right) \right\} d\mathbf{x}_{ij}. \qquad (4.11)
\end{aligned}$$

To transform to polar coordinates we include the Jacobian

$$\begin{aligned}
\mu(d\mathbf{x}_{ij}) =& J(\mathbf{x}_{ij} \to r, \mathbf{s}) \mu(dr \times d\mathbf{s}) \\
=& \frac{r}{s_j} r^{-4} \left\{ \frac{-\lambda_{ij}^2 - \log \frac{s_i}{s_j}}{\lambda_{ij}^3 s_i^3 s_j} \phi \left(\lambda_{ij} - \frac{1}{\lambda_{ij}} \log \frac{s_i}{s_j} \right) + \right. \\
& \left. \frac{-\lambda_{ij}^2 - \log \frac{s_j}{s_i}}{\lambda_{ij}^3 s_i s_j^3} \phi \left(\lambda_{ij} - \frac{1}{\lambda_{ij}} \log \frac{s_j}{s_i} \right) \right\} dr d\mathbf{s} \\
=& -2 r^{-3} dr \; \frac{1}{2 s_j} \left\{ \frac{\lambda_{ij}^2 + \log \frac{s_i}{s_j}}{\lambda_{ij}^3 s_i^3 s_j} \phi \left(\lambda_{ij} - \frac{1}{\lambda_{ij}} \log \frac{s_i}{s_j} \right) + \right. \\
& \left. \frac{\lambda_{ij}^2 + \log \frac{s_j}{s_i}}{\lambda_{ij}^3 s_i s_j^3} \phi \left(\lambda_{ij} - \frac{1}{\lambda_{ij}} \log \frac{s_j}{s_i} \right) \right\} d\mathbf{s} \\
=& -2 r^{-3} dr \; h_{\lambda_{ij}}(\mathbf{s}) d\mathbf{s} \qquad (4.12)
\end{aligned}$$

Here we note that $h_{\lambda_{ij}}(\mathbf{s})$ is a density on the one dimensional L_2-ball in the positive quadrant of $\mathbb{R}^2$ and thus it is equivalent to

$$h_{\lambda_{ij}}(s_i) = \frac{1}{2\sqrt{1-s_i^2}}\left\{ \frac{\lambda_{ij}^2 + \log\frac{s_i}{\sqrt{1-s_i^2}}}{\lambda_{ij}^3 s_i^3\sqrt{1-s_i^2}}\phi\left(\lambda_{ij} - \frac{1}{\lambda_{ij}}\log\frac{s_i}{\sqrt{1-s_i^2}}\right) + \right.$$
$$\left. \frac{\lambda_{ij}^2 + \log\frac{\sqrt{1-s_i^2}}{s_i}}{\lambda_{ij}^3 s_i\sqrt{1-s_i^2}^3}\phi\left(\lambda_{ij} - \frac{1}{\lambda_{ij}}\log\frac{\sqrt{1-s_i^2}}{s_i}\right)\right\}. \tag{4.13}$$

The ij^{th} parameter of the HR distribution has ij^{th} TPD parameter

$$\begin{aligned}\sigma_{ij} &= \int_0^1 s_i\sqrt{1-s_i^2}h_{\lambda_{ij}}(s_i)ds_i \\ &= \frac{1}{2}\int_0^1 s_i\left\{ \frac{\lambda_{ij}^2 + \log\frac{s_i}{\sqrt{1-s_i^2}}}{\lambda_{ij}^3 s_i^3\sqrt{1-s_i^2}}\phi\left(\lambda_{ij} - \frac{1}{\lambda_{ij}}\log\frac{s_i}{\sqrt{1-s_i^2}}\right) + \right. \\ &\quad \left. \frac{\lambda_{ij}^2 + \log\frac{\sqrt{1-s_i^2}}{s_i}}{\lambda_{ij}^3 s_i\sqrt{1-s_i^2}^3}\phi\left(\lambda_{ij} - \frac{1}{\lambda_{ij}}\log\frac{\sqrt{1-s_i^2}}{s_i}\right)\right\}ds_i. \end{aligned} \tag{4.14}$$

The integral in (4.14) is unknown and thus evaluating the map between the HR dependence parameter and the TPD requires numerical approximation. Let $\lambda(\cdot)$ be the inverse function which takes a TPD value and returns the bivariate HR dependence parameter (called λ_{ij} up until this point) and $\Lambda(\cdot)$ matrix version of $\lambda(\cdot)$. Figure 4.3 plots the function $\lambda(\sigma)$ which demonstrates the bijection between the TPD and the dependence parameter of a bivariate HR distribution. This plot was created through numerical integration and linear interpolation. Figure 4.4 shows an example HR point cloud under weak and strong dependence respectively. Each point cloud has 10000 points. Combining the TPD function from one of the TLETS models (4.2), (2.4), or (4.4) with $\lambda(\cdot)$ gives a map between the TLETS parameters and the HR parameter that has the equivalent TPD. Let $\boldsymbol{\theta}$ be the parameter vector for the TLETS model that is being fit (e.g., a $TL-ARMA(1,1)$ model has $\boldsymbol{\theta} = (\phi, \theta)$). This map is the link between the two models and results in a bivariate HR distribution (like 4.8) which is a function of the TLETS parameters:

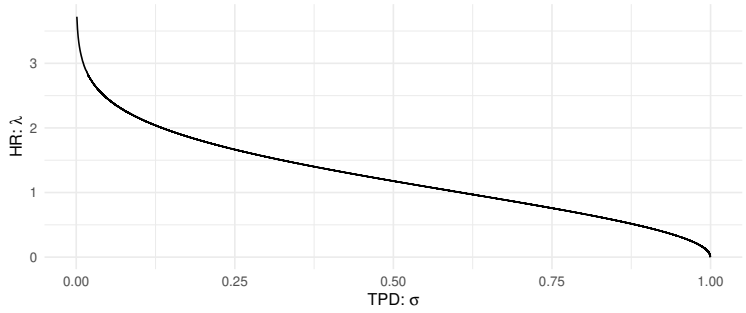
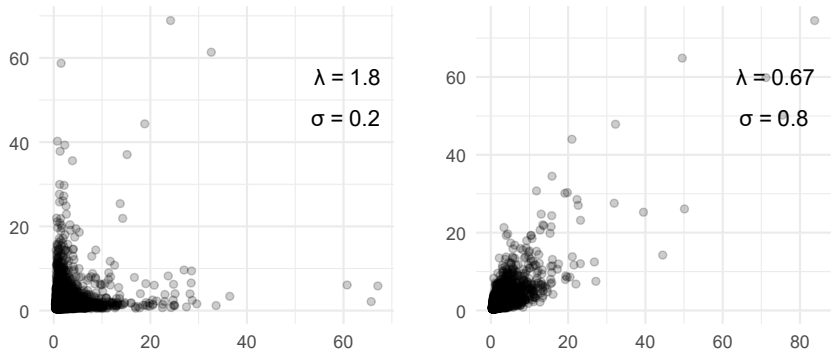


Figure 4.3: Link between Hüsler Reiss dependence parameter (λ) and TPD dependence parameter (σ).



(a) Weak Dependence **(b)** Strong Dependence

Figure 4.4: Hüsler-Reiss point clouds (10000 points) with dependence parameter (λ) and TPD parameter (σ).

$$g_{\lambda\{\sigma(h,\boldsymbol{\theta})\}}(x_n, x_{n+h}) = G_{\lambda\{\sigma(h,\boldsymbol{\theta})\}}(x_n, x_{n+h})\left(\frac{\partial}{\partial x_n}V\frac{\partial}{\partial x_{n+h}}V - \frac{\partial^2}{\partial x_n \partial x_{n+h}}V\right) \tag{4.15}$$

$$\text{where } V = V\left\{x_n, x_{n+h}, \lambda[\sigma(h,\boldsymbol{\theta})]\right\}.$$

4.4 Composite Likelihood

Our method combines the pieces shown above using a composite likelihood approach. Composite likelihood approaches go as far back as Lindsay (1988) and are a useful alternative to full likelihood inference when the full likelihood is analytically unavailable or computationally infeasible (see, e.g., Varin et al., 2011). The idea is to use a combination of valid likelihoods (e.g.,

bivariate or conditional likelihoods) in the place of the full likelihood. We follow Padoan et al. (2010) and use the bivariate margins as our likelihood components.

We adapt the composite likelihood to the time series context in the following manner. Let $\boldsymbol{\theta}$ be the set of parameters for the full likelihood. Let $\mathbf{x_i} = (x_{i+1}, x_{i+2}, \ldots, x_N)^T$ where N is the length of the observed time series. Let L_h be the likelihood for the lag-h margin and w_h be user specified weights. The bivariate composite likelihood for time series is

$$L_{cl}(\boldsymbol{\theta}|\mathbf{x}) = \prod_{h=1}^{\infty} \prod_{n=1}^{N-h} [L_h(\boldsymbol{\theta}|x_n, x_{n+h})]^{w_h}. \tag{4.16}$$

Here the first product is over lags (i.e., the bivariate margins we are considering) and the second product is over the $N - h$ observations of lag-h points. Time series are infinite dimensional and thus the first product has infinitely many terms. In practice we consider some large maximum lag value h_{max} which is equivalent to so-called tapered weights which set $w_h := 0, \forall h > h_{max}$. A user could consider other weights such as weights which are inversely proportional to the lag (e.g., h^{-1}) but we have not found that useful in simulations and thus we consider unit weights for all $h < h_{max}$ in the following. When using the method for model selection h_{max} must be the same in all model fits.

This composite likelihood approach is necessary as the analytic form of the h_{max}-dimensional Hüsler-Reiss likelihood becomes unwieldy for even moderate dimensions. While necessary to proceed in this manner, we think that this is a sensible approach because the parameters that we are interested in are inherently bivariate and the composite likelihood enjoys many of the same benefits of a full likelihood approach.

Replacing L_h with the bivariate HR density (4.15) the composite log-likelihood is

$$\begin{aligned} \ell_{cl}(\boldsymbol{\theta}|\mathbf{x}) &= \sum_{h=1}^{h_{max}} \sum_{n=1}^{N-h} w_h \log g_{\lambda\{\sigma(h,\boldsymbol{\theta})\}}(x_n, x_{n+h}) \\ &= \sum_{h=1}^{h_{max}} w_h \sum_{n=1}^{N-h} \left\{ -V + \log\left(\frac{\partial}{\partial x_n} V \frac{\partial}{\partial x_{n+h}} V - \frac{\partial^2}{\partial x_n \partial x_{n+h}} V \right) \right\} \end{aligned} \tag{4.17}$$

where the derivatives are defined in (4.9) and (4.10). Equation (4.17) is what we term the bivariate HR composite proxy-likelihood (proxy-likelihood for short). Many likelihood methods are able to determine an analytic form for the likelihood estimator by solving the system of equations generated by setting the score function equal to zero. This is not available to us as we do not have closed forms for some components of the score function which will be discussed further in the next section.

4.5 Inference and Model Selection

We acknowledge that our composite likelihood approach reuses data as it treats each bivariate margin as if it were independent of all other margins. This results in a surface which, likely, has more curvature than the true likelihood surface suggesting more confidence in parameter estimates than indicated by the data. We correct for this with the sandwich variance estimator (Godambe, 1960) which is used in an estimating equations approach and in the composite likelihood approach (see, e.g., Varin, 2008). We briefly review the estimating equations approach here in an attempt to highlight the assumptions that we are making in this section. For a more thorough discussion see, e.g., Boos & Stefanski, 2013. The estimating equation approach says that if one can come up with a function $\psi(\mathbf{x}_i, \boldsymbol{\theta})$ for a single point $\mathbf{x}_i$ such that the unique solution to $E[\psi(\mathbf{x}_i, \boldsymbol{\theta})] = \mathbf{0}$ is the parameter of the distribution $\boldsymbol{\theta}_0$, then we can use ψ to develop an estimator $\hat{\boldsymbol{\theta}}$ which is a function of the mean of ψ evaluated over several points such that, by the weak law of large numbers, $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. Estimators developed this way are called M-estimators. Furthermore, standard Taylor expansion arguments for $\frac{1}{n}\sum_{i=1}^{n} \psi(\mathbf{x}_i, \boldsymbol{\theta})$ around $\boldsymbol{\theta}_0$ demonstrate the asymptotic normality of $\sqrt{(}n)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ with variance $\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\mathbf{J}(\boldsymbol{\theta}_0)\{\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\}^T$ where $\mathbf{H}(\boldsymbol{\theta_0}) = E[-\psi'(\mathbf{x}_i, \boldsymbol{\theta}_0)]$ and $\mathbf{J}(\boldsymbol{\theta}_0) = E[\psi(\mathbf{x}_i, \boldsymbol{\theta}_0)\psi(\mathbf{x}_i, \boldsymbol{\theta}_0)^T]$.

We naturally consider the derivative of (4.17) (the score function) as ψ and thus $\mathbf{H}(\boldsymbol{\theta})$ is the Hessian and $\mathbf{J}(\boldsymbol{\theta})$ is the covariance of the score function. Our context has a few challenges which must be acknowledged.

First, we do not know the data-generating likelihood; our method is designed to be useful as it captures the second order tail dependence without knowledge of the likelihood. For this reason we cannot show that the unique solution to the expected value of our ψ is $\boldsymbol{\theta}_0$. As such we do not claim to have developed an M-estimator but we think that it is illuminating to think of our method in that framework and that the correction to the variance estimator may be useful for model selection as is done by Varin & Vidoni (2005).

Second, (4.17) includes the entire observed time series and thus we do not have replications. Without replications we cannot average to get the WLLN result. The consistency of the estimator relies on a consistent estimator for

$$\mathbf{J}(\hat{\boldsymbol{\theta}}_{MPL}) = \text{Var}\big[S(\hat{\boldsymbol{\theta}}_{MPL}, \mathbf{X})\big] = \text{Var}\left[\sum_{h=1}^{h_{max}} \sum_{n=1}^{N-h} S_h(\hat{\boldsymbol{\theta}}_{MPL}, X_n, X_{n+h})\right]$$

but this is challenging due to the lack of replicates. We cannot pull a sum out of the variance because neither sum is over independent random variables. We follow Heagerty & Lumley (2000) who suggest using the entire time series in the estimator $\hat{\boldsymbol{\theta}}$ but then using subsets of the domain $1, \ldots, N$ to compute the score. This allows us to estimate $\mathbf{J}(\boldsymbol{\theta})$ as the covariance of the estimated scores computed over each sub-domain. Let $m = 1, \ldots, M$ index the subseries of length n_m which we treat as independent replicates. With these quasi-replicates we consider

$$\mathbf{J}(\hat{\boldsymbol{\theta}}_{MPL}) \approx \sum_{m=1}^{M} \text{Var}\left[\sum_{h=1}^{h_{max}} \sum_{n=1}^{n_m-h} S_h(\hat{\boldsymbol{\theta}}_{MPL}, X_{m,n}, X_{m,n+h})\right]$$

which allows us to use the natural estimator (the empirical variance over the subseries).

Finally, asymptotic normality of the estimator has to rely on a CLT for α-mixing processes as our process is clearly not iid but the sub-series should satisfy α-mixing requirements. This is of least importance because any claims of approximate normality are subject to the prior caveats and thus we do not claim to have shown normality.

In practice we split our data into a small number of subseries ($M \in 10, \cdots, 20$) as we need to have tail dependence information in each. Let $\hat{\boldsymbol{\theta}}_{MPL} = (\hat{\theta}_{1,MPL}, \ldots, \hat{\theta}_{K,MPL})$ be the maximum

proxy-likelihood (MPL) estimator for $\boldsymbol{\theta}$. The composite likelihood versions of the AIC (Varin & Vidoni, 2005) and BIC (Gao & Song, 2010) are the natural tool for comparison across models:

$$\text{CLAIC} = -2\ell_{cl}(\hat{\boldsymbol{\theta}}_{MPL}|\mathbf{x}) + 2\text{tr}\{\mathbf{J}(\hat{\boldsymbol{\theta}}_{MPL})\mathbf{H}^{-1}(\hat{\boldsymbol{\theta}}_{MPL})\} \tag{4.18}$$

$$\text{CLBIC} = -2\ell_{cl}(\hat{\boldsymbol{\theta}}_{MPL}|\mathbf{x}) + \log(n)\text{tr}\{\mathbf{J}(\hat{\boldsymbol{\theta}}_{MPL})\mathbf{H}^{-1}(\hat{\boldsymbol{\theta}}_{MPL})\}. \tag{4.19}$$

While having straightforward uncertainty estimation through confidence intervals is always desirable, and is a common motivator for likelihood methods, we do not prioritize uncertainty quantification. One could use the estimated matrices $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}}_{MPL})$ and $\hat{\mathbf{H}}(\hat{\boldsymbol{\theta}}_{MPL})$ in a sandwich variance estimator. Standard arguments could then be used to develop confidence intervals for the parameters in the models that we fit. Initial investigation suggests that coverage of intervals made using these standard arguments is poor. This is not a big concern for us because our primary concern is the second-order dependence property (the TPD) not the parameters. If uncertainty quantification is desired we suggest using bootstrapping.

The composite likelihood model selection techniques require us to estimate the Hessian ($\mathbf{H}$) and the variance of the score ($\mathbf{J}$). We derive the score equations below which allows for natural covariance estimation. We could write down an estimator of the Hessian through differentiating the score function but it is easily obtained from many standardized computer optimization routines which is what we use in practice.

4.5.1 Score function

The contribution of one point to the score function for some $\theta_k \in \boldsymbol{\theta}$ is, by the chain rule, of the form

$$\frac{\partial}{\partial\theta_k}\ell_{cl}[\lambda\{\sigma(\boldsymbol{\theta})\}|x_n, x_{n+h}] = \frac{\partial}{\partial\lambda_h}\ell_{cl}(\lambda_h|x_n, x_{n+h})\frac{\partial}{\partial\sigma}\lambda(\sigma)\frac{\partial}{\partial\theta_k}\sigma(h, \boldsymbol{\theta}). \tag{4.20}$$

Obtaining the first and third terms is done below but the middle term, $\frac{\partial}{\partial\sigma}\lambda(\sigma)$, is obtained numerically as we do not have an analytic form of the inverse of (4.14). Notice that lag-h and lag-h' points will both contribute to the score for θ_k if the third term, $\frac{\partial}{\partial\theta_k}\sigma(\cdot,\boldsymbol{\theta})$, is non-zero for h and h'.

The first term in the contribution of a single point to (4.20) is the only component that includes the data. It is in this component that the likelihood surface aligns the data with the HR distribution. One could imagine simply finding the critical points of the HR likelihood surface, computing the resulting TPD values, and finding the TLETS parameters that provide the closest match to those TPD values. While attractive this process does not satisfy the goals that we set out to achieve (namely natural uncertainty quantification and model selection). The parameter space that we are interested in is the space of valid TLETS parameters which is a subspace of $\mathbb{R}_+^{h_{max}}$, the space that we would be optimizing over to find the best HR parameters. The set of TPD parameters that we would find would not map directly to a set of TLETS parameters and thus we would almost certainly not be at a critical point of the TLETS parameterized surface. Thus even though this is the only component that directly involves the data, the solution is a weighted sum of these components.

To derive the analytic form of the first component (assuming unit weights) let $a_h = a(x_n, x_{n+h}, \lambda_h) = \lambda_h - \log(x_n/x_{n+h})/\lambda_h$ and $a_{h'} = a(x_{n+h}, x_n, \lambda_h)$ is defined by symmetry so that a_h and $a_{h'}$ are analogous to a_{ij} and a_{ji} used above. The first component of the score for one lag-h point is

$$\frac{\partial}{\partial\lambda_h}\ell_{cl}(\lambda_h|x_n, x_{n+h}) = \\ -x_n^{-2}\frac{\partial}{\partial\lambda_h}\Phi(a_h) - x_{n+h}^{-2}\frac{\partial}{\partial\lambda_h}\Phi(a_{h'}) + \frac{\frac{\partial}{\partial\lambda_h}\left(\frac{\partial}{\partial x_n}V\frac{\partial}{\partial x_{n+h}}V - \frac{\partial^2}{\partial x_n\partial x_{n+h}}V\right)}{\frac{\partial}{\partial x_n}V\frac{\partial}{\partial x_{n+h}}V - \frac{\partial^2}{\partial x_n\partial x_{n+h}}V}. \quad (4.21)$$

The derivatives in (4.21) are

$$\frac{\partial}{\partial \lambda_h}\Phi(a_h) = \left(1 + \frac{1}{\lambda_h^2}\log\frac{x_n}{x_{n+h}}\right)\phi(a_h) \tag{4.22}$$

$$\begin{aligned}\frac{\partial}{\partial \lambda_h}\left(\frac{\partial}{\partial x_n}V\frac{\partial}{\partial x_{n+h}}V\right) =& \frac{\partial}{\partial x_{n+h}}V\left(\frac{\partial}{\partial \lambda_h}\frac{\partial}{\partial x_n}V\right) + \frac{\partial}{\partial x_n}V\left(\frac{\partial}{\partial \lambda_h}\frac{\partial}{\partial x_{n+h}}V\right)\\ \frac{\partial}{\partial \lambda_h}\frac{\partial}{\partial x_n}V =& \frac{-2}{x_n^3}\left(1 + \frac{1}{\lambda_h^2}\log\frac{x_n}{x_{n+h}}\right)\phi(a_h) +\\ & \frac{\lambda_h^2 + 1 - \frac{1}{\lambda_h^4}\log^2\frac{x_n}{x_{n+h}}}{x_n^3}\phi(a_h) -\\ & \frac{\lambda_h^2 + 1 - \frac{1}{\lambda_h^4}\log^2\frac{x_{n+h}}{x_n}}{x_n x_{n+h}^2}\phi(a_{h'})\end{aligned} \tag{4.23}$$

$$\begin{aligned}\frac{\partial}{\partial \lambda_h}\frac{\partial^2}{\partial x_n \partial x_{n+h}}V =& \frac{1 + \frac{1}{\lambda_h^2} + \left(\frac{1}{\lambda_h^2} + \frac{3}{\lambda_h^4}\right)\log\frac{x_n}{x_{n+h}} - \frac{1}{\lambda_h^4}\log^2\frac{x_n}{x_{n+h}} - \frac{1}{\lambda_h^6}\log^3\frac{x_n}{x_{n+h}}}{x_n^3 x_{n+h}}\phi(a_h) +\\ & \frac{1 + \frac{1}{\lambda_h^2} + \left(\frac{1}{\lambda_h^2} + \frac{3}{\lambda_h^4}\right)\log\frac{x_{n+h}}{x_n} - \frac{1}{\lambda_h^4}\log^2\frac{x_{n+h}}{x_n} - \frac{1}{\lambda_h^6}\log^3\frac{x_{n+h}}{x_n}}{x_n^3 x_{n+h}}\phi(a_{h'})\end{aligned} \tag{4.24}$$

The third term in contribution of a single point to (4.20) is model dependent. We give the forms for each of our motivating models here. The $TL - AR(1)$ has one parameter typically denoted ϕ and has TPD function (4.2) which has derivative

$$\frac{\partial}{\partial \theta_1}\sigma(h, \theta_1) = \frac{\partial}{\partial \phi}\phi^h = h\phi^{h-1}. \tag{4.25}$$

The $TL - MA(q)$ has q parameters $(\theta_1, \ldots, \theta_q)$ and TPD given by (2.4) which has derivative

$$\frac{\partial}{\partial \theta_k}\sigma(h, \theta_1, \ldots, \theta_q) = \frac{\theta_{k+h}^{(0)} + \theta_{k-h}^{(0)}}{\sum_{l=0}^q \theta_l^2} + \frac{2\theta_k * \sum_{l=0}^q \theta_l^{(0)}\theta_{l+h}^{(0)}}{\left(\sum_{l=0}^q \theta_l^2\right)^2} \tag{4.26}$$

where one or both of $\theta_{k+h}^{(0)}$ and $\theta_{k-h}^{(0)}$ may be zero. The $TL - ARMA(1,1)$ model has two parameters typically denoted ϕ for the AR component and θ for the MA component. The TPD is given by (4.4)

which has derivative with respect to ϕ

$$\frac{\partial}{\partial\phi}\sigma(h,\theta,\phi)=\begin{cases}\frac{\phi^{h-2}(h\phi+h\theta-\theta)(1+\phi\theta)+\phi^{h-1}\theta(\phi+\theta)}{1+2\theta\phi+\theta^2}- & \\ \quad\frac{2\theta(\phi+\theta)\phi^{h-1}(1+\phi\theta)}{(1+2\theta\phi+\theta^2)^2} & \text{if } \phi>0,\phi+\theta>0\\ 0 & \text{if } \phi>0,\phi+\theta<0\\ \frac{2(\phi+\theta)\phi^h+(\phi+\theta)^2h\phi^{h-1}}{1-\phi^4+(\phi+\theta)^2}- & \\ \quad\frac{(\phi+\theta)^2\phi^h(4\phi^3+2\phi+2\theta)}{\{1-\phi^4+(\phi+\theta)^2\}^2} & \text{if } \phi<0,\phi+\theta>0,h \text{ is even}\\ \frac{\phi^{h-2}\{(h\phi+h\theta-\theta)(1-\phi^4)+4\phi^5+4\theta\phi^4\}}{1-\phi^4+(\phi+\theta)^2}- & \\ \quad\frac{2\phi^{h-1}(\phi+\theta)(1-\phi^4)(2\phi^3+\phi+\theta)}{\{1-\phi^4+(\phi+\theta)^2\}^2} & \text{if } \phi<0,\phi+\theta>0,h \text{ is odd}\\ \frac{\phi^{h-2}\{(h\phi+h\theta-\theta)(1+\theta\phi^3)+3\theta\phi^3(\phi+\theta)\}}{1+\theta^2\phi^2+2\theta\phi^3}- & \\ \quad\frac{2\theta\phi^h(\phi+\theta)(1+\theta\phi^3)(\theta+3\phi)}{(1+\theta^2\phi^2+2\theta\phi^3)^2} & \text{if } \phi<0,\phi+\theta<0,h \text{ is even}\\ 0 & \text{if } \phi<0,\phi+\theta<0,h \text{ is odd.}\end{cases}$$

The derivative of (4.4) with respect to θ is

$$\frac{\partial}{\partial\theta}\sigma(h,\theta,\phi)=\begin{cases}\frac{\theta\phi^{h-1}(1+\phi\theta)(1+\phi)}{1+2\theta\phi+\theta^2}-\frac{2(\phi+\theta)^2\phi^{h-1}(1+\phi\theta)}{(1+2\theta\phi+\theta^2)^2} & \text{if } \phi>0,\phi+\theta>0\\ 0 & \text{if } \phi>0,\phi+\theta<0\\ \frac{2(\phi+\theta)\phi^h}{1-\phi^4+(\phi+\theta)^2}-\frac{2(\phi+\theta)^3\phi^h}{(1-\phi^4+(\phi+\theta)^2)^2} & \text{if } \phi<0,\phi+\theta>0,h \text{ is even}\\ \frac{\phi^{h-1}(1-\phi^4)}{1-\phi^4+(\phi+\theta)^2}-\frac{2(\phi+\theta)^2\phi^{h-1}(1-\phi^4)}{(1-\phi^4+(\phi+\theta)^2)^2} & \text{if } \phi<0,\phi+\theta>0,h \text{ is odd}\\ \frac{\phi^{h-1}(1+2\theta\phi^3+\phi^4)(1+\theta^2\phi^2+2\theta\phi^3)}{1+\theta^2\phi^2+2\theta\phi^3}- & \\ \quad\frac{2(\theta\phi^2+\phi^3)(\phi+\theta)\phi^{h-1}(1+\theta\phi^3)}{(1+\theta^2\phi^2+2\theta\phi^3)^2} & \text{if } \phi<0,\phi+\theta<0,h \text{ is even}\\ 0 & \text{if } \phi<0,\phi+\theta<0,h \text{ is odd.}\end{cases}$$

4.6 Censoring approach to parameter estimation

Key to any analysis of extreme events is the idea that we need data which are extreme to inform about the tails of the process. In other words, we do not want data from the bulk of the distribution to bias the inference in the tails. As in any statistical analysis, we seek to retain as much information as possible while ignoring the information in the bulk of the distribution that would cause bias. Two common approaches are Euclidean censoring (see, e.g., Smith et al., 1997 and Huser et al., 2016) and only including data which are large where "large" is based on the magnitude of the radial component. This second approach is commonly done in the estimation of the TPD (see, e.g., Cooley & Thibaud, 2019). We briefly explore these two censoring schema as our method can be employed with either.

Huser et al. (2016) show that using Euclidean threshold censoring in pairwise likelihood estimation has lower bias and RMSE than other proposed likelihood estimators. The central idea behind censoring is to transform the joint distribution of the data into a joint distribution of threshold exceedance indicators and threshold exceedances. Consider the data vector $(X_1, \cdots, X_n)$ which, after transformation, is marginally Frechet(2). Let u_q be the q-quantile of a Frechet(2) distribution which will be our censoring threshold. Following Smith et al. (1997) we let $\delta_i = \mathbb{I}(X_i > u_q)$ and $Y_i = \max(0, X_i - u_q)$. We then determine the joint distribution of $(\delta_i, Y_i, \delta_j, Y_j)$ from the bivariate HR density on Frechet(2) margins and denote this censored density $g^c_{\lambda_{ij}}(\cdot)$:

$$
\begin{aligned}
&g^c_{\lambda_{ij}}(\delta_i, y_i, \delta_j, y_j) \\
&= \begin{cases}
G_{\lambda_{ij}}(u_q, u_q) & \delta_i = \delta_j = 0 \\
\frac{\partial}{\partial x_i} G_{\lambda_{ij}}(u_q + y_i, u_q) & \delta_i = 1, \delta_j = 0 \\
\frac{\partial}{\partial x_j} G_{\lambda_{ij}}(u_q, u_q + y_j) & \delta_i = 0, \delta_j = 1 \\
\frac{\partial^2}{\partial x_i \partial x_j} G_{\lambda_{ij}}(u_q + y_i, u_q + y_j) & \delta_i = \delta_j = 1.
\end{cases} \qquad (4.27) \\
&= \begin{cases}
G_{\lambda_{ij}}(u_q, u_q) & \delta_i = \delta_j = 0 \\
\left[\frac{2}{(u_q + y_i)^3} \Phi(a_{ij}) + \frac{1}{\lambda_{ij}(u_q + y_i)^3} \phi(a_{ij}) \right. & \\
\qquad \left. - \frac{1}{\lambda_{ij}(u_q + y_i) u_q^2} \phi(a_{ji}) \right] G_{\lambda_{ij}}(u_q + y_i, u_q) & \delta_i = 1, \delta_j = 0 \\
\left[\frac{2}{(u_q + y_j)^3} \Phi(a_{ji}) + \frac{1}{\lambda_{ij}(u_q + y_j)^3} \phi(a_{ji}) \right. & \\
\qquad \left. - \frac{1}{\lambda_{ij}(u_q + y_j) u_q^2} \phi(a_{ij}) \right] G_{\lambda_{ij}}(u_q, u_q + y_j) & \delta_i = 0, \delta_j = 1 \\
g_{\lambda_{ij}}(u_q + y_i, u_q + y_j) & \delta_i = \delta_j = 1.
\end{cases} \qquad (4.28)
\end{aligned}
$$

Where the derivatives in (4.27) are in (4.9) and $\partial^2 G_{\lambda_{ij}} / \partial x_i \partial x_j(\cdot) = g_{\lambda_{ij}}(\cdot)$ (from (4.8)).

This censored likelihood requires a new first component of the score function (4.20). While this censoring approach has four cases (both dimensions are small, the i^{th} component is small, the j^{th} component is small, or both dimensions are large) we only have two new cases to derive. The two one-small cases are the same by symmetry and the both large case is the same as (4.21). The case when both dimensions are small is

$$
\frac{\partial}{\partial \lambda} \ell^c_{cl}(\lambda | \delta_i = \delta_j = 0) = -\frac{\partial}{\partial \lambda} V(u_q, u_q) = -2 u_q^{-2} \phi(\lambda). \qquad (4.29)
$$

When one component is small (WLOG we assume $\delta_i = 1$) the score contribution is

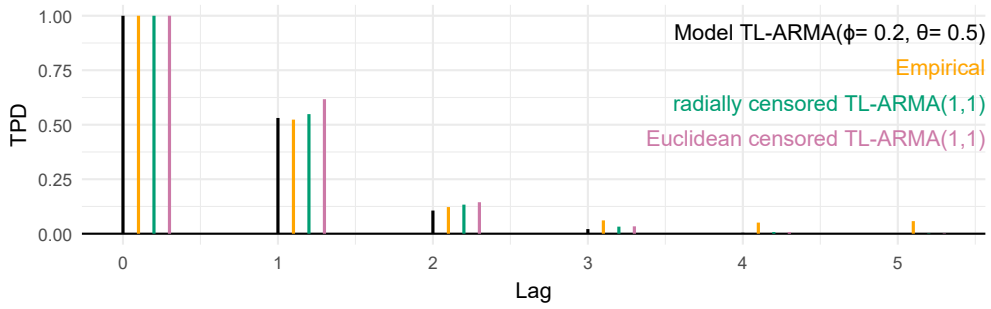
$$
\begin{aligned}
&\frac{\partial}{\partial\lambda}\ell^c_{cl}(\lambda|\delta_i=1,\delta_j=0)\\
&=\frac{\partial}{\partial\lambda}\log\left[\frac{2}{(u_q+y_i)^3}\Phi(a_{ij})+\frac{1}{\lambda_{ij}(u_q+y_i)^3}\phi(a_{ij})-\frac{1}{\lambda_{ij}(u_q+y_i)u_q^2}\phi(a_{ji})\right]\\
&=\left[\frac{2}{(u_q+y_i)^3}\Phi(a_{ij})+\frac{1}{\lambda_{ij}(u_q+y_i)^3}\phi(a_{ij})-\frac{1}{\lambda_{ij}(u_q+y_i)u_q^2}\phi(a_{ji})\right]^{-1}\\
&\quad*\left\{[1-\lambda^{-2}+(2\lambda^{-2}+\lambda^{-4}\log(x_i/u_q))\log(x_i/u_q)]x_i^{-3}\phi(a_{ij})\right.\\
&\qquad\left.+[\lambda^{-1}+\lambda-\lambda^{-3}\log^2(u_q/x_i)]x_i^{-1}u_q^{-2}\phi(a_{ji})]\right\} \qquad (4.30)
\end{aligned}
$$

Radial censoring is employed in Jiang et al. (2020); Wixson & Cooley (2023) and others. This method ignores all points which are not large when estimating the tail dependence which relies on the independence between the radial and angular components in regularly varying random vectors on polar coordinates. Consider $(X_i, X_j) \in RV_\alpha$ and transform to pseudo-polar coordinates as was done in the derivation of the TPD from an HR distribution: let $R = ||(X_i, X_j)||_2 = \sqrt{X_i^2 + X_j^2}$ and $\mathbf{S} = (S_i, S_j) = (X_i, X_j)/R$ so that R is the radial component of each point and $\mathbf{S}$ is the pseudo-angular component. Radial censoring involves fixing some large quantile of the radial distribution r_0 and including points in the analysis only if the radial component is larger than r_0.

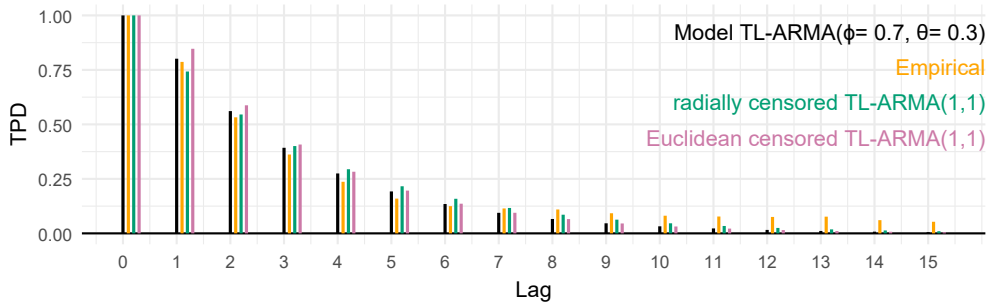
While both approaches have their strengths we found that the simpler radial approach which closely matches the estimation of second order dependence (our target) often performs better and is less computationally expensive. Figure (4.5) shows two examples of time series of length 10000 which were fit using the two different censoring schemes. In general the radially censored model fits the early lags better (because the TPD value is smaller) and the Euclidean censored model decays to zero faster.

4.7 Simulations

Our method targets a second-order summary of the joint distribution and uses a model with a continuous angular measure as a proxy for the discrete angular measure of the models that we



(a) Weak Dependence



(b) Strong Dependence

Figure 4.5: Comparison of radial and Euclidean censoring techniques.

wish to fit. We readily admit that the models that we want to fit are overly simple and thus there will be some discrepancies between real data and the fitted models. An interesting aspect of our method is that we believe that most data will look more like the proxy model than the models that we are fitting. This angular measure mismatch makes simulation studies challenging because the method was developed to target the TPD not the ability to distinguish between similar models (e.g., a $TL-AR(1)$ and a $TL-ARMA(1,1)$ with small θ). We want a method to perform model selection because we want to know which model best recreates the observed TPD, not because we think we can find the true model. In all simulations below we gernerate time series of length 10000, use the 0.95 quantile of the radial components as the threshold, and use 20 subseries to estimate the covariance of the score.

Figure (4.6) shows how similar two fitted models can be and demonstrates that both models fit the data quite well. In this case both have two parameters though the causal representation of the $TL-ARMA$ model has infinitely many MA coefficients. The $TL-ARMA$ model fit best

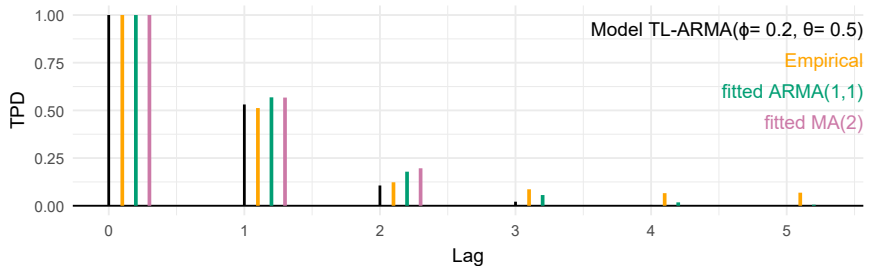


Figure 4.6: Model TPD in black, empirical TPD from a time series of length 10000 generated from a $TL - ARMA(\phi = 0.2, \theta = 0.5)$ in yellow, and TPD from fitted models. The fitted models are an $TL - ARMA(1,1)$ in green, and $TL - MA(2)$ in pink.

using both CLAIC (151580.1 compared to 151582.1) and CLBIC (by 6.25 units). Repeating this experiment we find that the CLAIC selects the $TL - ARMA$ over the $TL - MA(2)$ model in 51 out of 100 simulated time series (CLBIC selects the $TL - ARMA$ 29 times). Our method will not be able to reliably determine that these data came from a $TL - ARMA$ model but will be able find the parameters from each model that maximize the proxy-likelihood and choose the best fitting model.

The $TL - ARMA(\phi = 0.6, \theta = 0.9)$ model has stronger dependence than the previous simulation; $\sigma(5) = 0.104$ and $\sigma(10) = 0.008$ whereas our previous simulation used a model where $\sigma(5) = 0.008$. Figure (4.7) highlights the longer range of dependence in this series and the similarity between this model and competitors. In this simulation study we fit $TL-MA$ models of orders $q = 1, \ldots, 20$, the $TL - AR(1)$ model, and the $TL - ARMA(1,1)$ model. The CLAIC selected the $TL - ARMA(1,1)$ model in 82 of the 100 simulations. The generating model is chosen 94 times with CLAIC if we allow for a more parsimonious model to be selected when it is withing two CLAIC units of the best score. The CLBIC selected the generating model 94 times. The estimated parameters from the fitted true model are biased as θ is underestimated and ϕ is overestimated.

In this simulation (and the following simulation) we consider a third method of penalizing the objective function. This third method is motivated by the poor performance of the CLAIC penalty in the applications to the wildfire data that are shown below. In those applications (Tables 4.2 and 4.3) the CLAIC penalty suggests unreasonable models so we consider the penalty applied in the basic form of the AIC. This approach penalizes our objective function by two units for

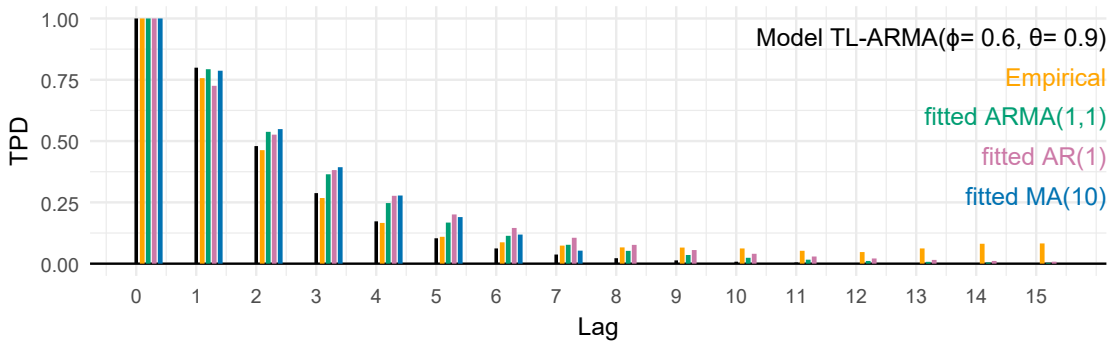


Figure 4.7: Model TPD in black, empirical TPD from a time series of length 10000 generated from a $TL-ARMA(\phi=0.6,\theta=0.9)$ in yellow, and TPD from fitted models. The fitted models are an $TL-ARMA(1,1)$ in green, $TL-AR(1)$ in pink, and a $TL-MA(10)$ in blue.

each parameter in the model. Akaike (1974) gives theoretical justification for this penalty but this justification requires the correct likelihood. With this penalty the method selects the generating model 62 times (75 times with the subjective relaxing of the criterion by two units).

Our final simulation generates data from a $TL-MA(15)$ with the coefficients set to the fitted values obtained from the innovations fit to present climate ERA5-derived FWI data in Colorado. This is the fitted model from Section 2.4. We selected this model as it is an example of a believable TPD from environmental data which does not decay as smoothly as transformed-linear models with some AR component. We fit the same models as in the previous set of simulations. Figure 4.8 displays the model, empirical, and fitted TPDs from one simulation. We highlight the broad agreement between the $TL-ARMA$ and $TL-MA$ models. In this set of 100 simulations the generating model was chosen a plurality of times with the CLAIC (32 times) and CLBIC (32 times). The generating model is selected in 92 of the simulations with the basic AIC penalty. Table 4.1 shows the number of times that each fitted model is selected by the three criterion. When the generating model was not chosen the selected models were either the $TL-ARMA$ model or a $TL-MA$ model of order 14 or greater. Each of these selected models can capture the longer-range dependence of the data which has non-zero TPD values out to lag-15.

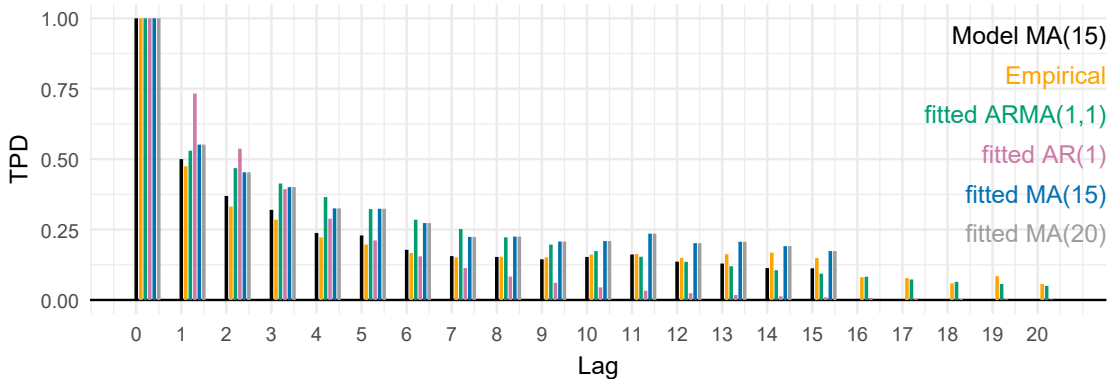


Figure 4.8: Model TPD in black, empirical TPD from a time series of length 10000 generated from the $TL - MA(15)$ fitted in Section 2.4 in yellow, and TPD from fitted models. The fitted models are an $TL - ARMA(1,1)$ in green, a $TL - AR(1)$ in pink, a $TL - MA(15)$ (the generating model) in blue, and a $TL - MA(20)$ in grey.

Table 4.1: Number of times (out of 100 simulations) that each model is selected by CLAIC, CLBIC, and the penalty from basic AIC, when data are generated from $TL - MA(15)$. Bold indicates the model with the most selections from that criterion.

Model	CLAIC	CLBIC	AIC penalty
AR(1)	0	0	0
MA(1)	0	0	0
⋮	⋮	⋮	⋮
MA(13)	0	0	0
MA(14)	5	6	0
MA(15)	**32**	**32**	**92**
MA(16)	24	20	4
MA(17)	16	15	2
MA(18)	8	5	0
MA(19)	9	4	0
MA(20)	3	3	2
ARMA(1,1)	3	15	0

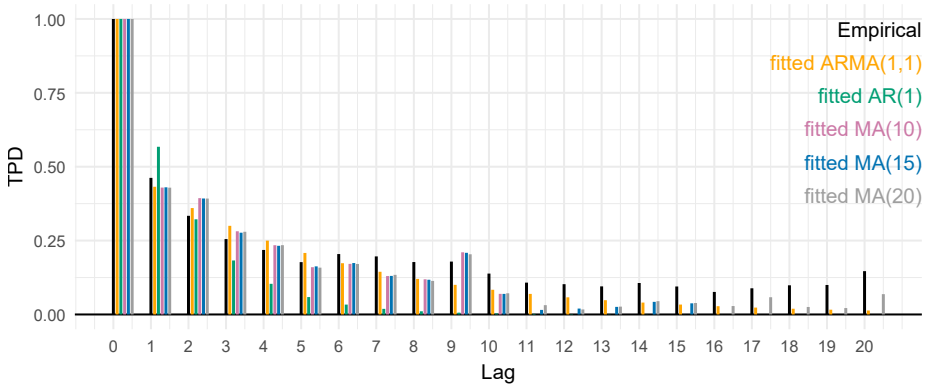


Figure 4.9: Tail Pairwise Dependence plot from past climate ERA5 FWI in Colorado (as in Figure 2.4). In black is the emperical TPD using the natural estimator (2.7). Each color is the model-based TPD from the respectively fitted model.

4.7.1 Wildfire Data

We applied our method to the ERA5 data from Colorado used to study wildfires risk in Chapter 2. This application demonstrates our motivation for the method as well as highlighting some areas for further investigation. Recall that we perform a bias reduction step in the data pre-processing (Section 2.4.2) before fitting models to this wildfire risk data. We take that pre-processed data and add $\epsilon = 0.0001$ to values that were set to zero as the HR density evaluated on the axes is always zero. We chose to include the bias reduction so that comparisons between this fitting and the innovations fitting in Chapter 2 is possible. As in the previous simulation studies we fit $TL - MA$ models of orders $q = 1, \ldots, 20$, the $TL - AR(1)$ model, and the $TL - ARMA(1, 1)$ model. We use the 0.95 quantile of the radial components as the censoring threshold and use 20 sub-series to estimate the covariance of the score. Each sub-series corresponds to a single wildfire season and thus the treatment of the data as replications matches what was done in Chapter 2 and is justifiable beyond being necessary for estimation of the covariance of the score. We expect that many environmental applications will, similarly, have natural sub-series that can be used.

Our method captures the second-order behavior in the past climate quite well (Figure 4.9). There is strong agreement between the fitted TPD values of the $TL - ARMA$ model and the $TL - MA$ models for many lags. This highlights that our method can be considered a TPD estimator which differs slightly from the natural estimator (2.7). The apparent longer term dependence

Table 4.2: Scores are listed as difference from the the best score where "0" indicates the model with the best score. Information from proxy-likelihood fitted models for Past Climate FWI time series from ERA5 data in Colorado. Models are transformed linear models; the TL is left of for brevity. The second column is the negative log-likelihood evaluated at the maximum proxy-likelihood estimate, the CL Penalty is the penalty from (4.18), CLAIC is given by (4.18), Params indicates the number of parameters in the given model, and Basic AIC is the objective function with the penalty from basic AIC.

Model	$-\ell_{cl}(\hat{\boldsymbol{\theta}}_{MPL})$	CL Penalty	CLAIC	Params	Basic AIC
AR(1)	28.35	154.39	127.15	1	45.06
MA(1)	67.79	51.37	0	1	123.95
MA(2)	39.27	1151.28	2142.77	2	68.90
MA(3)	27.41	1597.83	3012.16	3	47.20
MA(4)	19.43	1436.93	2674.41	4	33.24
MA(5)	15.99	1241.34	2276.33	5	28.35
MA(10)	3.18	4027.73	7823.51	10	12.74
MA(15)	2.99	4540.46	8848.57	15	22.35
MA(20)	0	6016.57	11794.81	20	26.37
ARMA(1,1)	4.82	23831.83	47434.97	2	0

indicated by the natural estimator is likely bias as discussed in Mhatre (2022) and in Section 2.4.2. Our method appears to have less bias in these later lags. This method indicates an increase in dependence at lag-9 which is not apparent in the innovations fit in Chapter 2. This is evidence that any dependence indicated by the proxy-likelihood is captured in $TL - MA$ models of higher orders. The $TL - AR$ fit highlights that when the model is not complex enough the fit is a balance between early and late lags. Here that means that the fitted $TL - AR$ overestimates dependence at lag-1 and underestimates it in the following lags.

Table 4.2 gives an overview of the model selection information that is readily obtained from our proxy-likelihood method. Values in the table columns entitled "$-\ell_{cl}(\hat{\boldsymbol{\theta}}_{MPL})$", "CLAIC", and "Basic AIC" are the difference between the value of the objective function for that model and the lowest (best) value of the objective function. The negative log-likelihood evaluated at the maximum proxy likelihood estimate, $-\ell_{cl}(\hat{\boldsymbol{\theta}}_{MPL})$, shows the expected behavior; more complex models have lower deviation from the best fitting model. In other words, as the order of the $TL - MA$ increases, the value of the negative composite log-likelihood decreases. It is no surprise that the most complicated model, the $TL - MA(20)$, has the lowest (best) value which was 50515.92.

Our proxy-likelihood suggests that, despite the poor fit, the $TL - AR$ model is a better fit than an $TL - MA(2)$ which has no dependence past lag-2 but a worse fit than the $TL - MA(3)$.

The penalties suggested by Varin (2008) and used in Padoan et al. (2010) are generally increasing with model complexity (Table 4.2, column 3, labeled "CL Penalty") though they are not monotonic. These penalties are very large and the penalty applied to the $TL - ARMA$ is an order of magnitude bigger than that applied to any other model. This results in the method selecting a $TL - MA(1)$ as the best fitting model which is not appropriate. A $TL - MA(1)$ would have a single non-zero TPD value at lag-1 which contradicts Figure 4.9. This observation led us to consider a criterion that uses the penalty from the basic AIC (Akaike, 1974). This simpler penalty suggests that the parsimonious $TL - ARMA(1,1)$ model with two parameters is the best model for these data, even better than highly parameterized $TL - MA$ models. Both penalized methods suggest a more parsimonious model than the $TL - MA(15)$ used in Chapter 2 which we suspected was possible but methods to fit $TL - ARMA$ models were not available at the time.

Investigation into the CLAIC penalties indicates that the curvature is higher in the $TL-ARMA$ model and the values in the covariance of the score matrix (the **J** component of (4.18)) are two to three orders of magnitude larger than those in the $TL - MA(2)$. When computing the covariance of the score we use subsets of the data (Section 4.5) and thus the peaks of these 20 log-likelihood surfaces will not be $\hat{\boldsymbol{\theta}}_{MPL}$. There is an expected balance between the $\mathbf{H}^{-1}$ component (the inverse of the curvature of the surface) and the **J** component (the covariance of the scores) as a more highly curved surface will often coincide with larger derivatives on the sub-surfaces. That balance seems to be off for our method as any practitioner would prefer the fit of the $TL - ARMA$ model than that of the $TL - MA(1)$. Estimation of the covariance of the score is hard. In addition to the usual complications with this estimation, our method uses psuedo-replicates which are made of of fixed-length sub-domains that have differing numbers of large points. This example suggests that the composite likelihood penalty proposed by Varin & Vidoni (2005) and used in Padoan et al. (2010) is not well calibrated for our method.

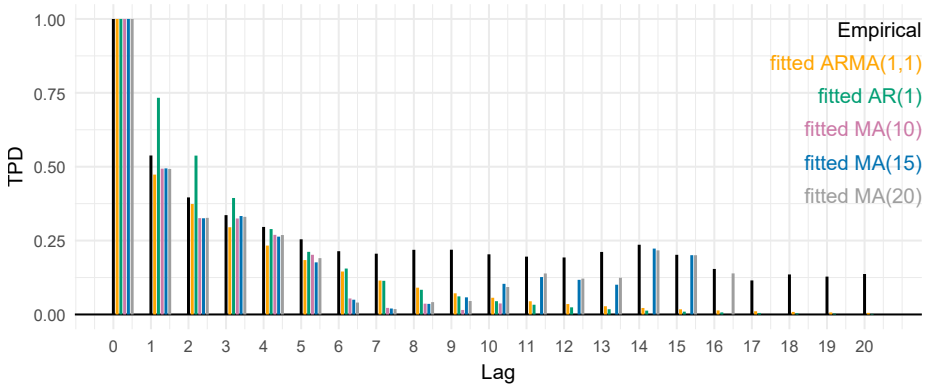


Figure 4.10: Tail Pairwise Dependence plot from present climate ERA5 FWI in Colorado (as in Figure 4.9)

We apply our method to the ERA5 data from Colorado under present climate and show the results in Figure 4.10 and Table 4.3. The fitted models seem odd in that fitted $TL - MA$ models suggest very little dependence from lags 6 to 9 and increasing dependence from lags 10 to 15. This highlights that using our method as a TPD estimator may result in estimates that are quite different than the natural estimates. One reason for this is that the natural TPD estimator (2.7) masks the radial component for large points (and ignores small points) whereas that radial information is included in the density-based fitting of this proxy-likelihood method. This is likely part of the lower estimated long-range dependence observed in our proxy-likelihood method when compared to the natural estimator which has known bias.

Table 4.3 demonstrates that the negative composite log-likelihood values perform as expected with the $TL - MA(20)$ achieving the smallest (best) value of 44248.02. Here, as with the past climate exploration, the CLAIC selects a model with one parameter which does not fit what we expect practitioners to choose when given Figure 4.10. In this case the $TL - MA(15)$ has the best score when using the basic AIC penalty, barely beating the $TL - ARMA(1,1)$ model by one unit. The subjective methods that were used in Chapter 2 also resulted in a $TL - MA(15)$ though the coefficients differ. Many practitioners would be happy to choose the model with 13 fewer parameters when the values of the objective function are this close.

Table 4.3: Information from proxy-likelihood fitted models for Past Climate FWI time series from ERA5 data in Colorado as in Table 4.2. Scores are listed as difference from the the best score where "0" indicates the model with the best score. **Bold** indicates more parsimonious models that are within 2 units of the best score.

Model	$-\ell_{cl}(\hat{\boldsymbol{\theta}}_{MPL})$	CL Penalty	CLAIC	Params	Basic AIC
AR(1)	25.19	106.18	0	1	19.71
MA(1)	68.36	1025.47	1924.91	1	106.04
MA(2)	48.94	295.40	425.95	2	69.22
MA(3)	31.03	12673.75	25146.83	3	35.40
MA(4)	19.94	1124.81	2026.75	4	15.20
MA(5)	14.49	1344.26	2454.78	5	6.31
MA(10)	14.02	1026.87	1819.05	10	15.37
MA(15)	**1.33**	12087.69	23915.31	15	0
MA(20)	0	9430.91	18599.09	20	7.32
ARMA(1,1)	14.85	11155.36	22077.71	2	**1.04**

4.8 Discussion

We have developed a maximum composite proxy-likelihood estimator for regularly varying models that have intractable likelihoods and applied it to the TLETS models of Mhatre & Cooley (2024). While developed in the context of these TLETS models adapting the method to other models is straightforward as all that is needed is the map between the parameters of the model and the TPD. Our method is able to capture the second-order tail behavior but, like all methods for the tail, is affected by the challenges inherent to estimating an asymptotic quantity (the TPD) at finite levels. The linking of two models through a summary measure implies an angular measure mismatch which is a necessary feature. We expect this feature to allows for better fitting to real data that do not have discrete angular measures but the mathematical implications of this feature are not well understood.

Application to the wildfire risk data used in Chapter 2 indicates that the composite likelihood penalties are not well calibrated and that using basic AIC penalties may be preferable as selected models coincide with what we expect subjective methods to select. While the basic AIC penalty is preferred in the real data application it performed worse in the second simulation study. Identification and justification of a more appropriate penalty is an area of active investigation.

While implementation of the method is intuitively simple optimization can be challenging. We have found that optimization routines which require the user to specify initial values are somewhat sensitive to the initial values in that they fail at some values and converge at others. In practice we have found the most success with randomly generating initial values, testing for convergence, and repeating if necessary.

Scripts for our method and for the simulations performed in this Chapter are available on `https://github.com/twixson/proxy-lhood`.

Chapter 5

Conclusion

In this dissertation we contributed to the full analysis pipeline for the modeling of extremal dependence. While these contributions could all be used in a single analysis they are only loosely connected and thus application can be widespread. Classification of extremal dependence (Chapter 3) is part of an exploratory step and thus assumptions are minimal. When modeling the dependence we rely on the mathematical framework of regular variation (Section 2.2). Specific modeling was done with TLETS models (Section 2.3) though adjusting for seasonality (Section 2.4.1) and the fitting methods of Chapter 4 are broadly applicable.

In Chapter 2 we quantified the increased probability of extreme wildfire seasons. The understanding of wildfire risk as a seasonal quantity necessitated the use of models which capture the temporal dependence in the tail because that is the part of the distribution that concerns us. A pre-requisite to the application of these models is that the data are plausibly tail stationary which is not often the case in environmental data. We handled the apparent seasonality in the tail of our data as a part of the marginal transformation which is a standard part of extremal analyses that rely on regular variation. Novel application of the TLETS models required developing a method to select the order of the model that would be used. We subjectively assessed model fit with diagnostic plots and ability to recreate summary statistics.

This project has several natural applied and methodological extensions. We focused on one location but clearly wildfire risk is a spatial phenomena. We used the FWI which is a function of several weather variables. Perhaps another summary or modeling the variables themselves would provide additional insight. Both of these extensions would require new models which are either spatio-temporal or multivariate extensions of TLETS models. In addition, the model fitting and selection is subjective. We have partially addressed this in Chapter 4.

In Chapter 3 we developed a finite-sample classifier for asymptotically defined tail dependence regimes. Traditional approaches to this challenging problem suffer from the lack of information in

the tail and require the practitioner to either specify a null or extrapolate beyond noisy quantiles. Our method is *a priori* agnostic to the dependence regime, is easily automatable, and was designed for the finite-sample case. There is no free lunch: this method requires the practitioner to determine whether their data are close enough to something in the training set to believe the output. As always, outside knowledge, expertise, and an understanding of the risks if the wrong regime is chosen should be an active part of the classification and model-selection process.

This project could be extended with a plethora of new experiments (simulation studies). We demonstrated that there is enough of a distinction between the models that, with simulated data from the limiting models, we can to reliably distinguish between them. One interesting extension would be to use methods like this to better understand when data are close enough to the limiting models to be able to make that distinction. While additional training and testing would expand the usefulness of the method one does have to consider at what point we have devoted enough resources to the distinction between AD and AI. Some researchers have determined that the best way forward is to develop models which can capture both regimes. These models are more complex which adds flexibility and challenges.

In Chapter 4 we take a stab at the problem of fitting models for extremes that have intractable likelihoods. Our approach consists of defining a model with a likelihood as a proxy for the intractable likelihood and linking the two models through a summary of the dependence which exists for both models. We use the composite likelihood approach of Padoan et al. (2010) and penalize the likelihood with basic and composite likelihood AIC penalties for model selection.

This project has room for improvement and extension. Classical time series analyses consider uncertainty quantification for the autocovariances. The analogous uncertainty quantification would be on the TPD values. This could be obtained through bootstrapping and an analytic form could be developed. Application of the method indicates that the CLAIC penalty is not well calibrated for this method which is an area of active investigation. Implementation of the method requires optimization which is challenging in even moderate dimensions. Our specific example of the

method with TLETS models could be advanced through an explicit form for the TPD from $TL - ARMA(p, q)$ models where $p, q > 1$.

Analyses in this dissertation were performed using `R` (R Core Team, 2022). We used the following packages: `cffdrs` (Wang et al., 2017), `chron` (James & Hornik, 2022), `evd` (Stephenson, 2002), `humidity` (Cai, 2019), `ismev` (Original S functions written by Janet E. Heffernan with R port and R documentation provided by Alec G. Stephenson., 2018), `keras` (Allaire & Chollet, 2023), `knitr` (Xie, 2014), `lattice` (Sarkar, 2008), `lubridate` (Grolemund & Wickham, 2011), `ncdf4` (Pierce, 2021), `texmex` (Southworth et al., 2020), and `tidyverse` (Wickham et al., 2019).

References

Abatzoglou, J. T., & Williams, A. P. (2016). Impact of anthropogenic climate change on wildfire across western US forests. *Proceedings of the National Academy of Sciences*, *113*(42), 11770-11775. doi: 10.1073/pnas.1607171113

Ahmed, M., Maume-Deschamps, V., & Ribereau, P. (2022). Recognizing a spatial extreme dependence structure: a deep learning approach. *Environmetrics*, *33*(4), Paper No. e2714, 17. doi: 10.1002/env.2714

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Allaire, J. J., & Chollet, F. (2023). keras: R interface to 'keras' [Computer software manual]. Retrieved from `https://tensorflow.rstudio.com/` (R package version 2.13.0.9000)

Asadi, P., Davison, A. C., & Engelke, S. (2015). Extremes on river networks. *The Annals of Applied Statistics*, *9*(4), 2023–2050. doi: 10.1214/15-AOAS863

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer, New York. doi: 10.1007/978-0-387-45528-0

Boos, D. D., & Stefanski, L. A. (2013). *Essential statistical inference: theory and methods* (Vol. 591). Springer.

Bopp, G. P., Shaby, B. A., & Huser, R. (2021). A hierarchical max-infinitely divisible spatial model for extreme precipitation. *Journal of the American Statistical Association*, *116*(533), 93–106. doi: 10.1080/01621459.2020.1750414

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to time series and forecasting* (2nd ed. ed.). New York, NY: Springer Nature.

Brown, B. M., & Resnick, S. I. (1977). Extreme values of independent stochastic processes. *Journal of Applied Probability*, *14*(4), 732–739.

Cai, J. (2019). humidity: Calculate water vapor measures from temperature and dew point [Computer software manual]. Retrieved from `https://github.com/caijun/humidity` (R package version 0.1.5)

Coles, S. G. (2001). *An introduction to statistical modeling of extreme values*. London: Springer-Verlag London Ltd.

Coles, S. G., Heffernan, J., & Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes (Boston)*, *2*(4), 339-365.

Coles, S. G., & Tawn, J. A. (1991). Modelling extreme multivariate events. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *53*(2), 377–392.

Cooley, D., Cisewski, J., Erhardt, R. J., Jeon, S., Mannshardt, E., Omolo, B. O., & Sun, Y. (2012). A survey of spatial extremes: Measuring spatial dependence and modeling spatial effects. *REVSTAT-Statistical Journal*, *10*(1), 135–165.

Cooley, D., Hunter, B. D., & Smith, R. L. (2019). Univariate and multivariate extremes for the environmental sciences. In *Handbook of environmental and ecological statistics* (pp. 153–180). CRC Press, Boca Raton, FL.

Cooley, D., & Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika*, *106(3)*, 587-604. doi: 10.1093/biomet/asz028

Davison, A. C., & Huser, R. (2015). Statistics of extremes. *Annual Review of Statistics and its Application*, *2*(1), 203–235.

Davison, A. C., Padoan, S. A., & Ribatet, M. (2012). Statistical Modeling of Spatial Extremes. *Statistical Science*, *27*(2), 161 – 186. doi: 10.1214/11-STS376

de Haan, L., & Ferreira, A. (2006). *Extreme value theory: an introduction* (Vol. 3). Springer.

Dey, D., & Yan, J. (Eds.). (2016). *Extreme value modeling and risk analysis : methods and applications*. Boca Raton: Chapman & Hall/CRC.

Draisma, G., Drees, H., Ferreira, A., & de Haan, L. (2004). Bivariate tail estimation: dependence in asymptotic independence. *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, *10*(2), 251–280. doi: 10.3150/bj/1082380219

Einmahl, J. H. J., de Haan, L., & Li, D. (2006). Weighted approximations of tail copula processes with application to testing the bivariate extreme value condition. *The Annals of Statistics*, *34*(4), 1987–2014. doi: 10.1214/009053606000000434

Einmahl, J. H. J., Kiriliouk, A., & Segers, J. (2018). A continuous updating weighted least squares estimator of tail dependence in high dimensions. *Extremes*, *21*, 205–233.

Embrechts, P., McNeil, A. J., & Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. In *Risk management: value at risk and beyond (Cambridge, 1998)* (pp. 176–223). Cambridge Univ. Press, Cambridge. doi: 10.1017/CBO9780511615337.008

Engelke, S., Malinowski, A., Kabluchko, Z., & Schlather, M. (2015). Estimation of Hüsler–Reiss distributions and Brown–Resnick processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *77*(1), 239–265.

European Centre for Medium-Range Weather Forecasts. (2023). *ERA5.* Retrieved from `https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation`

Fougères, A.-L., Mercadier, C., & Nolan, J. P. (2013). Dense classes of multivariate extreme value distributions. *Journal of Multivariate Analysis*, *116*, 109–129.

Gao, X., & Song, P. X.-K. (2010). Composite likelihood bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, *105*(492), 1531–1540.

Geffroy, J. (1958). Contribution à la théorie des valeurs extrêmes. *Publications de l'Institut de Statistique de l'Université de Paris*, *7*(3-4), 37–121.

Gissibl, N., & Klüppelberg, C. (2018). Max-linear models on directed acyclic graphs. *Bernoulli*, *24*(4A), 2693 – 2720. doi: 10.3150/17-BEJ941

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, *31*(4), 1208–1211.

Goldstein, H., & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society. Series A, Statistics in society*, *158*(1), 175-177.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press, Cambridge, MA.

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25. Retrieved from `https://www.jstatsoft.org/v40/i03/`

Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association*, *55*, 698–707. Retrieved from `http://links.jstor.org/sici?sici=0162-1459(196012)55:292<698:BED>2.0.CO;2-5&origin=MSN`

Gumbel, E. J., & Goldstein, N. (1964). Analysis of empirical bivariate extremal distributions. *Journal of the American Statistical Association*, *59*, 794–816. Retrieved from `http://links.jstor.org/sici?sici=0162-1459(196409)59:307<794:AOEBED>2.0.CO;2-Q&origin=MSN`

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 ieee international conference on computer vision (iccv)* (p. 1026-1034). IEEE.

Heagerty, P. J., & Lumley, T. (2000). Window subsampling of estimating functions with application to regression models. *Journal of the American Statistical Association*, *95*(449), 197–211.

Heffernan, J. E., & Tawn, J. A. (2004). A conditional approach for multivariate extreme values. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *66*(3), 497–546. (With discussions and reply by the authors) doi: 10.1111/j.1467-9868.2004.02050.x

Hicke, J., Lucatello, S., L.D., M., Dawson, J., Aguilar, M. D., Enquist, C., ... Miller, K. (2022). North America [Book Section]. In H. O. Pörtner et al. (Eds.), *Climate change 2022: Impacts, adaptation and vulnerability: Contribution of working group ii to the sixth assessment report of the intergovernmental panel on climate change* (p. 1929-2042). Cambridge, UK and New York, USA: Cambridge University Press. doi: 10.1017/9781009325844.016

Huser, R., Davison, A. C., & Genton, M. G. (2016). Likelihood estimators for multivariate extremes. *Extremes (Boston)*, *19(1)*, 79–103.

Huser, R., Opitz, T., & Thibaud, E. (2017). Bridging asymptotic independence and dependence in spatial extremes using Gaussian scale mixtures. *Spatial Statistics*, *21*, 166–186. doi: 10.1016/j.spasta.2017.06.004

Huser, R., & Wadsworth, J. L. (2019). Modeling spatial processes with unknown extremal dependence class. *Journal of the American Statistical Association*, *114*(525), 434–444. doi: 10.1080/01621459.2017.1411813

Huser, R., & Wadsworth, J. L. (2022). Advances in statistical modeling of spatial extremes. *Wiley Interdisciplinary Reviews: Computational Statistics*, *14*(1), e1537.

Hüsler, J., & Reiss, R.-D. (1989). Maxima of normal random vectors: between independence and complete dependence. *Statistics & Probability Letters*, *7*(4), 283–286.

James, D., & Hornik, K. (2022). chron: Chronological objects which can handle dates and times [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=chron` (R package version 2.3-57. S original by David James, R port by Kurt Hornik.)

Janßen, A., & Wan, P. (2020). k-means clustering of extremes. *Electronic Journal of Statistics*, *14*(1), 1211 – 1233. doi: 10.1214/20-EJS1689

Jarrett, K., Kavukcuoglu, K., Ranzato, M., & LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? In *2009 ieee 12th international conference on computer vision* (p. 2146-2153). doi: 10.1109/ICCV.2009.5459469

Jiang, Y., Cooley, D., & Wehner, M. F. (2020). Principal component analysis for extremes and application to us precipitation. *Journal of Climate*, *33*(15), 6441–6451.

Jézéquel, A., Dépoues, V., Guillemot, H., Trolliet, M., Vanderlinden, J.-P., & Yiou, P. (2018). Behind the veil of extreme event attribution. *Climatic change*, *149*(3-4), 367-383.

Kabluchko, Z., Schlather, M., & De Haan, L. (2009). Stationary max-stable fields associated to negative definite functions. *Annals of Probability*, *37*, 2042-2065.

Kiriliouk, A., & Zhou, C. (2022). *Estimating probabilities of multivariate failure sets based on pairwise tail dependence coefficients.* arXiv. doi: 10.48550/ARXIV.2210.12618

Kulik, R., & Soulier, P. (2020). *Heavy-tailed time series*. New York, NY: Springer.

Ledford, A. W., & Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, *83*(1), 169–187. doi: 10.1093/biomet/83.1.169

Ledford, A. W., & Tawn, J. A. (1997). Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society. Series B. Methodological*, *59*(2), 475–499. doi: 10.1111/1467-9868.00080

Lindsay, B. G. (1988). Composite likelihood methods. *Comtemporary Mathematics*, *80*(1), 221–239.

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml* (Vol. 30, p. 3).

Marcon, G., Padoan, S., Naveau, P., Muliere, P., & Segers, J. (2017). Multivariate nonparametric estimation of the Pickands dependence function using Bernstein polynomials. *Journal of statistical planning and inference*, *183*, 1–17.

Mhatre, N. (2022). *Transformed-linear models for time series extremes* (Unpublished doctoral dissertation). Colorado State University.

Mhatre, N., & Cooley, D. (2021). Transformed-linear models for time series extremes. *arXiv*. doi: 10.48550/arXiv.2012.06705

Mhatre, N., & Cooley, D. (2023). Transformed-linear innovations algorithm for modeling and forecasting of time series extremes. *arXiv preprint arXiv:2309.10061*.

Mhatre, N., & Cooley, D. (2024). Transformed-linear models for time series extremes. *Journal of Time Series Analysis*, *45*(5), 671–690.

Mikosch, T., & Wintenberger, O. (2024). *Extreme value theory for time series. models with power-law tails.* Springer.

National Interagency Fire Center. (2023). *Remote Automatic Weather Stations (RAWS).* Retrieved from `https://www.nifc.gov/about-us/what-is-nifc/remote-automatic-weather-stations`

Nolan, J. P. (2020). Univariate stable distributions. *Springer Series in Operations Research and Financial Engineering*, *10*, 978–3.

Original S functions written by Janet E. Heffernan with R port and R documentation provided by Alec G. Stephenson. (2018). ismev: An introduction to statistical modeling of extreme values [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=ismev` (R package version 1.42)

Padoan, S. A., Ribatet, M., & Sisson, S. A. (2010). Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, *105*(489), 263–277.

Patricola, C. M., & Wehner, M. F. (2018). Anthropogenic influences on major tropical cyclone events. *Nature (London)*, *563*(7731), 339-346.

Pickands, J. (1981). Proceedings of the 43rd session of the International Statistical Institute. In *Multivariate extreme value distributions* (pp. 859–878). International Statistical Institute Amsterdam.

Pierce, D. (2021). ncdf4: Interface to unidata netcdf (version 4 or earlier) format data files [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=ncdf4` (R package version 1.19)

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ramos, A., & Ledford, A. (2005). Regular score tests of independence in multivariate extreme values. *Extremes. Statistical Theory and Applications in Science, Engineering and Economics*, *8*(1-2), 5–26. doi: 10.1007/s10687-005-4857-4

Resnick, S. I. (2007). *Heavy-tail phenomena: probabilistic and statistical modeling* (Vol. 10). Springer Science & Business Media.

Resnick, S. I. (2008a). *Extreme values, regular variation, and point processes* (Vol. 4). Springer Science & Business Media.

Resnick, S. I. (2008b). Multivariate regular variation on cones: application to extreme values, hidden regular variation and conditioned limit laws. *Stochastics: An International Journal of Probability and Stochastics Processes*, *80*(2-3), 269–298.

Reyes-Velarde, A. (2019, January). *California's Camp fire was the costliest global disaster last year, insurance report shows.* Retrieved from `https://www.latimes.com/local/lanow/la-me-ln-camp-fire-insured-losses-20190111-story.html`

Sainsbury-Dale, M., Zammit-Mangion, A., & Huser, R. (2024). Likelihood-free parameter estimation with neural Bayes estimators. *The American Statistician*, *78*(1), 1–14.

Sarkar, D. (2008). *Lattice: Multivariate data visualization with r.* New York: Springer. Retrieved from `http://lmdvr.r-forge.r-project.org` (ISBN 978-0-387-75968-5)

Sibuya, M. (1960). Bivariate extreme statistics. I. *Annals of the Institute of Statistical Mathematics*, *11*, 195–210. doi: 10.1007/bf01682329

Smith, R. L., Tawn, J. A., & Coles, S. G. (1997). Markov chain models for threshold exceedances. *Biometrika*, *84*(2), 249–268.

Southworth, H., Heffernan, J. E., & Metcalfe, P. D. (2020). texmex: Statistical modelling of extreme values [Computer software manual]. (R package version 2.4.8)

State of California. (2023). *Stats and events.* Retrieved from `https://www.fire.ca.gov/stats-events/`

State of Colorado. (2023). *Historical wildfire information.* Retrieved from `https://dfpc.colorado.gov/sections/wildfire-information-center/historical-wildfire-information`

Stephenson, A. G. (2002, June). evd: Extreme value distributions. *R News*, *2*(2). Retrieved from https://CRAN.R-project.org/doc/Rnews/

Tawn, J. A. (1988). Bivariate extreme value theory: models and estimation. *Biometrika*, *75*(3), 397–415. doi: 10.1093/biomet/75.3.397

Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. *University of Toronto, Technical Report*, *6*.

Van Wagner, C. E. (1987). *Development and structure of the canadian forest fire weather index system*. Canadian Forestry Service.

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, *92*(1), 1–28.

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 5–42.

Varin, C., & Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, *92*(3), 519–528.

Wadsworth, J. L., & Tawn, J. A. (2012). Dependence modelling for spatial extremes. *Biometrika*, *99*(2), 253–272. doi: 10.1093/biomet/asr080

Wadsworth, J. L., Tawn, J. A., Davison, A. C., & Elton, D. M. (2017). Modelling across extremal dependence classes. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *79*(1), 149–175. doi: 10.1111/rssb.12157

Wang, X., Wotton, B. M., Cantin, A., Parisien, M.-A., Anderson, K., Moore, B., & Flannigan, M. D. (2017). cffdrs: An r package for the canadian forest fire danger rating system. *Ecological Processes*, *6*(1), 5. Retrieved from https://ecologicalprocesses.springeropen.com/articles/10.1186/s13717-017-0070-z

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi: 10.21105/joss.01686

Wixson, T. P., & Cooley, D. (2023). Attribution of seasonal wildfire risk to changes in climate: A statistical extremes approach. *Journal of Applied Meteorology and Climatology*, *62*(11), 1511–1521.

Wixson, T. P., Shaby, B. A., Philtron, D. L., Consortium, I. P. D. G., Lima, L. A., Wyman, S. K., ... Finkbeiner, S. (2024). *A three-groups non-local model for combining heterogeneous data sources to identify genes associated with Parkinson's disease.* Retrieved from `https://arxiv.org/abs/2406.05262`

Xie, Y. (2014). knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing reproducible computational research.* Chapman and Hall/CRC. Retrieved from `http://www.crcpress.com/product/isbn/9781466561595` (ISBN 978-1466561595)

Zhang, Z. (2008). Quotient correlation: a simple based alternative to Pearson's correlation. *The Annals of Statistics*, *36*(2), 1007–1030. doi: 10.1214/009053607000000866