Inference of Extreme Value Distributions using Bayesian Neural Networks

Gabriel Haeck

A Thesis

in

The Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Science (Mathematics) at

Concordia University

Montréal, Québec, Canada

September 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Gabriel Haeck**

Entitled: **Inference of Extreme Value Distributions using Bayesian Neural Networks**

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Mathematics)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

______________________ Supervisor
Dr. Mélina Mailhot

______________________ Examiner
Dr. Xiaowen Zhou

Approved by ______________________
Dr. Lea Popovic, Graduate Program Director

____
Dr. Pascale Sicotte, Dean of Faculty

____
Date

Abstract

Inference of Extreme Value Distributions using Bayesian Neural Networks

Accurate prediction of extreme weather events are crucial from a societal point of view, where the consequences of said events can have major financial and demographic impacts upon society. Extreme Value Theory (EVT) provides a statistical framework for the modelling of such extreme events. On the other hand, Bayesian Neural Networks (BNNs) extend traditional neural networks by incorporating Bayesian inference, which provides a probabilistic approach to learning and prediction in any given regression task. In this thesis, we extend the methodology of a recently introduced BNN and integrate it with EVT to be able to infer the parameters of Generalised Extreme Value (GEV) distributions. We then apply our methodology to annual maximal rainfall in Eastern Canada, where we infer and interpolate GEV parameter estimates across the interpolation region. The obtained results demonstrate that our approach outperforms Polynomial Regression and Inverse Distance Weighting methods in predicting extreme rainfall events.

Acknowledgments

I would like to first and foremost express my gratitude towards my supervisor, Dr. Mailhot, for her continual support through the many highs and lows of my Master's studies. I am especially thankful to her for introducing me to academic world as early as my undergraduate studies, helping spark my research ideas and stimulating my intellectual curiosity. Her constant optimism and joyfulness have made working with her a delight. I also extend thanks to Dr. James-Alexandre Goulet and Dr. Bhargob Deka for the many insightful discussions and for their patience when answering my numerous questions.

I would also like to thank my parents, Linda and Gaétan, for their unconditional love and for believing in my abilities, and my brother, Nicolas, for being yet the funniest and most selfless person I know.

Lastly, I am profoundly grateful for the love and support of my fiancée, Catherine, with whom I have had the joy of sharing nearly a decade of life. Her unwavering support, not only through the challenges of my Master's journey but in all aspects of my life, has been a true blessing.

I also give special thanks to my two dogs, my golden retriever Playa and newfoundland Moomoo, for being bundles of pure love and perpetual rays of sunshine in my life.

Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

Symbols

$\equiv$	Mathematical equivalence. Used for notational convenience.
$\perp\!\!\!\perp$	Independence between random variables.
$\mathbf{X}^\intercal$	Transpose operator of a matrix
X, x	(Random Variables) Capital letters represent random variables, lowercase letters a realisation.
$n!$	Factorial of $n \in \mathbb{N}$, $n! = n \cdot (n-1) \cdots 2 \cdot 1$, $1! = 1$.
${}_kC_i$	Number of combinations, ${}_kC_i = k! \Big/ (k-i)!i!$.
$\Gamma(x)$	The Gamma function, $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \, \mathrm{d}t$.
$\mathrm{sgn}(x)$	The sign function. Returns 1 if $x > 0$, -1 otherwise.

Abbreviations

AGVI	Approximate Gaussian Variance Inference
AGSI	Approximate Gaussian Skewness Inference
BNN	Bayesian Neural Network
CLT	Central Limit Theorem
EVT	Extreme Value Theory
GEV	Generalised Extreme Value (distribution)
GMA	Gaussian Multiplicative Approximation
GPD	Generalised Pareto Distributon
IDW	Inverse Distance Weight (interpolation)
i.i.d.	Independent and identically distributed
MAE	Mean Absolute Error
MGF	Moment Generating Function
MSE	Mean Squared Error
MVN	Multivariate Normal (distribution)
PWM	Probability Weighted Moments
RMSE	Root Mean Squared Error
RTS	Rauch-Tung-Striebel (algorithm)
TAGI	Tractable Analytical Gaussian Inference

Chapter 1

Introduction

The increasing trends in extreme rainfall events, driven by climate change, are becoming more apparent globally and pose significant challenges in all spheres of life, from the economic infrastructure to the geographical and sociological aspects of society. For example, the hydrological impacts of precipiation are studied in Tabari (2020) and Barbería et al. (2014) examine the social impact of extreme rainfall. Understanding and predicting these events is crucial for mitigating their adverse effects on financial, demographic, sociological and environmental systems. Traditional models have made substantial progress in weather forecasting, yet they often struggle with accurately predicting extreme events due to their complex and localised nature. In this context, the development of advanced spatial models becomes more and more important.

Extreme Value Theory (EVT) is a branch of statistics that focuses on understanding the behaviour of the extreme events of a data set, such as the events discussed above. Unlike traditional statistical methods, which analyse the overall distribution, EVT specifically examines the tail ends of the distribution where rare and significant events occur. By providing tools to model and quantify the likelihood of extreme outcomes, EVT helps in the assessment and mitigation of risks associated with these rare but impactful events.

The origins of EVT stem from the work of Fisher and Tippett (1928), who first lay the foundation for understanding the distribution of extreme values. Later, Gumbel (1958) formalises these ideas and establishes EVT as a distinct field of study by developing the Gumbel distribution, one of the three fundamental types of extreme value distributions.

Leadbetter et al. (1983) then advances the applicability of EVT by developing broader depictions of extremal behaviour, and de Haan and Ferreira (2010) provide a modern, advanced and in-depth study of the subject. Today, EVT sees application across multiple fields of study.

In finance, EVT is used to model rare events in asset pricing and porfolio management, such as significant changes in investments, market crashes, major loan defaults and many more cases. By understanding how extreme losses are modelled, financial analysts can prepare for and mitigate the impact of these financial events (Poon et al. (2004)).

In the world of insurance, EVT is of use in excess-of-loss reinsurance contracts, where the ceding company is compensated when losses exceed a certain threshold. Reinsurers evaluate the likelihood and impact of these extreme losses and can better determine the offered coverage (McNeil (1997)). EVT is also of great asset in modelling catastrophic losses that occur from large-scale natural catastrophes (hurricanes, forest fires, floods, etc.) and as such help insurers estimate possible severity claims of important magnitude and pose appropriate premiums (Embrechts et al. (1997)).

In environmental sciences, the scope of applications is vast. For example, EVT is used in flood risk assessment to model extreme precipitation and river levels (Katz et al. (2002)), to model the occurrences and magnitudes of droughts (Katz and Brown (1992)), to model both heatwaves and cold waves (Perkins et al. (2012)) and to predict the impacts of wildfires which have recently seen a dangerous increase in frequency in Canada (Canada (2024)). All of these environmental applications are of great importance to bodies of Government globally, which need to financially and logistically prepare in case such extreme environmental events occur.

Neural networks and Bayesian Neural Networks (BNNs) are important components of machine learning and more broadly of modern computation. These models have revolutionised numerous fields by providing powerful tools for pattern recognition, prediction, and decision-making. BNNs extend traditional neural networks by incorporating Bayesian inference, which provides a probabilistic approach to learning and predicting.

In BNNs, the model parameters are treated as random variables with specified prior distributions. Bayesian inference updates these priors with data to obtain posterior distributions, which represent the updated beliefs about the model parameters given the observed data (Neal (1996)). As such, one of the key benefits of BNNs is the ability to provide uncertainty estimates for predictions. By incorporating prior knowledge and regularization through Bayesian inference, they can mitigate overfitting, especially when dealing with small datasets or noisy data (Blundell et al. (2015)).

In practice, BNNs can be applied to virtually any field where computational tasks are required, whether a regression or classification problem. For example, in the world of medical diagnosis, in Leibig et al. (2017) the authors use a BNN to obtain uncertainty estimates on diagnosis classification. This helps medical experts predict the risk of obtaining a false positive or false negative diagnosis. In Natural Language Processing, having a method to measure uncertainty in predictions helps build more reliable translation models (Fortunato et al. (2019)).

The main contribution of this thesis is the integration of BNNs to EVT by taking advantage of uncertainty quantification possible with BNN to develop a BNN that can obtain the expected value, variance and skewness of predicted outputs and by which we can obtain parameter estimates of the GEV distribution. We then explore the application of our developped framework to predict and interpolate the behavior of rainfall across Eastern Canada. More specifically, the present application involves the processing of available meteorological data, the design and training of a neural network that integrates with EVT to identify and model extreme events at a given individual location, followed by the spatial interpolation over a given geographical location.

The thesis is structed as follows. In Chapter 2, we explain EVT from the univariate point of view and look at the main approaches to EVT, namely the block maxima and threshold exceedance methods. In Chapter 3, we review the foundational BNN framework we use to build our methodology, namely the Tractable Approximate Gaussian Inference (TAGI) and TAGI-V neural networks, which stem from the work of Goulet et al. (2021) and Deka et al. (2024) respectively. The main contribution of the thesis is in Chapter 4, where we extend TAGI-V to accomodate for the quantification of the skewness of predicted outputs and how it relates to EVT by obtaining GEV parameter estimates. The framework we develop is named TAGI-S, shorthand for TAGI-Skewness. In Chapter 5, we turn our attention to numerical applications of TAGI-S, namely the spatial interpolation of GEV parameters using TAGI-S. We first provide spatial interpolation for simulated data sets and then apply our interpolation methodology to extreme rainfall modelling in Eastern Canada. Chapter 6 concludes the thesis and presents possible extensions of TAGI-S and ongoing work.

The scientific contribution of this thesis can be broken down into two main aspects. The technical contribution is rooted in the development of a BNN that integrates with EVT to obtain GEV parameter estimates. On a practical aspect, we contribute to the growing body of knowledge in the field of weather prediction by demonstrating the potential of integrating EVT with BNNs to enhance our ability to anticipate and respond to extreme weather events.

Chapter 2

Extreme Value Theory

In this chapter, we review Extreme Value Theory in the univariate setting, the multivariate point of view being out of scope of the present thesis. In Section 2.1, we look at the classical construction of EVT. Then, in Section 2.2, we look at two different approaches to model extremes: the block maxima approach and the threshold exceedance approach. Lastly, in Section 2.3 we look at two estimation methods of interest that will be used throughout the remainder of the thesis.

2.1 Classical Extreme Value Theory

We review the classic approach to EVT, which is based on the study of the asymptotic behaviour of maximums (or minimums) of sequences of random variables.

2.1.1 Framework, Extremal Types Theorem

Let $\{X_1, X_2, \ldots, X_n\}$ be a finite sequence of independent and identically distributed (i.i.d.) random variables. Central Limit Theory studies the sums $S_n = \sum_{i=1}^{n} X_i$ as $n \to \infty$. On the other hand, the classical approach to EVT is concerned with the statistical behaviour and properties of the maximum of said sequence[1], namely

$$M_n = \max\left\{X_1, X_2, \ldots, X_n\right\}. \tag{2.1}$$

[1] or the minimum of the sequence, since $\min\{X_1, X_2, \ldots, X_n\} = -\max\{-X_1, -X_2, \ldots, -X_n\}$.

Assuming an underlying distribution function $F(\cdot)$, one can theoretically obtain the distribution for M_n from the i.i.d. assumption,

$$\begin{aligned}\mathbb{P}\left(M_n \leq x\right) &= \mathbb{P}\left(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x\right), \\ &= \prod_{i=1}^{n} \mathbb{P}\left(X_i \leq x\right) \\ &= \left(F(x)\right)^n \equiv F^n(x). \qquad (2.2)\end{aligned}$$

However, since $F(\cdot)$ is unknown, it is not evident how one can use Equation (2.2). As mentionned in Coles (2001), a possible solution is to first estimate $F(\cdot)$ itself from standard techniques and then consider the estimated distribution function into Equation (2.2). However, small errors in estimating $F(\cdot)$ can lead to large errors in $F^n(\cdot)$.

Another approach is to look at the asymptotic behaviour of Equation (2.2) as $n \to \infty$ to find distributions that can be approximated using extreme data. This asymptotic reasoning is equivalent to approximating the distribution of sample means by the Gaussian distribution used in the Central Limit Theorem.

If we let x^* be the right endpoint to the distribution function $F(\cdot)$, that is $x^* = \sup\{x \mid F(x) < 1\}$, then the sequence M_n of Equation (2.1) converges in probability to x^*, $M_n \xrightarrow{\text{p}} x^*$ as $n \to \infty$, since $F^n(x) \to 0$ when $x < x^*$ and $F^n(x) \to 1$ otherwise. Thus, M_n degenerates to the point mass x^*. To obtain a non-degenerate limit distribution, we employ a linear normalisation of the sequence M_n by sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$M_n^* = \frac{\max\left\{X_1, X_2, \ldots, X_n\right\} - b_n}{a_n} = \frac{M_n - b_n}{a_n} \qquad (2.3)$$

has a legitimate limiting distribution $G(\cdot)$, that is

$$F^n\left(a_n \cdot x + b_n\right) \to G(x)$$

as $n \to \infty$. In other words, instead of looking for a limiting distribution for M_n, we consider the limitting distribution of M_n^*. As it turns out, all the possible limit distributions of M_n^* can be broken down into three distributions. The extremal types theorem (Fisher and Tippett (1928), Gnedenko (1943)) states the three possible distributions.

Theorem 1 (Extremal Types Theorem)**.** *Let $\{X_1, X_2, \ldots, X_n\}$ be a sequence of i.i.d. random variables with common distribution function $F(\cdot)$, and let $M_n = \max\{X_1, X_2, \ldots, X_n\}$. If there are sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that*

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = \mathbb{P}\left(M_n^* \leq x\right) \to G(x) \tag{2.4}$$

as $n \to \infty$ for a non-degenerate distribution $G(\cdot)$, then $G(\cdot)$ belongs to one of the three following families:

$$I : G(x) = \exp\left\{-\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\}\right\}, \quad \infty < x < \infty, \quad \textit{(Gumbel Distribution)} \tag{2.5}$$

$$II : G(x) = \begin{cases} 0, & x \leq \mu \\ \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^{-\alpha}\right\}, & x > \mu \end{cases} \qquad \textit{(Fréchet Distribution)} \tag{2.6}$$

$$III : G(x) = \begin{cases} \exp\left\{-\left[-\left(\frac{x-\mu}{\sigma}\right)^{\alpha}\right]\right\}, & x < \mu \\ 0, & x \geq \mu \end{cases} \qquad \textit{(Weibull Distribution)} \tag{2.7}$$

for location parameters $\mu \in \mathbb{R}$, scale parameter $\sigma > 0$ and in the case of Equation (2.6) and Equation (2.7) shape parameter $\alpha > 0$.

Proof. See Fisher and Tippett (1928), Gnedenko (1943). □

The three families of distributions together are known as the extreme value distributions. Equation (2.5) is known as the Gumbel distribution, Equation (2.6) as the Fréchet distribution and Equation (2.7) as the Weibull distribution. Theorem 1 states that as long as the underlying distribution $F(\cdot)$ respects Equation (2.4), the maximum M_n of a sequence of i.i.d. random variables can be normalised such that M_n^* has a limitting distribution that will always be one of the Gumbel, Fréchet or Weibull distribution.

To link M_n and M_n^* together, we note that assuming Equation (2.4), for a sufficiently large enough value of n we can write

$$\mathbb{P}\left(M_n^* \leq x\right) = \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \approx G(x)$$

and then

$$\mathbb{P}\left(M_n \leq a_n \cdot x + b_n\right) = F^n\left(a_n \cdot x + b_n\right) \approx G\left(a_n \cdot x + b_n\right), \tag{2.8}$$

where we can write $G\left(a_n \cdot x + b_n\right) = G^*(x)$, which is a different member of the extreme value distribution. As such, Equation (2.8) provides us with a bridge from M_n^* to M_n, the latter being the object of interest. We will present the remainder of the theoretical notions in this chapter through the lense of M_n directly, as for example Corollary 1.1 of the next section.

When working on obtaining an estimate of the parameters of an extreme value distribution, according to Coles (2001), in early applications one would have to first determine a robust method to choose which of the Gumbel, Fréchet or Weibull distribution to use and secondly assume that subsequent inferences have the correct distribution. As we will see in the next section, a better alternative exists in which we merge the three extreme value distributions into one main distribution with a shape parameter specifying tail behaviour.

2.1.2 Generalised Extreme Value Distribution

When working with the family of extreme value distributions, instead of differentiating between the Gumbel, Fréchet and Weibull distributions, it is often more convenient to work with what is called the Generalised Extreme Value (GEV) distribution, which is a reformulation of Theorem 1 that encompasses each of the three extreme value distributions into one. We present the GEV distribution as a corollary of Theorem 1.

Corollary 1.1 (Generalised Extreme Value Distribution)**.** *Take the same assumptions as Theorem 1. If there are sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that*

$$F^n\left(a_n \cdot x + b_n\right) \rightarrow G(x) \tag{2.9}$$

as $n \to \infty$ for a non-degenerate distribution $G(\cdot)$, then $G(\cdot)$ is a GEV distribution with parameters $\Lambda = \{\mu, \sigma, \xi\}$ defined by

$$G(x;\Lambda) \equiv G(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \qquad 1+\xi\frac{(x-\mu)}{\sigma} > 0, \tag{2.10}$$

where $\mu \in \mathbb{R}$, $\sigma > 0$ and $\xi \in \mathbb{R}$[2]*. When a random variable X follows Equation (2.10), we write $X \sim GEV(\Lambda)$.*

Moreover, the class of distributions $F(\cdot)$ that satisfy Equation (2.9) are said to belong to the domain of attraction of $G(\cdot)$.

Proof. See either de Haan and Ferreira (2010), Fisher and Tippett (1928) or Gnedenko (1943). □

The GEV distribution has three parameters $\Lambda = \{\mu, \sigma, \xi\}$, namely the location parameter $\mu \in \mathbb{R}$, the scale parameter $\sigma > 0$ and the shape parameter $\xi \in \mathbb{R}$. The shape parameter ξ is sometimes referred to as the extreme value index, as seen in de Haan and Ferreira (2010). Different values of ξ in Equation (2.10) correspond to either the Gumbel, Fréchet, Gumbel or Weibull distribution. The Fréchet distribution occurs when $\xi > 0$, the Weibull distribution when $\xi < 0$ and the Gumbel distribution arises as $\xi \to 0$.

For example, assuming $\xi > 0$, we obtain the Fréchet distribution from the GEV distribution by letting $\alpha = 1/\xi$ in Equation (2.10):

$$\begin{aligned}
G(x) &= \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}\\
&= \exp\left\{-\left[1+\frac{1}{\sigma\alpha}(x-\mu)\right]^{-\alpha}\right\}\\
&= \exp\left\{-\left[\frac{x-\mu+\alpha\sigma}{\sigma\alpha}\right]^{-\alpha}\right\}\\
&= \exp\left\{-\left(\frac{x-\mu^*}{\sigma^*}\right)^{-\alpha}\right\},
\end{aligned}$$

which is Fréchet with $\mu^* = \mu - \sigma\alpha$ and $\sigma^* = \sigma\alpha$.

We plot the GEV distribution for a fixed location parameter $\mu = 30$, fixed scale

[2] We write $G(\cdot\,;\Lambda)$ as $G(\cdot)$ when the context of the parameters Λ is clear.

parameter $\sigma = 2$ and different shape parameter values $\xi = 0.5, 0, -0.5$, each value representing either the Fréchet, Gumbel or Weibull distribution respectively. The Fréchet distribution is represented in green, the Gumbel distribution in blue and the Weibull distribution in red.

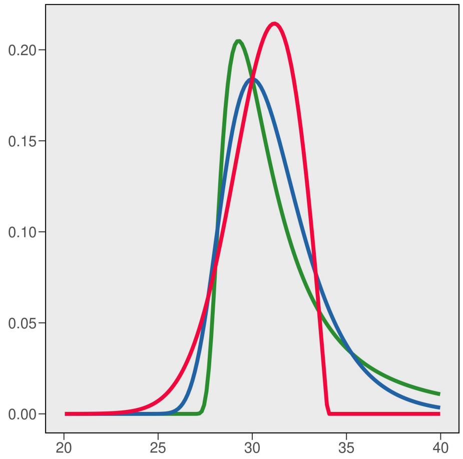


Figure 2.1: GEV Distribution for Fixed Values of $\mu = 30$, $\sigma = 2$ and Different Values of $\xi = 0.5$ (Fréchet) , $\xi = 0$ (Gumbel) and $\xi = -0.5$ (Weibull)

For a random variable $X \sim \text{GEV}(\Lambda)$, the mean, variance, skewness and kurtosis of the GEV distribution are given as follows, see Muraleedharan et al. (2009):

$$\mathbb{E}(X) = \mu + \frac{\sigma}{\xi}\left(g_1 - 1\right), \qquad \text{for } \xi < 1 \tag{2.11}$$

$$\text{Var}\left(X\right) = \left(g_2 - g_1^2\right)\frac{\sigma^2}{\xi^2}, \tag{2.12}$$

$$\text{Skew}(X) = \text{sgn}(\xi)\cdot\frac{g_3 - 3g_2g_1 + 2g_1^3}{\left(g_2 - g_1^2\right)^{3/2}}, \tag{2.13}$$

$$\text{Kurtosis}(X) = \frac{g_4 - 4g_1g_3 + 6g_2g_1^2 - 3g_1^4}{\left(g_2 - g_1\right)^2} - 3,$$

where $g_k = \Gamma\left(1 - k\cdot\xi\right)$. We have that $\mathbb{E}(X) = \infty$ when $\xi \geq 1$.

A clear advantage of having one common distribution when performing inference is that instead of having to choose between the Gumbel, Fréchet or Weibull distributions, we can apply inference directly on the extreme value index ξ. As such, one can let the data itself determine which member of the extreme value family is the most appropriate.

The quantile function for the GEV distribution returns the $(1-p)$-th quantile of the distribution:

$$G^{-1}(p;\Lambda) = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - (-\log(1-p))^{-\xi}\right], & \xi \neq 0 \\ \mu - \sigma \log(-\log(1-p)), & \xi = 0 \end{cases} \tag{2.14}$$

for $p \in (0,1]$. It is writen $G^{-1}(p;\Lambda)$ to indicate that the returned quantile is also a function of the GEV parameters. Following Coles (2001), the quantile of the GEV distribution is often called the *return level* or *return period* associated with level ${}^{1}\!/_{p}$, meaning that we expect the $(1-p)$-th quantile to be exceeded only once every ${}^{1}\!/_{p}$ years. For example, letting $p = 0.05$ means that the 95th quantile is expected to be exceeded once every ${}^{1}\!/_{0.05} = 20$ years.

2.2 Modelling Extremes

In this section, we present the two main approaches to estimating extreme values: the block maxima approach and the threshold exceedance (or peaks-over-threshold) method. The first method stems from the previous section, Section 2.1, whereas the latter method leads to the Generalised Pareto Distribution.

2.2.1 Block Maxima

The block maxima approach consists of breaking down the observed data into equal, independent and disjoint periods of observations and to then consider the maximum of each created block as the observed data. A typical application of this approach is to take blocks to represent a time period of one year and to take the maximum of each year as the set of observations we seek to fit a GEV distribution to.

Formally, let $\{X_1, X_2, \ldots, X_{n\cdot k}\}$ be a $n \cdot k$ sequence of i.i.d. random variables with underlying distribution function $F(\cdot)$ with $n, k \in \mathbb{N}$. For $j = 1, 2, \ldots, n$, we create k blocks of size n out of the $n \cdot k$ observations

$$M_j = \max_{(j-1)n < i \leq jn} \left\{X_{(j-1)n+1}, \ldots, X_i, \ldots, X_{jn}\right\} \tag{2.15}$$

such that $M_1 = \max\{X_1, \ldots, X_n\}$, $M_2 = \max\{X_{n+1}, \ldots, X_{2n}\}$, $\ldots$, $M_k = \max\left\{X_{(k-1)n+1}, \ldots, X_{kn}\right\}$.

Assuming $F(\cdot)$ to be in the domain of attraction of some GEV distribution, the sequence $\{M_1, M_2, \ldots, M_k\}$ of maximums will converge in distribution to the GEV distribution of Equation (2.10) as $k \to \infty$.

A visual representation of the block maxima method is presented in Figure 2.2, where we consider $k = 5$ blocks of size $n = 40$ with the maximum of each block being represented in blue, which form the sequence of observations $\{M_1, \ldots, M_5\}$ of Equation (2.15).

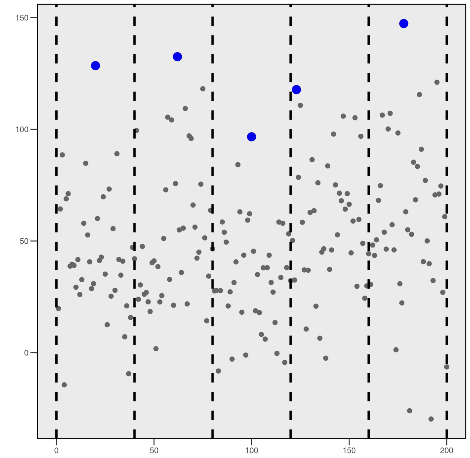


Figure 2.2: Visualisation of the Block Maxima Method

When working with block maxima, an important aspect that must be considered is that of block size selection. If block sizes are too small, then the approximation by the limit in Corollary 1.1 will be poor and create bias in estimations. On the other hand, blocks that are too large will evidently generate too few observations and lead to high model variance.

2.2.2 Threshold Exceedance

In the threshold exceedance method, we define extreme observations as all observations that occur above a specified threshold. As such, the threshold exceedance method considers all relevant high observations, whereas the block maxima approach by its nature may omit high values and include low values. Figure 2.3 shows a visual representation of the method for an arbitrary threshold value, where the points in blue are the considered data.

If we take a sequence $\{X_1, X_2, \ldots, X_n\}$ of i.i.d. random variables, we consider all values

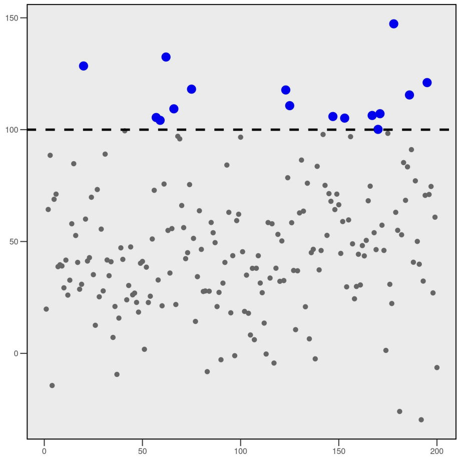


Figure 2.3: Visualisation of the Threshold Exceedance Method

that exceed a threshold u to be extreme events. In other words, for any arbitrary random variable $X \in \{X_1, X_2, \ldots, X_n\}$, we are interested in the behaviour of X above the threshold u: we want to know how the conditional excess distribution $X \mid X > u$, denoted $F_u(\cdot)$ behaves. By considering the excess $y = x - u$, we can write

$$\begin{aligned} F_u(y) &= \mathbb{P}\left(X - u \leq y \middle| X > u\right) \\ &= \frac{F(u+y) - F(u)}{1 - F(u)} \\ &= \frac{F(x) - F(u)}{1 - F(u)}, \end{aligned} \tag{2.16}$$

for $0 < y < x^* - u$, x^* being the right endpoint.

In a similar way to Equation (2.2) when explaining the context of the GEV distribution, the distribution $F(\cdot)$ in Equation (2.16) is not known. However, Pickands (1975) and Balkema and de Haan (1974) find that for a large quantity of distributions $F(\cdot)$, the conditional excess function can be approximated by a non-degenerate distribution named the Generalised Pareto Distributon.

Theorem 2 (Generalised Pareto Distribution)**.** *Consider a random variable X, with distribution function $F(\cdot)$ and conditional excess distribution $F_u(\cdot)$ with threshold u defined* *as in Equation (2.16). If there are functions $a(u) > 0$ and $b(u)$ such that*

$$F_u\left(a(u) \cdot y + b(u)\right) \to G\left(y\right)$$

as $y \to \infty$, where $G(\cdot)$ is a non-degenerate distribution, then the limitting distribution is the Generalised Pareto Distribution (GPD) defined by

$$G(y) = \begin{cases} 1 - \left(1 + \frac{\xi \cdot y}{\sigma}\right)^{-1/\xi}, & \xi \neq 0 \\ 1 - \exp\left\{-\frac{y}{\sigma}\right\}, & \xi = 0 \end{cases} \tag{2.17}$$

where $\sigma > 0$ is a scale parameter, $\xi \in \mathbb{R}$ is a shape parameter, and domains $y \geq 0$ when $\xi \geq 0$ and $0 \leq y \leq -\frac{\sigma}{\xi}$ when $\xi < 0$.

Proof. See either Pickands (1975) or Balkema and de Haan (1974). □

Pickands (1975) also shows that the family of distribution functions $F(\cdot)$ that satisfy Theorem 2 are the same distributions that lie in the domain of attraction of the GEV distribution, see Corollary 1.1. We remark that ξ in Equation (2.17) is the same shape parameter of the corresponding GEV distribution of Equation (2.10).

Selecting a threshold value u implies a tradeoff between bias and variance. If we select u to be too small, the asymptotic assumption will likely not hold and bring bias to the estimation of the model, whereas too high a threshold will imply too few observed values and lead to high variance. Coles (2001) proposes the usage of either a mean residual life plot to choose u or to estimate the GPD parameters at multiple values of u, where an appropriate threshold is u^* such that all other thresholds $u > u^*$ will yield approximately constant values of ξ.

2.2.3 Block Maxima vs. Threshold Exceedance

Recall that the threshold method picks up all relevant high observations, whereas the block maxima approach of Section 2.2.1 considers the maximum of each block of data. By definition of the block maxima framework, changing block sizes would affect the GEV parameters obtained, but would not on the other hand affect the estimation of the GPD parameters. The block maxima approach may also miss some relevant high observations and at the same time include some low values. Figure 2.4 showcases this issue, where the blue values are the block maximum values, the purple values are observations that block maxima will ignore but that would be considered by the threshold exceedance method and the red observations are values considered too low, that is under the specified threshold, that will still be considered by the block maxima.

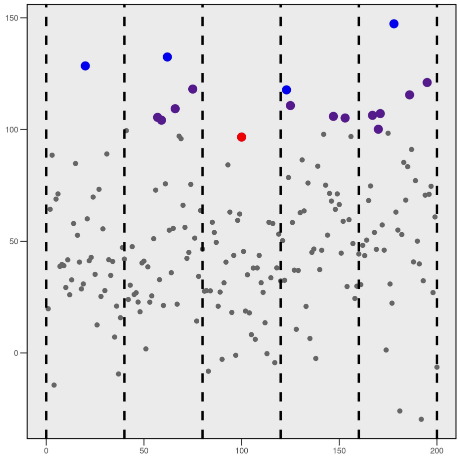


Figure 2.4: Difference Between Block Maxima and Threshold Exceedance

Based on these comments, the threshold method seems to be a more logical option to model extreme data. However, the block maxima approach still has its advantages. For one, in many situations the only information available is in the form of blocks of maximums[3]. Secondly, block periods may appear much more naturally and we can thus ignore the issue of having to select a valid threshold u. Lastly, block maxima can be better suited when dealing with heterogeneous observations. For example, if we have cyclical effects in a data set over multiple years, the threshold method might disregard observations that come from low trending years, whereas the block maxima approach would still consider the values that come from a low point in the cyclical data. We refer the reader to Coles (2001) for more details on both methodologies.

[3] Our application in Section 5.3 is an example of this case.

2.3 Estimation

In this section, we look at two specific methods of infering the parameters $\Lambda = \{\mu, \sigma, \xi\}$ of Equation (2.10) that are of interest in this thesis, namely the method of moments and the method of L-moments. Of course, there are many other possible estimation methods for the inference of the GEV parameters, two other popular methods being the Maximum Likelihood and Probability Weighted Moments (PWMs) approaches, the latter developed by Hosking et al. (1985). For an in-depth coverage of parameter estimation methods, the interested reader can consult Chapter 3 of de Haan and Ferreira (2010).

2.3.1 Method Of Moments

A straightfoward way to estimate Λ is to directly use the moments of the GEV distribution defined by Equation (2.11) through Equation (2.13). We note that the skewness depends only on the shape parameter $\{\xi\}$, the variance depends on $\{\sigma, \xi\}$ and the mean depends on all three parameters $\{\mu, \sigma, \xi\}$. As such, given that we have the mean, variance and skewness of a data set to which we seek to fit a GEV distribution to, we can iteratively obtain each parameter.

To obtain estimates $\hat{\Lambda} = \left\{\hat{\mu}, \hat{\sigma}, \hat{\xi}\right\}$ with the method of moments, we implement to following simple procedure:

1. Numerically solve for Equation (2.13), which will yield $\hat{\xi}$;

2. Given the obtained value of $\hat{\xi}$, numerically solve for Equation (2.12) to get $\hat{\sigma}$;

3. With $\hat{\xi}$ and $\hat{\sigma}$, numerically solve Equation (2.11) to obtain $\hat{\mu}$.

The procedure evidently requires numerical methods due to the Gamma functions $g_k = \Gamma\left(1 - k \cdot \xi\right), k \in \mathbb{N}$, that are present in the moments of the GEV distribution.

2.3.2 L-Moments

The method of L-moments was first presented by Hosking (1990). According to Hosking and Wallis (1997), L-moment estimators are less prone to bias and approximate parameters better than Maximum Likelihood for small sample sizes. This quality is the reason why we consider L-moments later in Section 5.3.

We first define the L-moments. Let X be a random variable with order statistics $X_{1:1} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ of a random sample of size n of X. The L-moment of order $r, r \in \mathbb{N}$, is denoted λ_r and defined as

$$\lambda_r = \frac{1}{r} \sum_{i=0}^{r-1} (-1)^i \, {}_{r-1}C_i \cdot \mathbb{E}(X_{r-i:r}) \,, \tag{2.18}$$

where ${}_{r-1}C_i = (r-1)! \big/ i!(r-1-i)!$. We also define the L-moment ratios as $\tau_r = \lambda_r \big/ \lambda_2$. The first three L-moments are given as:

$$\begin{aligned} \lambda_1 &= \mathbb{E}(X) \,, \\ \lambda_2 &= \frac{1}{2}\mathbb{E}(X_{2:2} - X_{1:2}) \,, \\ \lambda_3 &= \frac{1}{3}\mathbb{E}(X_{3:3} - 2X_{2:3} + X_{1:3}) \,. \end{aligned} \tag{2.19}$$

The justification of being able to use L-moments to describe distributions stems from the following proposition by Hosking (1990).

Proposition 1. *The L-moments of a real-valued distribution X exist if and only if X has finite mean. A distribution whose mean exists is characterized by its L-moments.*

Proof. See Hosking (1990). □

We can interpret the L-moments in a very similar way to how moments are interpreted normally: λ_1 describes the first moment of X, λ_2 measures the dispersion of the distribution, τ_3 the skewness, etc (Silva Lomba and Fraga Alves (2020)). As an example, for λ_2 in Equation (2.19), if the two values of the order statistics are close, then λ_2 will be smaller and hence measure the dispersion/scale of the distribution accordingly.

For the GEV distribution, the first two L-moments and third L-moment ratio are given by

$$\lambda_1 = \xi + \frac{\sigma}{\mu}\left[1 - \Gamma\left(1+\mu\right)\right], \tag{2.20}$$

$$\lambda_2 = \frac{\sigma}{\mu}\left(1 - 2^{-\mu}\right)\Gamma\left(1+\mu\right), \tag{2.21}$$

$$\tau_3 = \frac{2\left(1-3^{-\mu}\right)}{\left(1-2^{-\mu}\right)} - 3 \tag{2.22}$$

However, in practice, the L-moments must be calculated from a sample drawn from an unknown distribution. Hosking (1990) therefore considers the sample L-moments, which estimate the true L-moments above. Sample L-moments are based on PWMs, which were introduced by Greenwood et al. (1979) as an alternative to convential moments to summarise distributions. To remedy the fact that PWMs are hard to interpret in terms of distributional characteristics, Hosking (1986) considers linear combinations of PWMs that can be interpreted as measures of location, scale and shape of distributions.

Assuming a sample $x_1, x_2, \ldots, x_n$, the L-moments are linear combinations of order statistics of the n-sized sample $x_{1:n} \leq x_{2:n} \leq \cdots \leq x_{n:n}$ of X. With the estimates $\hat{\beta}_r$ of the PWMs

$$\hat{\beta}_r = \frac{1}{n}\sum_{i=1}^{n} {}_{n-i}C_r \cdot x_{i:n} \cdot \left({}_{n-1}C_r\right)^{-1} \tag{2.23}$$

for $r \in \{0, 1, \ldots, n-1\}$, the first three sample L-moment estimates, which we denote by ℓ_i for $i \in \{1, 2, 3\}$, are given by

$$\ell_1 = \hat{\beta}_0,$$
$$\ell_2 = \hat{\beta}_0 - 2\hat{\beta}_1,$$
$$\ell_3 = \hat{\beta}_0 - 6\hat{\beta}_1 + 6\hat{\beta}_2.$$

We additionally define the sample L-moment ratios as $t_r = {}^{\ell_r}/_{\ell_2}$ for $r \in \mathbb{N}$, where ℓ_2 is used to standardize the moments.

To obtain parameter estimates of any valid distribution, we equate the first three true L-moments of Equation (2.18) to their respective sample L-moments. That is, we set $\lambda_1 = \ell_1, \lambda_2 = \ell_2$ and $\tau_3 = t_3$ and solve for the appropriate parameters. For the GEV distribution, doing so yields parameter estimates denoted $\Lambda^{(\mathrm{LM})} = \left\{\mu^{(\mathrm{LM})}, \sigma^{(\mathrm{LM})}, \xi^{(\mathrm{LM})}\right\}$ given by

$$\hat{\Lambda}^{(\mathrm{LM})} = \begin{cases} \hat{\xi}^{(\mathrm{LM})} = 7.8590 \cdot c + 2.9554 \cdot c^2, \\ \hat{\sigma}^{(\mathrm{LM})} = \dfrac{\hat{\beta}_0 \cdot \hat{\xi}^{(\mathrm{LM})}}{\left(1 - 2^{-\hat{\xi}^{(\mathrm{LM})}}\right) \Gamma\left(1 + \hat{\xi}^{(\mathrm{LM})}\right)}, \\ \hat{\mu}^{(\mathrm{LM})} = \hat{\beta}_0 - \dfrac{\hat{\sigma}^{(\mathrm{LM})}}{\hat{\xi}^{(\mathrm{LM})}} \left[1 - \Gamma\left(1 + \hat{\xi}^{(\mathrm{LM})}\right)\right], \end{cases} \tag{2.24}$$

where

$$c = \frac{2}{3 + t_3} - \frac{\log(2)}{\log(3)},$$
$$t_3 = \frac{\ell_3}{\ell_2} = \frac{\hat{\beta}_0 - 6\hat{\beta}_1 + 6\hat{\beta}_2}{\hat{\beta}_0 - 2\hat{\beta}_1},$$

and where $\hat{\beta}_r$ are the PWM estimates in Equation (2.23). More details about the method of L-Moments are available in Hosking (1990).

Chapter 3

Bayesian Neural Networks

In this chapter, we explain the framework of the *Tractable Approximate Gaussian Inference* (TAGI) Bayesian Feed-Forward Neural Network and the *Approximate Gaussian Variance Inference* (AGVI) methodology, which stem from the work of Goulet et al. (2021) and Deka et al. (2024) respectively. These frameworks serve as the basis for the neural network model we develop in the subsequent chapter.

Section 3.1 presents the bayesian feed-forward mechanism. The Gaussian Multiplicative Approximation is covered in Section 3.1.1. Section 3.2 deals with the second half of the TAGI framework, where parameter inference is presented in Section 3.2.1, and how network hyperparameters are obtained is shown in Section 3.2.2. We conclude the coverage of TAGI in Section 3.2.3 with a numerical example.

We then cover TAGI-V in Section 3.3, where we present the AGVI framework in Section 3.3.1, which extends TAGI and allows heteroscedastic observation error to be accomodated for. A numerical example is provided in Section 3.3.2.

3.1 Approximate Gaussian Feedforward Neural Network

We start our coverage of Bayesian Neural Networks (BNNs) with how forward propagation occurs in the TAGI BNN. The principal factor at play is the Gaussian Multiplicative Approximation (GMA) in Section 3.1.1, which permits the analytical calculation of the mean and variance of multiplied terms in the network.

Consider a Feedforward Neural Network (FNN) with $\mathsf{L} \in \mathbb{N}$ layers and associated vector of covariates $\boldsymbol{X} = \{X_1, X_2, \ldots, X_{n_{\boldsymbol{X}}}\}^{\mathsf{T}}$ of dimension $n_{\boldsymbol{X}}$ and outputs $\boldsymbol{Y} = \{Y_1, Y_2, \ldots, Y_{n_{\boldsymbol{Y}}}\}^{\mathsf{T}}$ of size $n_{\boldsymbol{Y}}$ such that $\boldsymbol{X} \in \mathbb{R}^{n_{\boldsymbol{X}}}$ and $\boldsymbol{Y} \in \mathbb{R}^{n_{\boldsymbol{Y}}}$. The relationship between covariates and outputs are described by

$$\boldsymbol{y} = \boldsymbol{z}^{(0)} + \mathbf{v}, \tag{3.1}$$

where $\boldsymbol{z}^{(0)} \equiv \boldsymbol{z}^{(0)}(\boldsymbol{x})$ is the output of the neural network, which is a function of the inputs $\boldsymbol{x}$, and $\mathbf{v}$ are observation errors following a Multivariate Gaussian random variable such that $\boldsymbol{V} \sim \mathrm{MVN}(0, \boldsymbol{\Sigma}_{\boldsymbol{V}})$. For now, we consider the error $\mathbf{V}$ to be homoscedastic, that is the observation variance σ_V^2 is assumed to be constant and is set seperately from the inference procedure described below. The case when observation variance can vary with respect to covariates is covered in Section 3.3 with the AGVI framework.

We model the relationship in Equation (3.1) using a FNN with L layers in which each ith layer of the FNN consists of A hidden units $z_j^{(i)}$ $\forall j \in \{1, 2, \ldots, \mathsf{A}\}$, where we activate each hidden unit with an activation function $\phi(\cdot)$ such that the activated unit is written $a_j^{(i)} = \phi\left(z_j^{(i)}\right)$. Thus, we go from the input layer with covariates $\boldsymbol{X}$ and prior information $\boldsymbol{\Theta}^{(\varnothing)} = \left\{\boldsymbol{W}^{(\varnothing)}, \boldsymbol{B}^{(\varnothing)}\right\}$ defined by

$$\boldsymbol{W}^{\varnothing} = \begin{pmatrix} w_{1,1}^{(\varnothing)} & w_{1,2}^{(\varnothing)} & \cdots & w_{1,\mathsf{A}}^{(\varnothing)} \\ w_{2,1}^{(\varnothing)} & w_{2,2}^{(\varnothing)} & \cdots & w_{2,\mathsf{A}}^{(\varnothing)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{\mathsf{A},1}^{(\varnothing)} & w_{\mathsf{A},2}^{(\varnothing)} & \cdots & w_{\mathsf{A},\mathsf{A}}^{(\varnothing)} \end{pmatrix} \quad \text{and} \quad \boldsymbol{B}^{(\varnothing)} = \begin{pmatrix} b_1^{(\varnothing)} \\ b_2^{(\varnothing)} \\ \vdots \\ b_{\mathsf{A}}^{(\varnothing)} \end{pmatrix}$$

to the first layer of the network by calculating

$$z_j^{(1)} = \sum_{k=1}^{n_{\boldsymbol{X}}} w_{j,k}^{(\varnothing)} x_k + b_j^{(\varnothing)}, \tag{3.2}$$

where we multiply the covariates with initialised input weights $w_{j,k}^{(\varnothing)}$ and add a bias term $b_j^{(\varnothing)}$. Here, we assume that the prior input covariates, weights and bias $\boldsymbol{\Theta}^{(\varnothing)}$ is multivariate

Gaussian. To go from a given layer i to $i+1$, we generalise Equation (3.2) as

$$z_j^{(i+1)} = \sum_{k=1}^{\mathsf{A}} w_{j,k}^{(i)} a_k^{(i)} + b_j^{(i)} \tag{3.3}$$

and then propagate uncertainties with $a_j^{(i+1)} = \phi\left(z_j^{(i+1)}\right)$, such that the vector of activated units $\boldsymbol{a}^{(i)} = \left\{a_1^{(i)}, a_2^{(i)}, \cdots, a_{\mathsf{A}}^{(i)}\right\}^{\intercal}$ follows a multivariate Gaussian distribution $\boldsymbol{a}^{(i)} \sim \mathrm{MVN}\left(\boldsymbol{\mu}_{\boldsymbol{a}}^{(i)}, \boldsymbol{\Sigma}_{\boldsymbol{a}}^{(i)}\right)$.

For the output layer, we simply consider

$$z_j^{(0)} = \sum_{k=1}^{\mathsf{A}} w_{j,k}^{(\mathsf{L})} a_k^{(\mathsf{L})} + b_j^{(\mathsf{L})}. \tag{3.4}$$

We can also use matrix notation to represent the set of weights and bias at each layer as

$$\boldsymbol{Z}^{(i+1)} = \mathbf{W}^{(i)} \cdot \mathbf{A}^{(i)} + \mathbf{B}^{(i)} \equiv \begin{pmatrix} z_1^{(i+1)} \\ z_2^{(i+1)} \\ \vdots \\ z_{\mathsf{A}}^{(i+1)} \end{pmatrix} = \begin{pmatrix} w_{1,1}^{(i)} & w_{1,2}^{(i)} & \cdots & w_{1,\mathsf{A}}^{(i)} \\ w_{2,1}^{(i)} & w_{2,2}^{(i)} & \cdots & w_{2,\mathsf{A}}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{\mathsf{A},1}^{(i)} & w_{\mathsf{A},2}^{(i)} & \cdots & w_{\mathsf{A},\mathsf{A}}^{(i)} \end{pmatrix} \cdot \begin{pmatrix} a_1^{(i)} \\ a_2^{(i)} \\ \vdots \\ a_{\mathsf{A}}^{(i)} \end{pmatrix} + \begin{pmatrix} b_1^{(i)} \\ b_2^{(i)} \\ \vdots \\ b_{\mathsf{A}}^{(i)} \end{pmatrix}. \tag{3.5}$$

As such, each neuron in any given layer is a weighted combination of the activated neurons from the previous layer with an added bias term. The weights $w_{j,k}^{(i)}$ and bias term $b_j^{(i)}$ are assumed Gaussian. As such, the parameters between any two layer $\boldsymbol{\Theta}^{(i)} = \{\boldsymbol{W}^{(i)}, \mathbf{B}^{(i)}\}, i \in \{1, 2, \ldots \mathsf{L}\}$ are Gaussian. The entirity of the network parameters is denoted by $\boldsymbol{\Theta} = \{\mathbf{W}, \boldsymbol{B}\}$ and is composed of all the weights and bias terms in the network, $\boldsymbol{W} = \left\{\boldsymbol{W}^{(\varnothing)}, \boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(\mathsf{L})}\right\}$ and $\boldsymbol{B} = \left\{\boldsymbol{B}^{(\varnothing)}, \boldsymbol{B}^{(1)}, \ldots, \boldsymbol{B}^{(\mathsf{L})}\right\}$.

Figure 3.1 graphically shows the FNN mechanics within the setup of Equation (3.1). A full graphical representation of the network can be found in Appendix A.

3.1.1 Gaussian Multiplicative Approximation (GMA)

Equation (3.2), Equation (3.3) and Equation (3.4) involve the multiplication of Gaussian random variables. However, the multiplication of Gaussian random variables is known to

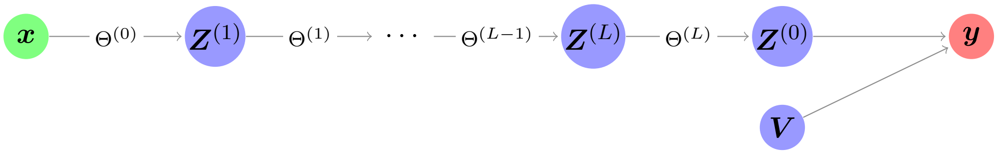


Figure 3.1: Compact Graphical Representation of the FNN

not be Gaussian[1]. To keep the propagation in the network tractable, Goulet et al. (2021) propose the GMA methodology to approximate the product of Gaussian random variables. GMA states that we can approximate the product of two Gaussian random variables by a Gaussian random variable where the mean and variance are defined in the following.

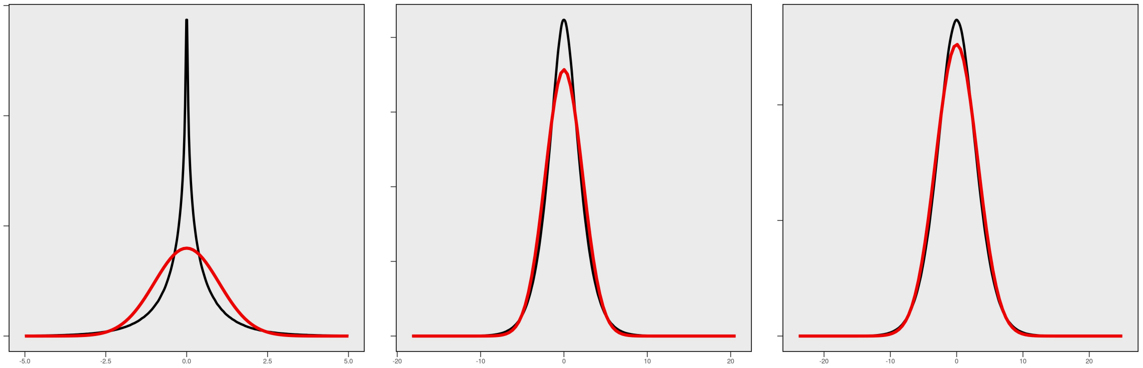
For a 4-dimensional random vector $\mathbf{X} = \{X_1, X_2, X_3, X_4\}^\intercal, \boldsymbol{X} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we can use the moment generating function to get the first two moments of any product $X_i{\cdot}X_j$ for $i, j = 1, 2, 3, 4$ and obtain

$$\begin{aligned}
\mathbb{E}(X_1X_2) &= \mu_1\mu_2 + \text{Cov}(X_1, X_2)\,, && (3.6)\\
\text{Cov}(X_3, X_1X_2) &= \text{Cov}(X_1, X_3)\,\mu_2 + \text{Cov}(X_2, X_3)\,\mu_1, && (3.7)\\
\text{Cov}(X_1X_2, X_3X_4) &= \text{Cov}(X_1, X_3)\,\text{Cov}(X_2, X_4) + \text{Cov}(X_1, X_4)\,\text{Cov}(X_2, X_3)\\
&\quad + \text{Cov}(X_1, X_3)\,\mu_2\mu_4 + \text{Cov}(X_1, X_4)\,\mu_2\mu_3\\
&\quad + \text{Cov}(X_2, X_3)\,\mu_1\mu_4 + \text{Cov}(X_2, X_4)\,\mu_1\mu_3, && (3.8)\\
\text{Var}(X_1, X_2) &= \sigma_1^2\sigma_2^2 + \text{Cov}(X_1, X_2)^2 + 2\text{Cov}(X_1, X_2)\,\mu_1\mu_2 + \sigma_1^2\mu_2^2 + \sigma_2^2\mu_1^2, && (3.9)\\
\mathbb{E}(X_1X_2X_3) &= \text{Cov}(X_1, X_2)\,\mu_3 + \text{Cov}(X_1, X_3)\,\mu_2\\
&\quad + \text{Cov}(X_2, X_3)\,\mathbb{E}(X_1) + \mu_1{\cdot}\mu_2{\cdot}\mu_3. && (3.10)
\end{aligned}$$

For instance, Figure 3.2 shows how GMA approximates the true distribution of $\sum_{i=1}^n X_i^2$ for $X_i \sim \text{Gaussian}(0, 1)$ for diffrent values of n, where the true distribution is shown in black and the approximation with the GMA is shown in red. We see that as n increases, the approximation quickly improves.

[1] For instance, the product of two standard Gaussian random variables follows the Chi-Squared distribution with degree one.

Although the product of two Gaussian random variables is not Gaussian, Equation (3.3) considers the sum of many pairwise Gaussian distributions. Given that all activation units a_k are independent, the CLT states that a large sum of independent product terms will be approximately Gaussian. Wu et al. (2019) showed that independence between activated units empirically holds, justifying the GMA.



(a) $n = 1$ (b) $n = 5$ (c) $n = 10$

Figure 3.2: GMA Approximation of $\sum_{i=1}^{n} X_i^2$ for Standard Gaussian Random Variables

To better illustrate how the GMA works in the context of the present BNN, Figure 3.2 can be interpreted as Figure 3.3, where for a fixed layer $i \in \{1, 2, \ldots, \mathsf{L}\}$ we show the passage from the activated neurons to a fixed hidden unit of the following layer $z_j^{(i+1)}$ for some fixed $j \in \{1, 2, \ldots, \mathsf{A}\}$. In interpreting Figure 3.2 as Figure 3.3, it is assumed that $w_{j,k}^{(i)} \sim \text{Gaussian}(0, 1)$, $a_k^{(i)} \sim \text{Gaussian}(0, 1)$ and $b_j^{(i)} = 0$ for all $k \in \{1, 2, \ldots, \mathsf{A}\}$. Figure 3.2a represents the case when $\mathsf{A} = 1$ in Equation (3.3), Figure 3.2b is when $\mathsf{A} = 5$ and Figure 3.2c is when $\mathsf{A} = 10$. This simplified example highlights the fact that although the GMA is a rough estimation of each multiplication $w_{j,k}^{(i)} \cdot a_k^{(i)}$, when we add each term together and increase the number of activation units A, we quickly tend towards a solid estimation.

As such, in the context of a complete neural network, with the GMA we approximate each term $w_{j,k}^{(i)} \cdot a_k^{(i)}$ in Equation (3.3) by a Gaussian random variable, which then allows us to analytically propagate uncertainty forward in the network.

To activate the jth given hidden unit at layer i, $z_j^{(i)}$, we apply an activation function $\phi(\cdot)$ such that the activated united is given by $a_j^{(i)} = \phi\left(z_j^{(i)}\right)$. Activation functions are needed to introduce non-linearity to the input of a neuron. There are many possible choices

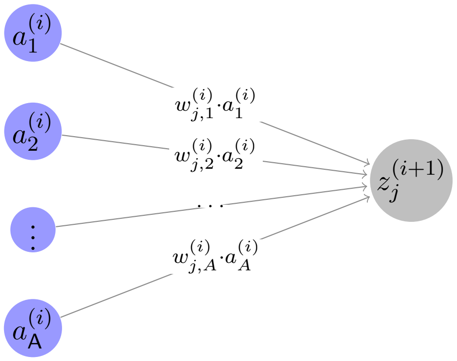

$a_1^{(i)}$
$a_2^{(i)}$
$\vdots$
$a_A^{(i)}$
$w_{j,1}^{(i)} \cdot a_1^{(i)}$
$w_{j,2}^{(i)} \cdot a_2^{(i)}$
$\cdots$
$w_{j,A}^{(i)} \cdot a_A^{(i)}$
$z_j^{(i+1)}$

Figure 3.3: Passage from an Activated Layer $i \in \{1, 2, \ldots, \mathsf{L}\}$ to the Hidden Unit $z_j^{(i+1)}$ for Some Fixed $j \in \{1, 2, \ldots, \mathsf{A}\}$

of activation functions, such as the binary step function, linear function, sigmoid function, tanh function, leaky ReLu function, softmax function, and more. Each function activates neurons in a different way and some functions can be more suitable to certain situations than others. See Sharma et al. (2020) for more details.

However, in our bayesian setting, each neuron is a random variable and it is therefore impossible to directly apply any of the activation functions mentionned above to the random variable $z_j^{(i)}$. It is proposed instead to used a locally linearised activation function, denoted $\tilde{\phi}(\cdot)$, which is locally linearised at the average value of each layer $\mathbb{E}(\mathbf{Z}) = \mu_{\boldsymbol{Z}}$ and is defined as

$$\tilde{\phi}\left(z\right) = \phi\left(\mu_{\mathbf{Z}}\right) + \frac{\partial \phi\left(\mu_{\mathbf{z}}\right)}{\partial z}\left(z - \mu_{\boldsymbol{Z}}\right). \tag{3.11}$$

It is important to note that since the linearization is done at different values of $\mu_{\mathbf{Z}}$, we maitain the non-linear dependency between inputs and outputs. We thus activate each neuron as $a_j^{(i)} = \tilde{\phi}\left(z_j^{(i)}\right)$.

3.2 Tractable Approximate Gaussian Inference for Bayesian Neural Networks (TAGI)

The previous section presented the general BNN setup alongside the GMA for forward propagation in the network. We now present the second characteristic that makes up the TAGI neural network: the inference of the weight $(\mathbf{W})$ and bias $(\mathbf{B})$ parameters $\boldsymbol{\Theta} =$

$\{\mathbf{W}, \mathbf{B}\}$. Section 3.2.1 deals with the inference of $\boldsymbol{\Theta}$ and Section 3.2.2 deals with obtaining the correct hyperparameters.

3.2.1 Inference in TAGI

With the assumptions that the set of all weights and bias terms in the network, $\boldsymbol{\Theta} = \{\mathbf{W}, \mathbf{B}\}$, are Gaussian, we can use properties of multivariate Gaussian random variables to obtain the posterior distribution of $\boldsymbol{\Theta}$. Given any n-dimensional random vector $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}^\intercal$, $\mathbf{X}$ is said to follow a Multivariate Gaussian random variable with mean vector and variance matrix

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix} \tag{3.12}$$

if it admits the following probability density function:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|} \exp\left\{ -\frac{1}{2} \left[(\mathbf{x} - \mu)^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mu) \right] \right\},$$

where $|\cdot|$ is the determinant and $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$ for all $i, j \in \{1, 2, \ldots, n\}$ (Johnson and Wichern (2002)). We can consider a partitioning of $\mathbf{X} = \{\boldsymbol{X}_1, \boldsymbol{X_2}\}^\intercal$, where $\mathbf{X}_1$ is of dimension n_1 and $\mathbf{X}_2$ of dimension n_2 such that $n_1 + n_2 = n$. Each partition of $\mathbf{X}$ has its respective mean vector and variance matrix $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ such that $\mathbf{X}_1 \sim \mathrm{MVN}(\boldsymbol{\mu}_{\mathbf{X}_1}, \boldsymbol{\Sigma}_{\mathbf{X}_1})$ and $\mathbf{X}_2 \sim \mathrm{MVN}(\boldsymbol{\mu}_{\mathbf{X}_2}, \boldsymbol{\Sigma}_{\mathbf{X}_2})$. Then, the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathrm{x}_2$ is also multivariate Gaussian with mean and variance

$$\boldsymbol{\mu}_{\mathbf{X}_1|\mathbf{X}_2} = \boldsymbol{\mu}_{\mathbf{X}_1} + \boldsymbol{\Sigma}_{\mathbf{X}_1\mathbf{X_2}} \boldsymbol{\Sigma}^{-1}_{\mathbf{X_2}} (\mathrm{x}_2 - \boldsymbol{\mu}_{\mathbf{X}_2}), \tag{3.13}$$

$$\boldsymbol{\Sigma}_{\mathbf{X}_1|\mathbf{X}_2} = \boldsymbol{\Sigma}_{\mathbf{X}_1} - \boldsymbol{\Sigma}_{\mathbf{X}_1\mathbf{X}_2} \boldsymbol{\Sigma}_{\mathbf{X}_2^{-1}} \boldsymbol{\Sigma}_{\mathbf{X}_2\mathbf{X}_1}. \tag{3.14}$$

Once forward propagation is performed as described in Section 3.1, we can perform inference on $\mathbf{\Theta}$. Given the Gaussian assumption on the weights and bias, the joint probability density function of $\mathbf{\Theta}$ and $\mathbf{Y}$ is Gaussian with mean and variance

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{\mathbf{\Theta}} \\ \boldsymbol{\mu}_{\mathbf{Y}} \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{\Theta}} & \mathbf{\Sigma}_{\mathbf{Y\Theta}}^{\mathsf{T}} \\ \mathbf{\Sigma}_{\mathbf{Y\Theta}} & \mathbf{\Sigma}_{\mathbf{Y}} \end{pmatrix}.$$

Equation (3.13) and Equation (3.14) state that we can theoretically obtain the posterior for $\mathbf{\Theta}$ in one sweep as

$$\boldsymbol{\mu}_{\mathbf{\Theta}|\mathbf{Y}} = \boldsymbol{\mu}_{\mathbf{\Theta}} + \mathbf{\Sigma}_{\mathbf{\Theta Y}}\mathbf{\Sigma}_{\mathbf{Y}}^{-1}\left(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}\right), \tag{3.15}$$

$$\mathbf{\Sigma}_{\mathbf{\Theta}|\mathbf{Y}} = \mathbf{\Sigma}_{\mathbf{\Theta}} - \mathbf{\Sigma}_{\mathbf{\Theta Y}}\mathbf{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{\Sigma}_{\mathbf{Y\Theta}}. \tag{3.16}$$

However, in practice calculating these quantities is not feasible since the involved matrices are far too dense. As a remedy, the authors consider a diagonal covariance structure for $\mathbf{\Theta}$ and use the fact given that all parameters of $\mathbf{\Theta}$ are independent between layers, each pairwise layer of hidden units are independent, meaning that we can use a layer-wise recursive inference procedure. More precisely, given the information at a layer i, the layers $i-1$ and $i+1$ are independent:

$$\mathbf{Z}^{(i+1)} \perp\!\!\!\perp \mathbf{Z}^{(i-1)} \,\Big|\, \mathbf{z}^{(i)}. \tag{3.17}$$

Thus, to perform inference recursively, we first obtain the posterior for the output layer to calculate

$$\mathbf{z}^{(0)}|\mathbf{y} \sim \mathrm{MVN}\Big(\boldsymbol{\mu}_{\mathbf{Z}^{(0)}|\mathbf{Y}}, \mathbf{\Sigma}_{\mathbf{Z}^{(0)}|\mathbf{Y}}\Big)$$

by means of Equation (3.13) and Equation (3.14). Then, the Rauch-Tung-Striebel (RTS) procedure developped by Rauch et al. (1965) is used to obtain the posterior distribution of each layer recursively. The RTS method works as follows. Note that for a given layer $i \in \{1, 2, \ldots \mathsf{L}\}$, for sake of readability we write $\left\{\mathbf{\Theta}^{(i)}, \mathbf{Z}^{(i)}\right\} \equiv \{\mathbf{\Theta}, \mathbf{Z}\}$ and $\left\{\mathbf{\Theta}^{(i+1)}, \mathbf{Z}^{(i+1)}\right\} \equiv$

$\{\mathbf{\Theta}^{+}, \mathbf{Z}^{+}\}$. First, we calculate

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{Z}|\mathbf{y}} &= \boldsymbol{\mu}_{\mathbf{Z}} + \boldsymbol{J}_{\mathbf{Z}}\left(\boldsymbol{\mu}_{\mathbf{Z}^{+}|\mathbf{y}} - \boldsymbol{\mu}_{\mathbf{Z}}\right), \\
\mathbf{\Sigma}_{\mathbf{Z}|\mathbf{y}} &= \mathbf{\Sigma}_{\mathbf{Z}} + \boldsymbol{J}_{\mathbf{Z}}\left(\mathbf{\Sigma}_{\mathbf{Z}^{+}|\mathbf{y}} - \mathbf{\Sigma}_{\mathbf{Z}^{+}}\right)\boldsymbol{J}_{\mathbf{Z}^{+}}, \\
\boldsymbol{J}_{\mathbf{Z}} &= \mathbf{\Sigma}_{\mathbf{Z}\mathbf{Z}^{+}}\mathbf{\Sigma}_{\mathbf{Z}^{+}}^{-1},
\end{aligned} \tag{3.18}$$

then given Equation (3.18) we infer the layer's parameters:

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{\Theta}|\mathbf{y}} &= \boldsymbol{\mu}_{\mathbf{Z}} + \boldsymbol{J}_{\mathbf{\Theta}}\left(\boldsymbol{\mu}_{\mathbf{Z}^{+}|\mathbf{y}} - \boldsymbol{\mu}_{\mathbf{Z}}\right), \\
\mathbf{\Sigma}_{\mathbf{\Theta}|\mathbf{y}} &= \mathbf{\Sigma}_{\mathbf{\Theta}} + \boldsymbol{J}_{\mathbf{\Theta}}\left(\mathbf{\Sigma}_{\mathbf{Z}^{+}|\mathbf{y}} - \mathbf{\Sigma}_{\mathbf{Z}^{+}}\right)\boldsymbol{J}_{\mathbf{\Theta}^{+}}, \\
\boldsymbol{J}_{\mathbf{\Theta}} &= \mathbf{\Sigma}_{\mathbf{\Theta}\mathbf{Z}^{+}}\mathbf{\Sigma}_{\mathbf{Z}^{+}}^{-1}.
\end{aligned} \tag{3.19}$$

Using the RTS procedure, for any given layer we only need to store the mean vectors $\boldsymbol{\mu}_{\mathbf{\Theta}}, \boldsymbol{\mu}_{\mathbf{Z}}$ and covariances $\mathbf{\Sigma}_{\mathbf{\Theta}}, \mathbf{\Sigma}_{\mathbf{Z}}, \mathbf{\Sigma}_{\mathbf{\Theta}\mathbf{Z}^{+}}$ and $\mathbf{\Sigma}_{\mathbf{Z}\mathbf{Z}^{+}}$. Figure 3.4 illustrates the procedure. The process is repeated for each observation value, where the posterior for $\mathbf{\Theta}$ becomes the prior for the next observation value.

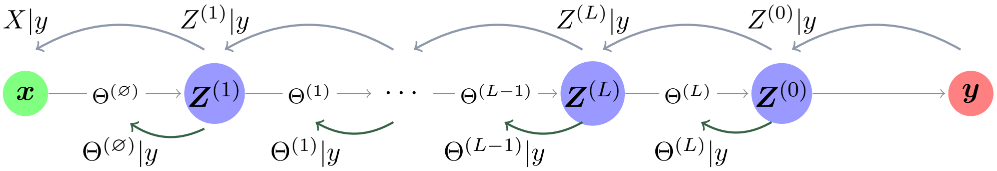


Figure 3.4: Layer-Wise Inference in TAGI

3.2.2 Hyperparameter Estimation

In a typical neural network setup, there is a great amount of different parameters in $\mathbf{\Theta}$, for which we do not have enough prior information to set proper hyperparameters $\boldsymbol{\vartheta}^{\varnothing} = \left\{\boldsymbol{\mu}_{\mathbf{\Theta}}^{(\varnothing)}, \mathbf{\Sigma}_{\mathbf{\Theta}}^{(\varnothing)}\right\}$. A solution for this problem is to run multiple *Epochs*, denoted E, where we let TAGI learn from the data set multiple times. To do so, we take the original data set of observed values $\mathcal{D}_{\text{obs}} = \{\mathcal{D}_{\text{fit}}, \mathcal{D}_{\text{val}}\}$ and split it into a training set $\mathcal{D}_{\text{fit}}$ and a validation set $\mathcal{D}_{\text{val}}$. We then recursively obtain hyperparameters through multiple epochs by using the posterior parameters of the previous iteration to get the hyperparameters of the following iteration. That is, we obtain the i-th posterior parameters $\boldsymbol{\vartheta}^{(i)} = \left\{ \boldsymbol{\mu}_{\boldsymbol{\Theta}}^{(i)}, \boldsymbol{\Sigma}_{\boldsymbol{\Theta}}^{(i)} \right\}$ and use $\boldsymbol{\vartheta}^{(i)}$ as the *prior* parameters for the $(i+1)$-th iteration. Thus, denoting by $\text{TAGI}_{\boldsymbol{\vartheta}}(\mathcal{D}_{\text{obs}}, \boldsymbol{\vartheta}^{(i)})$ the hyperparameter output using $\boldsymbol{\vartheta}^{(i)}$ and observed data set $\mathcal{D}_{\text{obs}}$, we recursively calculate

$$\boldsymbol{\vartheta}^{(i+1)} = \text{TAGI}_{\boldsymbol{\vartheta}}(\mathcal{D}_{\text{obs}}, \boldsymbol{\vartheta}^{(i)}). \tag{3.20}$$

To decide how many iterations of Equation (3.20) to consider, the following stopping algorithm is proposed. Let ℓ_{E} be the log-likelihood value of the validation set $\mathcal{D}_{\text{val}}$ at a given epoch $\mathsf{E} \in \mathbb{N}$. Let ℓ^* be the current best validation likelihood value and E^* the current optimal number of epochs. Define the *difference* parameter $\delta > 0$, *patience* parameter $\eta \in \mathbb{N}$ with maximum value η^* and let $\mathsf{E}_{\max}$ be the maximum number of allowed epochs. Each of δ, η^* and $\mathsf{E}_{\max}$ must be set by the user. Then, Algorithm 1 formally describes how we obtain the optimal number of epochs.

Algorithm 1 Obtaining Optimal Number of Epochs

```
Calculate $\ell_1$ and set $\ell^* = \ell_1$, $\mathsf{E}^* = 1$.
Set $\eta = 0$.
for $\mathsf{E} = 2$ to $\mathsf{E}_{\max}$ do
    Calculate $\ell_{\mathsf{E}}$.
    if $\ell_{\mathsf{E}} = \infty$ then
        Update $\ell^*$ to $\ell_{\mathsf{E}}$
    else if $\ell_{\mathsf{E}} - \ell_{\mathsf{E}-1} > \delta$ then          ▷ likelihood gets better by at least $\delta$
        Update $\ell^* = \ell_{\mathsf{E}}$ and $\mathsf{E}^* = \mathsf{E}$;
        Reset $\eta = 0$.                                      ▷ Restart patience counter
    else if $\ell_{\mathsf{E}} - \ell_{\mathsf{E}-1} \leq \delta$ then
        Update counter to $\eta + 1$;
        if $\eta \geq \eta^*$ then            ▷ Too many iterations without enough improvement
            Set $\mathsf{E}^* = \mathsf{E}$;
            Stop the procedure.
```

Algorithm 1 can be summarised as follows. We run the first epoch of the neural network and keep the first log-likelihood value as the current best, $\ell^* = \ell_1$. At any given epoch E, we check if the log-likelihood improves by at least δ by calculating

$$\ell_{\mathsf{E}} - \ell_{\mathsf{E}-1}. \tag{3.21}$$

If Equation (3.21) is larger than δ, that means that we have a significant enough improvement in the likelihood to update our current best likelihood to $\ell^* = \ell_{\mathsf{E}}$. We also re-set the patience counter to zero. If Equation (3.21) is smaller than δ, then we increase by one the counter η. If we reach the maximum allowed patience criteria of $\eta = \eta^*$, we stop the procedure and the current epoch E is the optimal number of epochs for the network[2]. If we have not yet reached the patience criteria, we run the next epoch to see if we have any significant likelihood improvement. The procedure continues until we either reach η^* or have considered the maximum amount of epochs decided beforehand, E_{max}.

It is important to note that the choice of δ, η^* and E_{max} will impact how TAGI performs inference. Imposing too high a value of δ will imply that Algorithm 1 will hardly accept improvements in the likelihood, and imposing too low a value of δ will cause Algorithm 1 to run too many iterations and likely induce overfitting. Likewise, E_{max} must be set high enough to ensure enough iterations of TAGI. As such, the user must be careful and adapt the settings of δ, η^* and E_{max} to the task at hand.

With a proper way of learning the network parameters $\boldsymbol{\Theta}$ over multiple epochs, the TAGI framework is complete. To train TAGI on a given data set of observed values $\mathcal{D}_{\text{obs}}$, it is important to note that we must *normalise* the data before feeding $\mathcal{D}_{\text{obs}}$ to TAGI.

3.2.3 Numerical Example

In this section we consider a regression task to showcase the capabilities and limitations of TAGI. The example demonstrates how TAGI can obtain accurate predictions with low amounts of data.

We apply TAGI to the function $y(x) = {}^{(5x)^3}/_{50} + V$, where $V \sim \text{Gaussian}(0, {}^3/_{50})$ and $x \in [-1, 1]$. This dataset is a slight modification[3] of the regression problem tackled by Hernández-Lobato and Adams (2015).

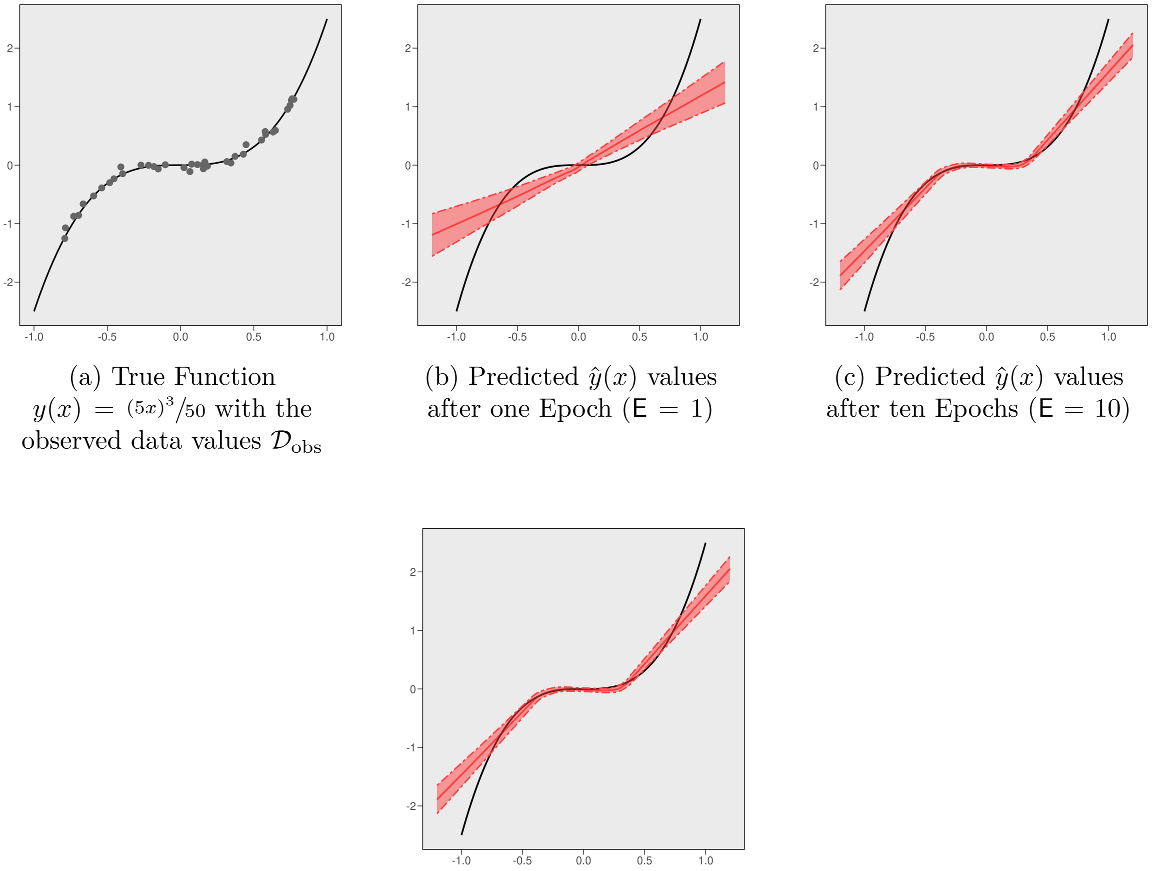
We use an observation set, denoted $\mathcal{D}_{\text{obs}} = \{\mathcal{D}_{\text{fit}}, \mathcal{D}_{\text{val}}\}$, which consists of 20 training values $\mathcal{D}_{\text{fit}} = \{\mathbf{X}_{\text{fit}}, \mathbf{Y}_{\text{fit}}\}$ and 20 validation points $\mathcal{D}_{\text{val}} = \{\mathbf{X}_{\text{val}}, \mathbf{Y}_{\text{val}}\}$ to train TAGI, and

[2] Colloquially speaking, “we have not seen a significant improvement in the likelihood for η^* iterations and thus stop the procedure”.

[3] The modification in question here is simply having already normalised the data. The original regression problem tackled by Hernández-Lobato and Adams (2015) predicts the function $y(x) = x^3$.

100 test values $\mathcal{D}_{\text{test}} = \{\mathbf{X}_{\text{test}}, \mathbf{Y}_{\text{test}}\}$ which we use to for prediction. As such, we train TAGI with $\mathcal{D}_{\text{obs}}$, which we write TAGI $(\mathcal{D}_{\text{obs}})$, and measure our performance with TAGI $(\mathcal{D}_{\text{test}})$.

The neural network is setup with one layer $(\mathsf{L} = 1)$ consisting of 100 activation units $(\mathsf{A} = 100)$ and ReLu activation function. To obtain the optimal number of epochs, we employ Algorithm 1 with difference parameter $\delta = 0.01$ and patience parameter $\eta = 5$. Doing so leads to the optimal number of epochs to be 31.



(a) True Function $y(x) = (5x)^3/50$ with the observed data values $\mathcal{D}_{\text{obs}}$

(b) Predicted $\hat{y}(x)$ values after one Epoch $(\mathsf{E} = 1)$

(c) Predicted $\hat{y}(x)$ values after ten Epochs $(\mathsf{E} = 10)$

(d) Predicted $\hat{y}(x)$ values at the optimal value of Epochs $(\mathsf{E} = 31)$

Figure 3.5: True Function $y(x)$ and Different Predicted TAGI Outputs Based on Different Number of Epochs $(\mathsf{E} = 1, 10, 31)$

Figure 3.5a presents the original function $y(x) = (5x)^3/50$ along with the observed values $\mathcal{D}_{\text{obs}}$ in gray, that is the values passed to the network for training. Figure 3.5b shows the predicted values $\hat{y}(x)$ outputted from the network where only one epoch is used $(\mathsf{E} = 1)$. The red band is the variance of the predicted output, namely $\text{Var}(\hat{y}(x))$, which is not to be confused with the variance of the observation error of V. Figure 3.5c and Figure 3.5d each show the same quantities as Figure 3.5b but but for different Epoch values, namely $\mathsf{E} = 10$ and $\mathsf{E} = 31$ respectively.

For each considered number of Epochs we calculate the MSE values of the test set $\mathcal{D}_{\text{test}}$, defined by

$$\text{MSE} = \mathbb{E}\left(\left(\boldsymbol{Y}_{\text{test}} - \hat{\boldsymbol{Y}}_{\text{test}}\right)^2\right). \tag{3.22}$$

We obtain an MSE value of 0.2189 when considering $\mathsf{E} = 1$, 0.1168 for $\mathsf{E} = 10$ and 0.0675 when taking the optimal number of epochs, $\mathsf{E} = 31$. As such, we see that as the number of Epochs increase, the MSE values decrease as expected. Overall, Figure 3.5 and the MSE values show that TAGI performs well given a relatively low amount of available data.

3.3 TAGI-V

In this section, we describe how TAGI is extended to be able to accomodate for observation error that varies with its inputs, which we refer to as heteroscedastic uncertainty. The methodology, named the Approximate Gaussian Variance Inference (AGVI) method, together with the framework of TAGI, lends itself to a modified neural network methodology referred to as TAGI-V. The AGVI framework stems from the work in Deka (2022) and Deka et al. (2024).

In Section 3.3.1 we formally present the logic of AGVI and follow with a numerical example in Section 3.3.2, which draws comparison to the numerical example of Section 3.2.3.

3.3.1 Approximate Gaussian Variance Inference (AGVI)

In Section 3.1 and Section 3.2, the model described by Equation (3.1) assumes that the error term $\mathbf{V}$ was homoscedastic: it is a constant that must be determined outside of the inference procedure. This is a key limitation to the TAGI methodology: it is impossible to take into account the possibility of having heteroscedastic uncertainty, that is observation errors that vary with the inputs. To remedy this, the authors Deka et al. (2024) use the AGVI method to be able to infer σ_V^2 and then obtain the variance of outputted values.

We present the AGVI method in the univariate case for sake of readability, but remark that the method can be extended to the Multivariate case using the diagonal covariance matrix $\mathbf{\Sigma_V} = \mathrm{diag}(\mathbf{V})$.

Assuming the model specification $y = z^{(0)} + v$, where $V \sim \text{Gaussian}(0, \sigma_V^2)$ is independent from $z^{(0)}$, the variance of Y is given by

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}\Big(Z^{(0)} + V\Big) \\ &= \mathrm{Var}\Big(Z^{(0)}\Big) + \mathrm{Var}(V) \\ &= \sigma^2_{Z^{(0)}} + \sigma^2_V. \end{aligned}$$

The objective of the AGVI procedure is to infer σ_V^2, since $\sigma^2_{Z^{(0)}}$ is already obtained through TAGI. To infer σ_V^2, the AGVI method can be summarised in two steps. First, we obtain the prior for σ_V^2 through the relationship between V, V^2 and $\mathbb{E}(V^2)$. Then, we use the posterior density of V to get the posterior information of σ_V^2.

We note that

$$\sigma_V^2 = \mathbb{E}\Big(V^2\Big) - (\mathbb{E}(V))^2 = \mathbb{E}\Big(V^2\Big) = \mu_{V^2},$$

since $\mathbb{E}(V) = 0$. This means that

$$V \sim \text{Gaussian}(0, \mu_{V^2}) \tag{3.23}$$

and that we must model V^2 to be able to infer σ_V^2.

As such, we start the AGVI method by modelling the squared error as $V^2 \sim \text{Gaussian}(\mu_{V^2}, \sigma^2_{V^2})$. Using the moments of the Gaussian distribution, we can re-write V^2 as

$$V^2 \sim \text{Gaussian}\Big(\mu_{V^2}, 2\mu^2_{V^2}\Big), \tag{3.24}$$

meaning that V^2 depends on μ_{V^2}. To maintain analytical tractability, we assume the parameter μ_{V^2} to itself follow a Gaussian random variable written $\overline{V^2} \sim \text{Gaussian}\Big(\mu_{\overline{V^2}}, \sigma^2_{\overline{V^2}}\Big)$ and thus re-write Equation (3.24) as

$$V^2 \Big| \overline{V^2} \sim \text{Gaussian}\left(\overline{v}^2, 2\left(\overline{v^2}\right)^2\right). \tag{3.25}$$

With the above established, we can graphically formulate the relationship between V, V^2 and $\mathbb{E}(V^2) \equiv \mu_{V^2}$ in Figure 3.6.

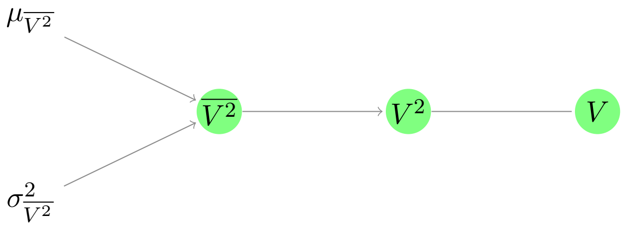


Figure 3.6: Relationship between $\overline{V^2}, V^2$ and V.

We thus obtain the prior density of V through the prior density of V^2. Following Deka et al. (2024), the prior moments of V^2 are given by $\mu_{V^2} = \mu_{\overline{V^2}}$ and $\sigma^2_{V^2} = 3\sigma^2_{\overline{V^2}} + 2\mu^2_{\overline{V^2}}$. With the distributional assumption $V \sim \text{Gaussian}(0, \sigma^2_V)$, we have that $\mathbb{E}(V) = 0$ and $\text{Var}(V) = \mathbb{E}(V^2) \equiv \mu_{\overline{V^2}}$. We can thus obtain the prior density of V as $V \sim \text{Gaussian}\Big(0, \mu_{\overline{V^2}}\Big)$.

Next, to model $\overline{V^2}$ (and hence V^2) as determined by the covariates, the authors add a branch to the output layer of the TAGI framework. That is, a second output node is added so that the output layer consists of $Z^{(0)}$, as before, and now $\overline{V^2}$. Figure 3.7 graphically represents the logic of the AGVI methodology. In this two-headed output layer, $\overline{V^2}$ has its own parameter $\Theta^{(L)}_{\overline{V^2}}$.

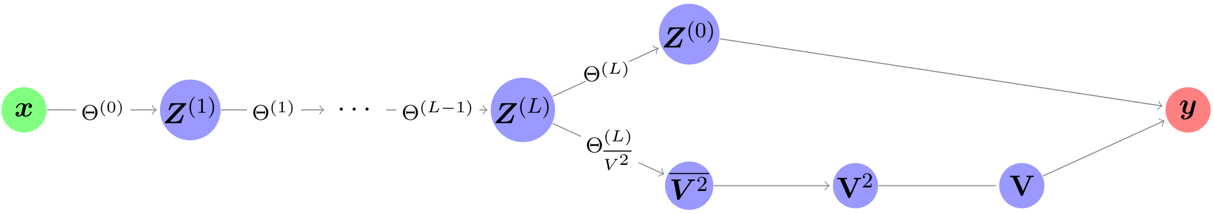


Figure 3.7: Graphical Representation of Forward Propagation with the AGVI Methodology

Recall that $\overline{V^2}$ must be positive, since it models the variance σ^2_V. As such, to ensure that $\overline{V^2}$ is positive, an exponential activation function is used. Let the activated form of

$\overline{V^2}$ be denoted $\widetilde{\overline{V^2}}$. The moments of $\widetilde{\overline{V^2}}$ are available from Goulet (2020) and are given as:

$$\mu_{\widetilde{\overline{V^2}}} = \exp\left\{\mu_{\overline{V^2}} + 0.5\sigma^2_{\overline{V^2}}\right\}, \tag{3.26}$$

$$\sigma^2_{\widetilde{\overline{V^2}}} = \exp\left\{2\mu_{\overline{V^2}} + \sigma^2_{\overline{V^2}}\right\}\cdot\left(\exp\left\{\sigma^2_{\overline{V^2}}\right\} - 1\right), \tag{3.27}$$

$$\mathrm{Cov}\left(\overline{V^2}, \widetilde{\overline{V^2}}\right) = \sigma^2_{\overline{V^2}} \cdot \mu_{\widetilde{\overline{V^2}}} \tag{3.28}$$

Thus, this concludes how forward proagation is done under AGVI. Next, we consider how to infer σ^2_V from the neural network. We first consider the posterior distribution of the joint output $\mathbf{h} = \left\{Z^{(0)}, V\right\}^\intercal$. Under the Gaussian assumptions and Equation (3.13) and Equation (3.14), we know that the distribution of $\mathbf{h}$ given y will be normal with conditional mean and variance given by

$$\boldsymbol{\mu}_{\boldsymbol{H}|y} = \mu_{\boldsymbol{H}} + \frac{\boldsymbol{\Sigma}_{\boldsymbol{H}Y}}{\sigma^2_Y}\left(y - \mu_Y\right), \tag{3.29}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{H}|y} = \boldsymbol{\Sigma}_{\boldsymbol{H}} - \frac{\boldsymbol{\Sigma}_{\boldsymbol{H}Y}\cdot\boldsymbol{\Sigma}^\intercal_{Y\boldsymbol{H}}}{\sigma^2_Y}. \tag{3.30}$$

From Equation (3.29) and Equation (3.30), we can derive the density for first $Z^{(0)}|y$ and $V|y$, then $V^2|y$ and finally $\overline{V^2}|y$. The authors obtain posterior moments of $\overline{V^2}$ and V^2 as:

$$\begin{aligned}
\mu_{V^2|y} &= \mu^2_{V|y} + \sigma^2_{V|y},\\
\sigma^2_{V^2|y} &= 2\left(\sigma^2_{V|y}\right)^2 + 4\sigma^2_{V|y}\cdot\mu^2_{V|y},\\
\mu_{\overline{V^2}|y} &= \mu_{\overline{V^2}} + k\left(\mu_{V^2|y} - \mu_{V^2}\right),\\
\sigma^2_{\overline{V^2}|y} &= \sigma^2_{\overline{V^2}} + k^2\left(\sigma^2_{V^2|y} - \sigma^2_{V^2}\right),
\end{aligned} \tag{3.31}$$

where $k = \sigma^2_{\overline{V^2}}/\sigma^2_{V^2}$. Figure 3.8 below graphically shows the posterior inference procedure for AGVI described by Equation (3.29) to Equation (3.31). From there, we take the posterior information $Z^{(0)}|y$ and $\overline{V^2}|y$ and follow the exact same inference procedure outlined by TAGI in Section 3.2.1.

Using the TAGI framework from Section 3.2 together with the AGVI procedure outlined above, it is possible to perform tractable inference of weight and bias parameters of the

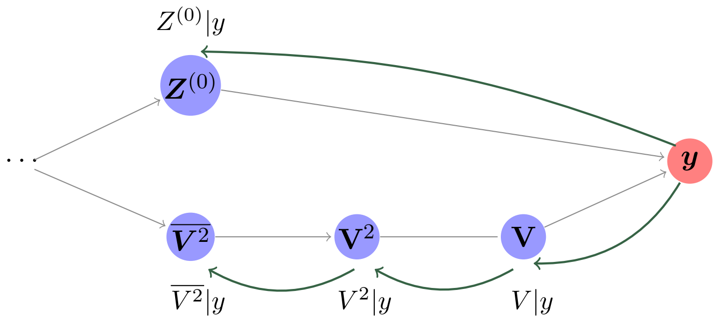


Figure 3.8: Graphical Representation of Posterior Inference for AGVI

network all the while capturing heteroscedastic uncertainty. The blending of TAGI and AGVI together is named TAGI-V.

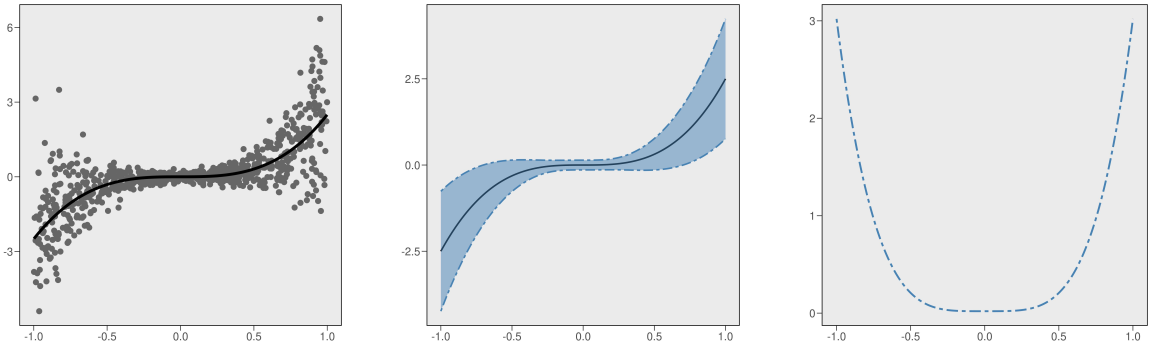
3.3.2 Numerical Example

We showcase how TAGI-V can extract heteroscedastic variance given a simulated dataset. We consider the same setup as in Section 3.2.3, that is the function $f(x) = {}^{(5x)^3}\!/_{30}$ on the domain $[-1, 1]$, with an added variance function following $v(x) = 3x^4 + 0.02$. As such, the observations carry non-constant error. Our objective is to see how well TAGI-V can detect said error.

3.3.2.1 First Simulation

We feed the neural network with 1,000 observation points $\mathcal{D}_{\text{obs}} = \{\mathcal{D}_{\text{fit}}, \mathcal{D}_{\text{val}}\}$ broken down into 800 fitting values and 200 validation values that follow $y(x) = f(x) + \sqrt{v(x)}$, where $v(x) \sim \text{Gaussian}\Big(0, \sqrt{v(x)}\Big)$. We also take 200 test data points ($\mathcal{D}_{\text{test}}$). We visualise the data set in Figure 3.9: Figure 3.9a plots $\mathcal{D}_{\text{obs}}$ with $y(x)$ in black, Figure 3.9b plots the values of $y(x) \pm \sigma_v$ and Figure 3.9c plots $v(x)$.

We train TAGI-V with $\mathcal{D}_{\text{obs}}$ using the same setup as the TAGI example in Section 3.2.3. That is, we use one layer ($\mathsf{L} = 1$) of 100 nodes ($\mathsf{A} = 100$), activated using the ReLu activation function. To determine the number of epochs used, we use Algorithm 1 with difference parameter $\delta = 0.01$ and patience parameter $\eta = 5$. With this algorithm we obtain $\mathsf{E} = 5$ and then get the following predicted values $\hat{y}(x)$ and $\hat{v}(x)$ with TAGI-V ($\mathcal{D}_{\text{test}}$)



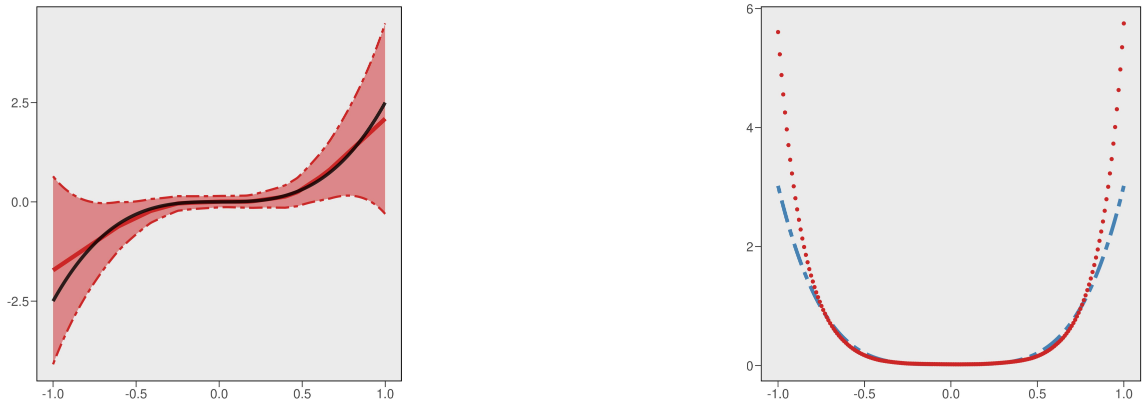
(a) Observed data $\mathcal{D}_{\text{obs}}$ with $y(x)$

(b) $y(x) = f(x) \pm \sigma_v$

(c) $v(x)$

Figure 3.9: Plots of $y(x)$ and variance function $v(x)$ in blue. The plot of $y(x)$ is added on top in black in the two leftmost plots.

as shown in Figure 3.10a and Figure 3.10b.



(a) $\hat{y}(x) = \hat{f}(x) \pm \sigma_{\hat{v}}$

(b) $v(x)$ and $\hat{v}(x)$

Figure 3.10: Plots of $\hat{y}(x)$ and variance function $\hat{v}(x)$ in red. The plot of $y(x)$ is added on top in black in the left plot and $v(x)$ is added in blue in the right plot

From the test set we obtain an MSE value (see Equation (3.22)) of 0.0296. We note that we obtain a lower MSE value than the ones in Section 3.2.3 since we use much more data for training. Analysing Figure 3.10, we can see comparing to Figure 3.9 that given the observed data with noise, TAGI-V is able to extract both the true function $f(x)$ and the variance function $v(x)$. Secondly, we note the accuracy of the predicted function $\hat{f}(x)$ given the fact that only five epochs were used. This highlights that TAGI-V can quickly extract a pattern, similarly to TAGI.

In the context of the TAGI neural network, the observation error σ_V is set as a hyperparameter, meaning that it must be set before running the neural network and cannot be adapted to the dataset as is the case here with TAGI-V. Thus, this example shows that capturing heteroscedastic error is out of scope for TAGI alone and that this issue is remedied with TAGI-V.

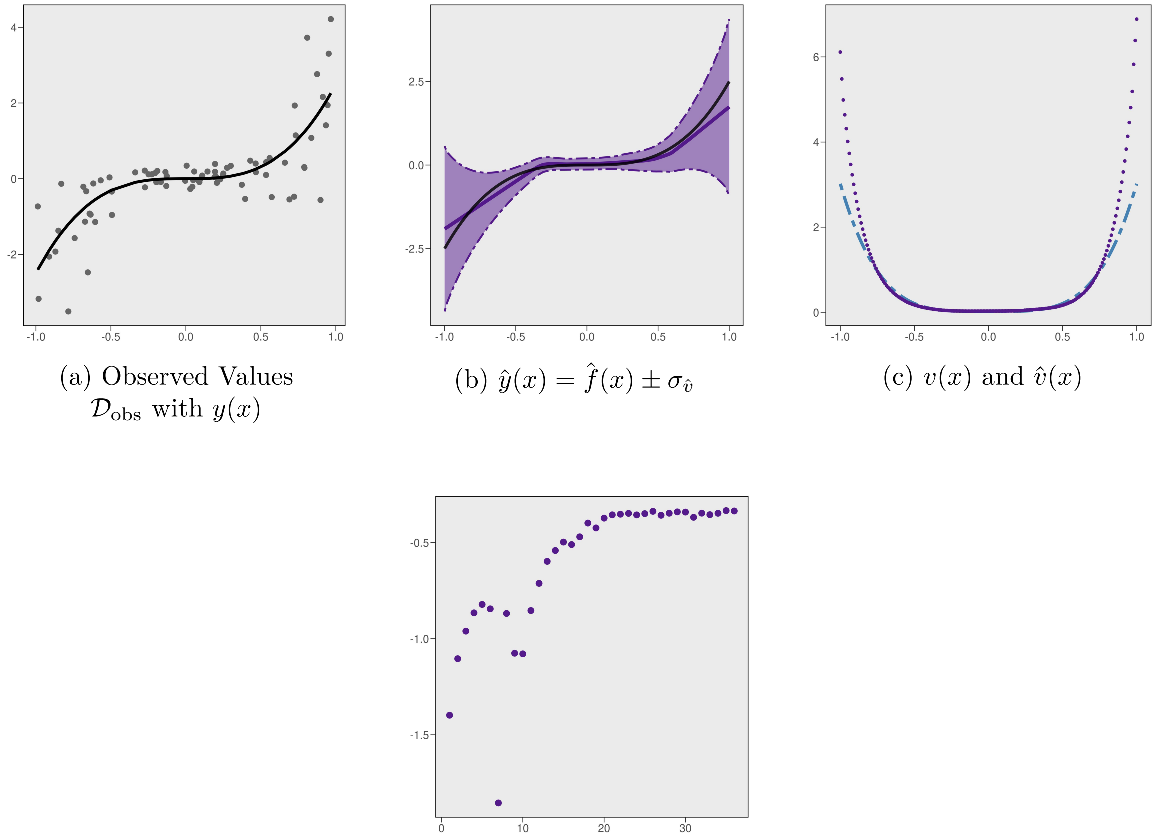
3.3.2.2 Second Simulation

We consider the same example as was described in the previous example in Section 3.3.2, the difference being that instead of having an observation set of $1,000$ points we now only take an observation data set $\mathcal{D}_{\text{obs}}$ of 100 observations, from which we take 70% to train, 20% to validate and 10% to test. The purpose is to show that TAGI-V can support low amounts of data and still be able to discern noise patterns from a dataset, without having an enormous penalty on computation.

Thus, we use 70 data points for the fitting data $\mathcal{D}_{\text{fit}}$, 20 data points for the validation set $\mathcal{D}_{\text{val}}$ and the remaining 10 values for the test set $\mathcal{D}_{\text{test}}$. The network setup is the same as Section 3.3.2, and training TAGI-V with $\mathcal{D}_{\text{obs}}$ we obtain the optimal number of epochs to be $\mathsf{E} = 26$.

Figure 3.11 presents the results of this modified example. More specifically, Figure 3.11a shows the observation data used with the true function in black, Figure 3.11b shows the predicted values and variance values from using the test set on the trained network, Figure 3.11c shows the predicted variance function along with the true variance function and Figure 3.11d depicts the validation likelihood values as a function of the number of epochs. We obtain a MSE value (see Equation (3.22)) of 0.0448, which shows good fit, however being slightly higher than the MSE value of the previous example. This is to be expected since we use less data for training.

We also notice that although the approximation of $f(x)$ of Figure 3.11b is not as smooth as in Figure 3.10a of Section 3.3.2, we are still able to adequately extract the underlying function and observation variance given the low amount of data fed to the network as showcased by Figure 3.11c. Comparing with Section 3.3.2, Figure 3.11d shows that TAGI-V takes more time to learn the pattern ($\mathsf{E} = 26$) compared to the original example ($\mathsf{E} = 5$), which is logical given the amount of data used in each example. We conclude that the



(a) Observed Values $\mathcal{D}_{\text{obs}}$ with $y(x)$

(b) $\hat{y}(x) = \hat{f}(x) \pm \sigma_{\hat{v}}$

(c) $v(x)$ and $\hat{v}(x)$

(d) Log-likelihood Values of Validation Set

Figure 3.11: Modified Example of TAGI-V with 80 Observation Data Points

advantage of TAGI being able to capture knowledge with low amounts of data still aplies for TAGI-V, with a relatively low increase in computational cost.

Chapter 4

TAGI-S

In this chapter, we present the main contribution of this thesis: we extend the methodology of TAGI and TAGI-V presented in Section 3.2 and Section 3.3 to be able to infer the skewness of outputted values. With the first three moments inferred, we can then use the method of moments defined in Section 2.3.1 to obtain estimates for the parameters of the GEV distribution. Since the method deals with the inference of skewness, we name the method TAGI-Skewness or more succinctly TAGI-S.

Let us recall 3.1, which presents the regression task

$$\mathbf{y} = \mathbf{z}^{(0)} + \mathbf{v}. \tag{4.1}$$

The observation error $\mathbf{V}$ of Equation (4.1) follows a Multivariate Gaussian distribution with mean $\boldsymbol{\mu}_{\mathbf{V}} = 0$ and variance $\boldsymbol{\Sigma}_{\mathbf{V}}$. The objective of TAGI-V in Section 3.3 (in the univariate case) was to obtain inference on σ_Y^2 through σ_V^2, which boiled down to estimating the mean of V^2 since

$$\begin{aligned}\mathrm{Var}(V) = \sigma_V^2 &= \mathbb{E}\left(V^2\right) - (\mathbb{E}(V))^2 \\ &= \mathbb{E}\left(V^2\right).\end{aligned}$$

Here, our objective is build upon the AGVI method and to obtain $\mathrm{Skew}(\mathbf{Y})$. We can view the evolution as follows: in the setting of Equation (4.1), TAGI lets us obtain $\mathbb{E}(\mathbf{Y})$, TAGI-V lets us measure Var($\mathbf{Y}$) and we now set to obtain Skew($\mathbf{Y}$). With these three moments estimated, in Section 4.4 we infer the parameters of the GEV distribution.

The essence of the procedure consists of establishing a relationship between $\mathbf{V}, \mathbf{V^2}$ and $\mathbf{V^3}$ using the GMA of Section 3.1.1 to be able to get the prior and posterior information of $\mathbf{V^3}$. Ultimately, the method presented here can be viewed as "adding another branch" to Figure 3.7, where we want to obtain information about $\mathbf{V}^3$.

4.1 Setting

For the remainder of the chapter, for sake of simplicity we present the methodology in the univariate setting. As in Section 3.3.1, we note that using the diagonal covariance matrix $\mathbf{\Sigma_V} = \mathrm{diag}(\mathbf{V})$ lets us translate the method to the multivariate case.

Before formally building the methodology, we first consider the context in which we work by obtaining the expression for Skew(Y), which helps to explain the relationship between the forthcoming methodology and both TAGI and TAGI-V.

We calculate the skewness of Y in the context of our regression model, where we employ the simplifying notation $Z^{(0)} \equiv Z$:

$$\begin{aligned}
\mathrm{Skew}(Y) &= \mathbb{E}\left(\left[\frac{Y-\mu_Y}{\sigma_Y}\right]^3\right) \\
&= \mathbb{E}\Big((Y-\mu_Y)^3\Big) \Big/ \sigma_Y^3 \\
&= \mathbb{E}\Big(Y^3 - 3Y^2\mu_Y + 3Y\mu_Y^2 - \mu_Y^3\Big) \Big/ \sigma_Y^3 \\
&= \Big\{\mathbb{E}\Big(Y^3\Big) - 3\mu_Y\mathbb{E}\Big(Y^2\Big) + 3\mu_Y^2\mathbb{E}(Y) - \mu_Y^3\Big\} \Big/ \sigma_Y^3 \\
&= \Big\{\mathbb{E}\Big(Y^3\Big) - 3\mu_Y\Big[\mathrm{Var}(Y) + \mathbb{E}(Y)^2\Big] + 3\mu_Y^3 - \mu_Y^3\Big\} \Big/ \sigma_Y^3 \\
&= \Big\{\mathbb{E}\Big(Y^3\Big) - 3\mu_Y\Big(\sigma_Z^2 + \sigma_V^2\Big) - 3\mu_Y^3 + 3\mu_Y^3 - \mu_Y^3\Big\} \Big/ \sigma_Y^3 \\
&= \Big\{\mathbb{E}\Big(Y^3\Big) - 3\mu_Y\sigma_Z^2 + 3\mu_Y\mu_{V^2} - \mu_Y^3\Big\} \Big/ \sigma_Y^3 && \text{since } \sigma_V^2 = \mu_{V^2} \\
&= \Big\{\mathbb{E}\Big(Y^3\Big) - 3\mu_Z\sigma_Z^2 + 3\mu_Z\mu_{V^2} - \mu_Z^3\Big\} \Big/ \sigma_Y^3. && \text{since } \mu_Y = \mu_Z \qquad (4.2)
\end{aligned}$$

We tackle the term $\mathbb{E}(Y^3)$ in Equation (4.2) by expanding $Y = Z + V$:

$$\begin{aligned}\mathbb{E}\Big(Y^3\Big) &= \mathbb{E}\Big((Z+V)^3\Big)\\ &= \mathbb{E}\Big(Z^3 + 3Z^2V + 3ZV^2 + V^3\Big)\\ &= \mathbb{E}\Big(Z^3\Big) + 3\mathbb{E}\Big(Z^2V\Big) + 3\mathbb{E}\Big(ZV^2\Big) + \mathbb{E}\Big(V^3\Big) \qquad (4.3)\end{aligned}$$

For each of the terms $\mathbb{E}(Z^3)$, $\mathbb{E}(Z^2V)$ and $\mathbb{E}(ZV^2)$ of Equation (4.3), we use Equation (3.10) of the GMA:

$$\begin{aligned}\mathbb{E}\Big(Z^3\Big) &= \mathbb{E}(Z\cdot Z\cdot Z) = 3\cdot \mathrm{Cov}(Z,Z)\cdot\mathbb{E}(Z)\\ &= 3\mu_Z\sigma_Z^2 \qquad (4.4)\\ \mathbb{E}\Big(Z^2V\Big) &= \mathbb{E}(Z\cdot Z\cdot V)\\ &= \mathrm{Cov}(Z,Z)\cdot\mathbb{E}(V) + \mathrm{Cov}(Z,V)\cdot\mathbb{E}(Z) + \mathrm{Cov}(Z,V)\cdot\mathbb{E}(Z)\\ &= 0 \qquad (4.5)\\ \mathbb{E}\Big(ZV^2\Big) &= \mathbb{E}(Z\cdot V\cdot V)\\ &= 2\cdot\mathrm{Cov}(Z,V)\cdot\mathbb{E}(V) + \mathrm{Cov}(V,V)\cdot\mathbb{E}(Z)\\ &= \sigma_V^2\mu_Z\\ &= \mu_Z\mu_{V^2}. \qquad (4.6)\end{aligned}$$

We replace Equation (4.4), Equation (4.5) and Equation (4.6) into Equation (4.3) to obtain

$$\mathbb{E}\Big(Y^3\Big) = 3\mu_Z\sigma_Z^2 + 3\mu_Z\mu_{V^2} + \mu_{V^3}, \qquad (4.7)$$

where we write $\mathbb{E}(V^3) \equiv \mu_{V^3}$. We can then replace once more Equation (4.7) into Equation (4.2) and get

$$\begin{aligned}\text{Skew}(Y) &= \left\{\mathbb{E}\left(Y^3\right) - 3\mu_Z\sigma_Z^2 + 3\mu_Z\mu_{V^2} - \mu_Z^3\right\} \Big/ \sigma_Y^3 && \text{since } \mu_Y = \mu_Z \\ &= \left\{3\mu_Z\sigma_Z^2 + 3\mu_Z\mu_{V^2} + \mu_{V^3} - 3\mu_Z\sigma_Z^2 + 3\mu_Z\mu_{V^2} - \mu_Z^3\right\} \Big/ \sigma_Y^3 \\ &= \frac{\mu_{V^3} - \mu_Z^3}{\sigma_Y^3}. \end{aligned} \tag{4.8}$$

From Equation (4.8), we see that in order to get the skewness of Y, we need $\mu_Z \equiv \mathbb{E}\left(Z^{(0)}\right)$ provided by TAGI, σ_Y^2 provided by AGVI and finally $\mathbb{E}(V^3)$. The following section provides the framework to obtain the prior and posterior information of V^3.

4.2 Approximate Gaussian Skewness Inference (AGSI)

Here, we formally state and develop what we name the *Approximate Gaussian Skewness Inference* (AGSI) procedure, which is the key step in merging BNNs to EVT. We remark that the AGSI framework is closely related to the AGVI procedure described in Section 3.3.1, since AGVI is still needed to obtain inference on V^3 through the relationship between V^3 and V^2.

We split the AGSI procedure into two main steps, the first step being obtaining the prior distribution of V^3 and the second step being the posterior inference of V^3 given observed data $Y = y$, which is equivalent to the inference of $\text{Skew}(Y)$ by means of Equation (4.8).

4.2.1 Prior Distribution of V^3

From Equation (4.8), we see that the skewness of Y depends on $\mu_{Z^{(0)}}$, σ_Y^2 and μ_{V^3}. The expected value of $Z^{(0)}$ and the modelling of $\sigma_Y^2 = \sigma_{Z^{(0)}}^2 + \sigma_V^2$ are handled respectively by TAGI and the AGVI procedure described in Section 3.2 and Section 3.3.1.

In a similar way to how AGVI begins its development, the AGSI method begins by modelling V^3 as a Gaussian distribution,

$$V^3 \sim \text{Gaussian}\left(\mu_{V^3}, \sigma_{V^3}^2\right). \tag{4.9}$$

We can write the variance of V^3 as

$$\begin{aligned}\mathrm{Var}\Big(V^3\Big) &= \mathbb{E}\Big(V^6\Big) - \Big(\mathbb{E}\Big(V^3\Big)\Big)^2 \\ &= 15 \cdot \sigma_V^6 \\ &= 15 \cdot \left(\sigma_V^2\right)^3 \\ &= 15 \cdot \mu_{V^2}^3,\end{aligned}$$

where we use the fact that

$$\mathbb{E}\Big(X^k\Big) = \sum_{i=0}^{k} {}_kC_i\, \mu^i \sigma^{k-i} \mathbb{E}\Big(Z^{k-i}\Big) \tag{4.10}$$

for $X \sim \mathrm{Gaussian}(\mu, \sigma^2)$, which is proved in Appendix B.1. Here we use Equation (4.10) with $\mu = \mu_V = 0$ and $\sigma = \sigma_V$. As such, we can write the distribution of V^3 in terms of μ_{V^3} and μ_{V^2}:

$$V^3 \Big| \mu_{V^2}, \mu_{V^3} \sim \mathrm{Gaussian}\Big(\mu_{V^3}, 15\mu_{V^2}^3\Big)\,. \tag{4.11}$$

To maintain analytical tractability, we assume that both hyperarameters μ_{V^2} and μ_{V^3} follow a Gaussian random variable with their respective means and variances. That is, we describe the means μ_{V^2} and μ_{V^3} with random variables that take the form $\mu_{V^2} \equiv \overline{V^2} \sim \mathrm{Gaussian}\Big(\mu_{\overline{V^2}}, \sigma^2_{\overline{V^2}}\Big)$ and $\mu_{V^3} \equiv \overline{V^3} \sim \mathrm{Gaussian}\Big(\mu_{\overline{V^3}}, \sigma^2_{\overline{V^3}}\Big)$. We can thus re-write Equation (4.11) as

$$V^3 \Big| \overline{v^2}, \overline{v^3} \sim \mathrm{Gaussian}\left(\overline{v^3}, 15\left(\overline{v^2}\right)^3\right)\,. \tag{4.12}$$

To better understand how the variables $V^2, V^3, \overline{V^2}$ and $\overline{V^3}$ are related, Figure 4.1 graphically presents the relationships between each random variable.

Figure 4.1 is very similar to Figure 3.6 from Section 3.3.1, where here we add another branch for V^3. The only other notable difference is the link from $\overline{V^2}$ to V^3, which stems from Equation (4.12). We still obtain the prior distribution of V through V^2, described by Equation (3.23).

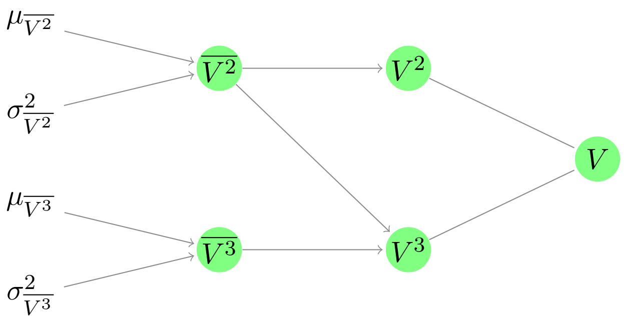


Figure 4.1: Relationship between $\overline{V^2}, V^2$, $\overline{V^3}, V^3$ and V.

The authors Deka et al. (2024) use the GMA to obtain prior mean and variance for V^2

$$\mu_{V^2} = \mu_{\overline{V^2}}, \tag{4.13}$$

$$\sigma^2_{V^2} = 3\sigma^2_{\overline{V^2}} + 2\mu^2_{\overline{V^2}}. \tag{4.14}$$

Given Equation (4.12), we represent the random variables $\overline{V^2}, V^3$ and $\overline{V^3}$ as

$$\begin{aligned}
\overline{V^2} &= \mu_{\overline{V^2}} + \sigma_{\overline{V^2}} \cdot \zeta, && \zeta \sim \text{Gaussian}(0,1) \\
V^3 &= \overline{V^3} + \sqrt{15}\,\overline{V^2}^{3/2} \cdot \epsilon, && \epsilon \sim \text{Gaussian}(0,1) \\
\overline{V^3} &= \mu_{\overline{V^3}} + \sigma_{\overline{V^3}} \cdot \nu, && \nu \sim \text{Gaussian}(0,1)
\end{aligned} \tag{4.15}$$

where each ζ, ϵ and ν are standard Gaussian random variables. We can then use the GMA to obtain the moments of V^3.

$$\begin{aligned}
\mathbb{E}\Big(V^3\Big) &= \mathbb{E}\Big(\overline{V^3} + \sqrt{15}\,\overline{V^2}^{3/2} \cdot \epsilon\Big) \\
&= \mathbb{E}\Big(\overline{V^3}\Big) + \sqrt{15} \cdot \mathbb{E}\Big(\overline{V^2}^{3/2} \cdot \epsilon\Big) \\
&= \mathbb{E}\Big(\overline{V^3}\Big) \\
&= \mu_{\overline{V^3}},
\end{aligned} \tag{4.16}$$

where we use the independence of $\overline{V^2}^{3/2}$ and ϵ to get $\mathbb{E}\left(\overline{V^2}^{3/2}\cdot\epsilon\right) = 0$. Next,

$$\begin{align}
\mathrm{Var}\left(\overline{V^3}\right) &= \mathrm{Var}\left(\overline{V^3} + \sqrt{15}\,\overline{V^2}^{3/2}\cdot\epsilon\right) \nonumber\\
&= \mathrm{Var}\left(\overline{V^3}\right) + 15\mathrm{Var}\left(\overline{V^2}^{3/2}\cdot\epsilon\right) \nonumber\\
&= \sigma^2_{\overline{V^3}} + 15\left[\mathbb{E}\left(\overline{V^2}^{3}\cdot\epsilon^2\right) - \left(\mathbb{E}\left(\overline{V^2}^{3/2}\right)\cdot\mathbb{E}(\epsilon)\right)^2\right] \nonumber\\
&= \sigma^2_{\overline{V^3}} + 15\cdot\mathbb{E}\left(\overline{V^2}^{3}\cdot\epsilon^2\right) \tag{4.17}\\
&= \sigma^2_{\overline{V^3}} + 15\left[\mathbb{E}\left(\overline{V^2}\right)\cdot\mathbb{E}\left(\epsilon^2\right) + \mathrm{Cov}\left(\overline{V^2}^{3}, \epsilon^2\right)\right] \tag{4.18}\\
&= \sigma^2_{\overline{V^3}} + 15\mathbb{E}\left(\overline{V^2}^{3}\right), \tag{4.19}
\end{align}$$

where we first use the GMA to treat $\overline{V^2}^3$ and ϵ^2 as Gaussian distributions, and then use GMA Equation (3.7) to obtain

$$\begin{aligned}
\mathrm{Cov}\left(\overline{V^2}^{3}, \epsilon\cdot\epsilon\right) &= \mathrm{Cov}\left(\overline{V^2}^{3}, \epsilon\right)\cdot\mathbb{E}(\epsilon) + \mathrm{Cov}\left(\overline{V^2}^{3}, \epsilon\right)\cdot\mathbb{E}(\epsilon)\\
&= 0.
\end{aligned}$$

To calculate the second term $\mathbb{E}\left(\overline{V^2}^{3}\right)$ in Equation (4.19), we once again use the GMA assumption where we assume that $\overline{V^2}^2$ is Gaussian and that $\overline{V^2}^3 = \overline{V^2}^2\cdot\overline{V^2}$ is also Gaussian

$$\begin{align}
\mathbb{E}\left(\overline{V^2}^{3}\right) &= \mathbb{E}\left(\overline{V^2}^{2}\cdot\overline{V^2}\right) \nonumber\\
&= \mathbb{E}\left(\overline{V^2}^{2}\right)\cdot\mathbb{E}\left(\overline{V^2}\right) + \mathrm{Cov}\left(\overline{V^2}^{2}, \overline{V^2}\right) \tag{4.20}\\
&= \left(\mu^2_{\overline{V^2}} + \sigma^2_{\overline{V^2}}\right)\mu_{\overline{V^2}} + \mathrm{Cov}\left(\overline{V^2}^{2}, \overline{V^2}\right) \nonumber\\
&= \left(\mu^2_{\overline{V^2}} + \sigma^2_{\overline{V^2}}\right)\mu_{\overline{V^2}} \nonumber\\
&\qquad + \left[\mathrm{Cov}\left(\overline{V^2}, \overline{V^2}\right)\cdot\mathbb{E}\left(\overline{V^2}\right) + \mathrm{Cov}\left(\overline{V^2}, \overline{V^2}\right)\cdot\mathbb{E}\left(\overline{V^2}\right)\right] \tag{4.21}\\
&= \left(\mu^2_{\overline{V^2}} + \sigma^2_{\overline{V^2}}\right) + 2\sigma^2_{\overline{V^2}}\mu_{\overline{V^2}} \nonumber\\
&= \mu^3_{\overline{V^2}} + 3\sigma^2_{\overline{V^2}}\mu_{\overline{V^2}}. \tag{4.22}
\end{align}$$

Combining Equation (4.19) and Equation (4.22) we obtain

$$\begin{aligned}\mathrm{Var}\left(\overline{V^3}\right) &= \sigma^2_{\overline{V^3}} + 15\cdot\mathbb{E}\left(\overline{V^2}^3\right)\\ &= \sigma^2_{\overline{V^3}} + 15\left[\mu^3_{\overline{V^2}} + 3\sigma^2_{\overline{V^2}}\mu_{\overline{V^2}}\right]\\ &= \sigma^2_{\overline{V^3}} + 15\mu^3_{\overline{V^2}} + 45\,\sigma^2_{\overline{V^2}}\,\mu_{\overline{V^2}}. \end{aligned}\tag{4.23}$$

As such, the prior moments of V^3 are given by Equation (4.16) and Equation (4.23):

$$\mu_{V^3} = \mu_{\overline{V^3}}, \tag{4.24}$$

$$\sigma^2_{V^3} = \sigma^2_{\overline{V^3}} + 15\mu^3_{\overline{V^2}} + 45\,\sigma^2_{\overline{V^2}}\,\mu_{\overline{V^2}}. \tag{4.25}$$

To have the moments of $\overline{V^3}$ depend on the covariates, we use the bayesian feedforward neural network setup of Section 3.1 with a three-noded output layer, comprised of $Z^{(0)}$ that models the expected response, $\overline{V^2}$ and $\overline{V^3}$. In this setup, each output node has its own parameter: $Z^{(0)}$ has parameter $\Theta^{(L)}$, $\overline{V^2}$ has its associated parameter $\Theta^{(L)}_{\overline{V^2}}$ and $\overline{V^3}$ has its associated parameter $\Theta^{(L)}_{\overline{V^3}}$. Graphically, the AGSI procedure is presented in Figure 4.2, where it can be viewed as an extention of Figure 3.7 of the AGVI method where we add a third branch to accomodate for $\overline{V^3}$. Recall that $\overline{V^2}$ must be positive and as such is activated using an exponential activation function given by Equation (3.26) to Equation (3.28). On the other hand, $\overline{V^3}$ is not restricted to being positive and as such any valid activation function can be used.

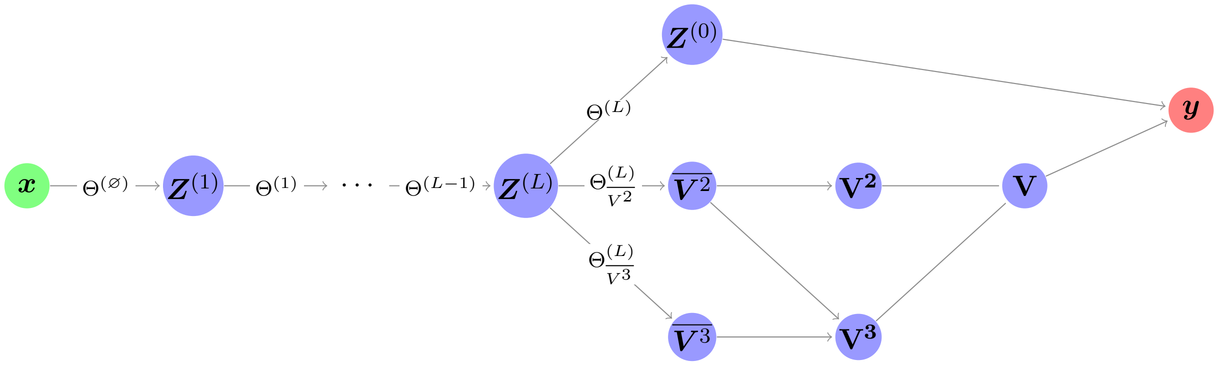


Figure 4.2: Graphical Representation of Forward Propagation with the AGSI Methodology

4.2.2 Posterior Distribution of V^3

Once forward propagation is complete, we obtain the posterior distribution of V^3 and $\overline{V^3}$ alongside the posterior of V^2 and $\overline{V^2}$. The final output of the forward propagation, as can be seen in Figure 4.2, is the vector $\mathbf{h} = \left\{Z^{(0)}, V\right\}^\intercal$. Given the Gaussian assumption, we can obtain the posterior $\mathbf{h}|y$ with Equation (3.13) and Equation (3.14) and get

$$\begin{aligned} \mu_{\mathbf{H}|y} &= \mu_{\mathbf{H}} + \frac{\Sigma_{\mathbf{H}Y}}{\sigma_Y^2}\left(y - \mu_Y\right), \\ \Sigma_{\mathbf{H}|y} &= \Sigma_{\mathbf{H}} - \frac{\Sigma_{\mathbf{H}Y} \cdot \Sigma_{\mathbf{H}Y}^\intercal}{\sigma_Y^2}. \end{aligned} \tag{4.26}$$

We can then use Equation (4.26) to first get $V|y$, and then the posterior for both $\overline{V^2}$ and $\overline{V^3}$. Deka et al. (2024) obtain posterior parameters for V^2 given by

$$\begin{aligned} \mu_{V^2|y} &= \mu_{V|y}^2 + \sigma_{V|y}^2, \\ \sigma_{V^2|y}^2 &= 2\left(\sigma_{V|y}^2\right)^2 + 4\sigma_{V|y}^2 {\cdot} \mu_{V|y}^2, \end{aligned} \tag{4.27}$$

and then posterior moments of $\overline{V^2}$:

$$\begin{aligned} \mu_{\overline{V^2}|y} &= \mu_{\overline{V^2}} + k_{V^2}\left(\mu_{V^2|y} - \mu_{V^2}\right), \\ \sigma_{\overline{V^2}|y}^2 &= \sigma_{\overline{V^2}}^2 + k_{V^2}^2\left(\sigma_{V^2|y}^2 - \sigma_{V^2}^2\right), \end{aligned} \tag{4.28}$$

where $k_{V^2} = \sigma^2_{\overline{V^2}}/\sigma^2_{V^2}$. We now develop the analogous versions of Equation (4.27) and Equation (4.28) for V^3 and $\overline{V^3}$. We start by considering the joint posterior distribution of $V^3, \overline{V^3}$:

$$V^3, \overline{V^3}\Big|y \sim \mathrm{MVN}\Big(\boldsymbol{\mu}_{\overline{V^3},V^3}, \boldsymbol{\Sigma}_{\overline{V^3},V^3}\Big), \tag{4.29}$$

with $\mathbf{\Sigma}_{\overline{V^3},V^3}$ in Equation (4.29) given by

$$\mathbf{\Sigma}_{\overline{V^3},V^3} = \begin{pmatrix} \sigma^2_{\overline{V^3}} & \operatorname{Cov}\left(\overline{V^3}, V^3\right) \\ \operatorname{Cov}\left(V^3, \overline{V^3}\right) & \sigma^2_{V^3} \end{pmatrix} = \begin{pmatrix} \sigma^2_{\overline{V^3}} & \sigma^2_{\overline{V^3}} \\ \sigma^2_{\overline{V^3}} & \sigma^2_{V^3} \end{pmatrix},$$

where we calculate

$$\begin{aligned}
\operatorname{Cov}\left(\overline{V^3}, V^3\right) &= \operatorname{Cov}\left(\overline{V^3}, \overline{V^3} + \sqrt{15}\,\overline{V^2}^{3/2}\cdot\epsilon\right) \\
&= \sigma^2_{\overline{V^3}} + \sqrt{15}\left[\operatorname{Cov}\left(\overline{V^3}, \overline{V^2}^{3/2}\right)\mathbb{E}(\epsilon) + \operatorname{Cov}\left(\overline{V^3}, \epsilon\right)\mathbb{E}\left(\overline{V^2}^{3/2}\epsilon\right)\right] \\
&= \sigma^2_{\overline{V^3}} + \sqrt{15}\left[\operatorname{Cov}\left(\overline{V^3}, \overline{V^2}^{3/2}\right)\cdot 0 + 0\cdot\mathbb{E}\left(\overline{V^2}^{3/2}\epsilon\right)\right] \\
&= \sigma^2_{\overline{V^3}}.
\end{aligned}$$

Next, we obtain $\mu_{V^3|y}$ and $\sigma^2_{V^3|y}$ using Equation (4.10):

$$\begin{aligned}
\mu_{V^3|y} &= \mu^3_{V|y} + 3\mu_{V|y}\cdot\sigma^2_{V|y}, \\
\sigma^2_{V^3|y} &= 15\left(\sigma^2_{V|y}\right)^3 + 36\mu^2_{V|y}\left(\sigma^2_{V|y}\right)^2 + 9\mu^4_{V|y}\sigma^2_{V|y}.
\end{aligned} \tag{4.30}$$

Then, given the properties of conditional Gaussian random variables, that is Equation (3.13) and Equation (3.14), we get the distribution for $\overline{V^3}|V^3, Y$ to be

$$\overline{V^3}\Big|V^3, Y \sim \operatorname{Gaussian}\left(\mu_{\overline{V^3}|V^3}, \sigma^2_{\overline{V^3}|V^3}\right),$$

with

$$\begin{aligned}
\mu_{\overline{V^3}|V^3} &= \mu_{\overline{V^3}} + \frac{\sigma^2_{\overline{V^3}}}{\sigma^2_{V^3}}\left(v^3 - \mu_{V^3}\right) \\
&= \mu_{\overline{V^3}} + k_{V^3}\left(v^3 - \mu_{V^3}\right), \\
\sigma^2_{\overline{V^3}|V^3} &= \sigma^2_{\overline{V^3}} - \frac{\sigma^2_{\overline{V^3}}}{\sigma^2_{V^3}}\cdot\sigma^2_{\overline{V^3}} \\
&= \sigma^2_{\overline{V^3}} - k^2_{V^3}\cdot\sigma^2_{V^3},
\end{aligned} \tag{4.31}$$

where $k_{V^3} = \sigma^2_{\overline{V^3}}/\sigma^2_{V^3}$. To obtain $\overline{V^3}|Y$, we first note that we can write

$$\begin{aligned} f\left(\overline{v^3}, v^3, y\right) &= f\left(v^3, y\right) \cdot f\left(\overline{v^3}|v^3, y\right) \\ &= f\left(v^3|y\right) \cdot f\left(y\right) \cdot f\left(\overline{v^3}|v^3, y\right) \\ \implies f\left(\overline{v^3}, v^3|y\right) &= f\left(v^3|y\right) \cdot f\left(\overline{v^3}|v^3, y\right), \end{aligned}$$

so that we can marginalise v^3 to get

$$f\left(\overline{v^3}|y\right) = \int f\left(v^3|y\right) \cdot f\left(\overline{v^3}|v^3, y\right) \mathrm{d}v^3.$$

Thus, using the Gaussian properties with a random mean and constant variance, the posterior moments for $\overline{V^3}$ are given as

$$\begin{aligned} \mu_{\overline{V^3}|y} &= \mathbb{E}\Big(\mu_{\overline{V^3}} + k_{V^3}\left(v^3 - \mu_{V^3}\right)\Big) \\ &= \mu_{\overline{V^3}} + k_{V^3}\left(\mu_{V^3|y} - \mu_{V^3}\right), \end{aligned} \tag{4.32}$$

$$\begin{aligned} \sigma^2_{\overline{V^3}|y} &= \sigma^2_{\overline{V^3}} - k^2_{V^3}\sigma^2_{V^3} + k^2_{V^3} \cdot \mathrm{Var}\Big(V^3|y\Big) \\ &= \sigma^2_{\overline{V^3}} + k^2_{V^3}\left(\sigma^2_{V^3|y} - \sigma^2_{V^3}\right). \end{aligned} \tag{4.33}$$

We summarise the inference step in the AGSI method graphically in Figure 4.3. The first step is to obtain the posterior of $Z^{(0)}$ and V from Equation (4.26), which is represented by the red lines. We then obtain the posterior distribution of $V^2|y$ and $V^3|y$ with Equation (4.27) and Equation (4.30), represented in blue. We lastly obtain the posterior for $\overline{V^2}$ and $\overline{V^3}$ through means of Equation (4.28), Equation (4.32) and Equation (4.33), which is shown in green.

Once the AGSI procedure is complete, inference on the rest of the network parameters $\boldsymbol{\Theta}$ is done exactly as outlined in Section 3.2.1. The combination of TAGI and AGSI lets us perform analytical inference of the network parameters as well as obtaining heteroscedastic information on the observation error V, by capturing its second and third moments. The

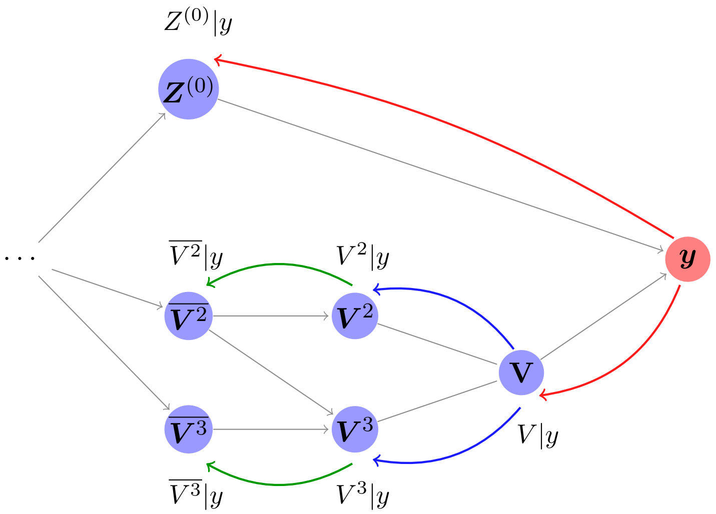


Figure 4.3: Graphical Representation of Posterior Inference for AGSI

blend of TAGI and AGSI together is formally referred to as TAGI-Skewness, or more simply put TAGI-S.

4.3 Inference of GEV parameters

In this section we create the bridge that connects EVT to the TAGI-S BNN by describing how we use the output of TAGI-S to obtain parameter estimates $\hat{\Lambda}$ of the GEV distribution. In broad terms, to go from the output of TAGI-S to Λ, we first denormalise the predicted response and then use the method of moments to calculate $\hat{\Lambda}$.

The framework of TAGI-S, like the TAGI and TAGI-V frameworks, expects us to provide normalised data and will also output normalised values. Thus, assuming an observed data set $\mathcal{D}_{\text{obs}}$ with responses $\mathbf{Y}_{\text{obs}} = \{Y_1, Y_2, \cdots, Y_n\}^\intercal$ with mean $\mu_{\mathbf{Y}_{\text{obs}}} \equiv \boldsymbol{\mu_Y}$ and standard deviatin $\sigma_{\mathbf{Y}_{\text{obs}}} \equiv \sigma_{\mathbf{Y}}$, TAGI-S expects to be trained with the normalised responses $\mathbf{Y}^*_{\text{obs}} = \{Y^*_1, Y^*_2, \ldots, Y^*_n\}^\intercal$ consisting of elements

$$Y^*_i = \frac{Y_i - \mu_{\mathbf{Y}}}{\sigma_{\mathbf{Y}}} \tag{4.34}$$

for $i \in \{1, 2, \ldots, n\}$. Then, given any provided input, TAGI-S will output the first three moments of the predicted value $\hat{Y}^*$, namely $\mathbb{E}\left(\hat{Y}^*\right)$, $\text{Var}\left(\hat{Y^*}\right)$ and $\text{Skew}\left(\hat{Y}^*\right)$. To go from

$\hat{Y}^*$ to $\hat{Y}$, from Equation (4.34) we write $Y = Y^* \cdot \sigma_{\mathbf{Y}} + \mu_{\mathbf{Y}}$. Then we can denormalise as follows:

$$\mathbb{E}\left(\hat{Y}\right) = \mathbb{E}\left(\hat{Y}^*\right) \cdot \sigma_{\mathbf{Y}} + \mu_{\mathbf{Y}}, \tag{4.35}$$

$$\begin{aligned} \mathrm{Var}\left(\hat{Y}\right) &= \mathrm{Var}\left(\hat{Y^*} \cdot \sigma_{\mathbf{Y}} + \mu_{\mathbf{Y}}\right) \\ &= \sigma_{\mathbf{Y}}^2 \cdot \mathrm{Var}\left(\hat{Y^*}\right), \end{aligned} \tag{4.36}$$

$$\begin{aligned} \mathrm{Skew}\left(\hat{Y}\right) &= \mathbb{E}\left(\left[\frac{Y - \mu_{\mathbf{Y}}}{\sigma_{\mathbf{Y}}}\right]^3\right) \\ &= \mathbb{E}\left(\left[\frac{\hat{Y^*} \cdot \sigma_{\mathbf{Y}} + \mu_{\mathbf{Y}} - \mu_{\mathbf{Y}}}{\sigma_{\mathbf{Y}}}\right]^3\right) \\ &= \mathbb{E}\left(\left(\hat{Y}^*\right)^3\right) \\ &= \mathrm{Skew}\left(\hat{Y^*}\right). \end{aligned} \tag{4.37}$$

As we can see from Equation (4.37), skewness does not change when denormalising.

Once we have the first three moments, we can use the method of moments described in Section 2.3.1 to obtain $\hat{\Lambda}^{(\text{TAGI-S})} = \left\{\hat{\mu}^{(\text{TAGI-S})}, \hat{\sigma}^{(\text{TAGI-S})}, \hat{\xi}^{(\text{TAGI-S})}\right\}$. For the remainder of this section, we drop the "TAGI-S" upperscript for better readability and write $\hat{\Lambda}^{(\text{TAGI-S})} \equiv \hat{\Lambda}$. We first obtain $\hat{\xi}$ through Equation (2.13) by numerically solving

$$\mathrm{Skew}\left(\hat{Y}\right) = \mathrm{sgn}(\xi) \cdot \frac{g_3 - 3 g_2 g_1 + 2 g_1^3}{\left(g_2 - g_1^2\right)^{3/2}}, \tag{4.38}$$

then get $\hat{\sigma}$ through solving Equation (2.12),

$$\mathrm{Var}\left(\hat{Y}\right) = \left(g_2 - g_1^2\right) \frac{\sigma^2}{\hat{\xi}^2}, \tag{4.39}$$

and finally obtain $\hat{\mu}$ through solving Equation (2.11):

$$\mathbb{E}\left(\hat{Y}\right) = \mu + \frac{\hat{\sigma}}{\hat{\xi}} \left(g_1 - 1\right), \tag{4.40}$$

where we recall that $g_k = \Gamma\left(1 - k \cdot \xi\right)$ for $k \in \mathbb{N}$.

We thus now have a link between the output of our BNN and the parameter estimation of the GEV distribution.

4.4 Numerical Examples

We now consider some numerical applications of TAGI-S in which we obtain estimates of various simulated GEV distributions. First, in Section 4.4.1 we showcase how TAGI-S can estimate the parameters the GEV distribution. Then in Section 4.4.2 we demonstrate how TAGI-S is able to detect changes in parameter values as a function of input. These examples are of relevance not only to highlight TAGI-S in action, but also give insight as to how we apply TAGI-S later in Section 5.3.

In each of the following examples, we assume the same network context. From the generated data set, we randomly take 70% of the data for training and 20% for validation, by which we form our observation data set $\mathcal{D}_{\text{obs}} = \{\mathcal{D}_{\text{fit}}, \mathcal{D}_{\text{val}}\}$. We take 10% for our test set $\mathcal{D}_{\text{test}}$. We train TAGI-S with $\mathcal{D}_{\text{obs}}$ using one layer ($\mathsf{L} = 1$) consisting of 100 nodes ($\mathsf{A} = 100$) and ReLu activation function. To determine the network parameters and the number of epochs, we use Algorithm 1 with *difference* parameter $\delta = 0.005$ and *patience* parameter $\eta = 10$, meaning that we will stop running new epochs when we get ten consecutive epochs without at least a 0.005 improvement in validation log-likelihood.

4.4.1 First Example

In this first example, we simulate values from a GEV distribution (see Equation (2.10)) with parameters $\mu = 50, \sigma = 1$ and $\xi = -0.10$. We run the same exercise for a simulated data set of a thousand values in Section 4.4.1.1 and a hundred simulated values in Section 4.4.1.2.

4.4.1.1 First Simulation

We simulate a thousand random values with the specified parameters $\mu = 50, \sigma = 1, \xi = -0.10$ in Figure 4.4 within the normalised domain $[-1.729, 1.729]$. Thus, we use 700 data points for training, 200 for validation and 100 for testing.

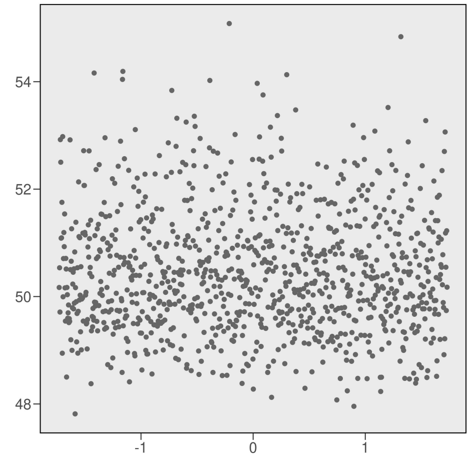
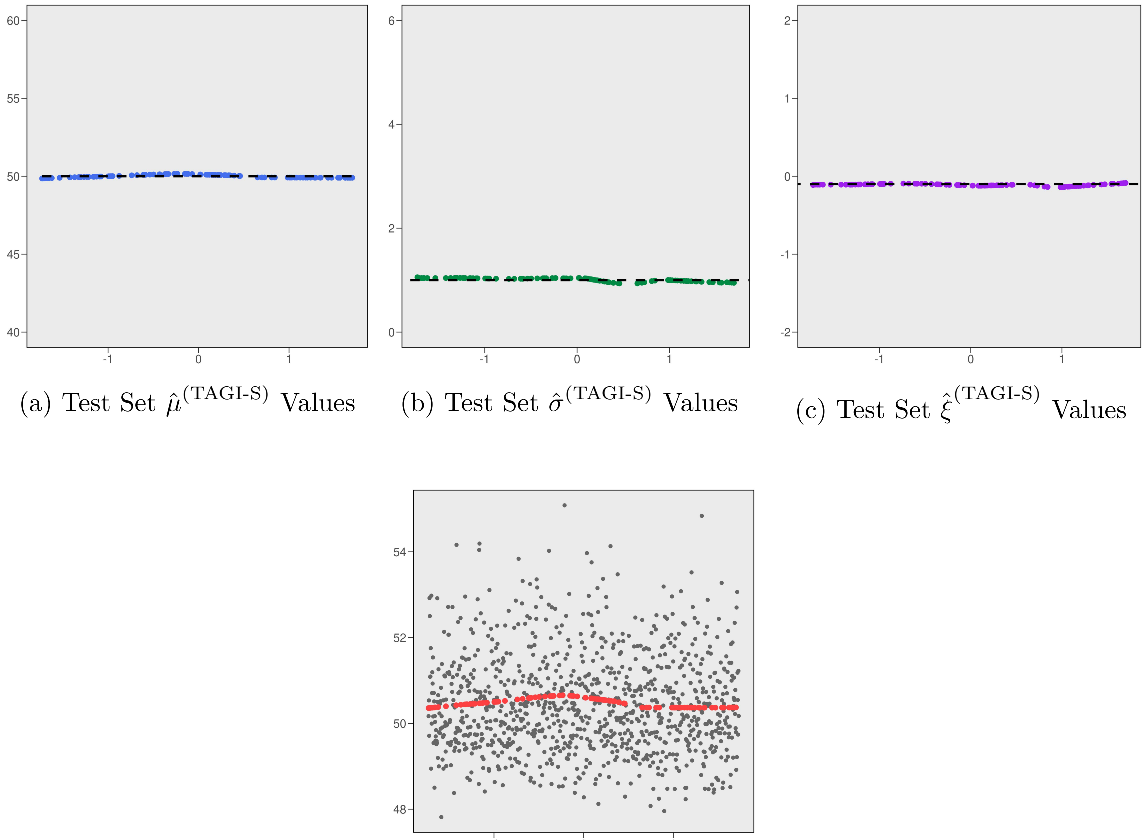


Figure 4.4: 1,000 Random Realisation of a GEV Distributon with $\mu = 50$, $\sigma = 1$ and $\xi = -0.10$.

Training TAGI-S on $\mathcal{D}_{\text{obs}}$ leads to five epochs used. With the network trained, from the test set $\mathcal{D}_{\text{test}}$ with a given value $x \in [-1.729, 1.729]$ and predicted value $\hat{y}$, we denormalise according to Equation (4.35), Equation (4.36) and Equation (4.37) to then obtain its associated GEV parameters, which we denote $\hat{\Lambda}^{\text{(TAGI-S)}} = \left\{\hat{\mu}^{\text{(TAGI-S)}}, \hat{\sigma}^{\text{(TAGI-S)}}, \hat{\xi}^{\text{(TAGI-S)}}\right\}$, by means of Equation (4.40), Equation (4.39) and Equation (4.38). Since we simulate from a GEV distribution with constant parameter values, we expect $\hat{\Lambda}$ to be close to Λ for every test value.

In Figure 4.5 we plot the values of $\hat{y}_i$ and $\hat{\Lambda}_i^{\text{(TAGI-S)}}$ for each test value with $i = 1, \ldots, 100$. In Figure 4.5a, Figure 4.5b and Figure 4.5c, the dashed lines represent the true parameter values, that is $\mu = 50, \sigma = 1$ and $\xi = -0.10$ respectively. Figure 4.5d presents the predicted values of the test set with the simulated data in gray.



(a) Test Set $\hat{\mu}^{(\text{TAGI-S})}$ Values

(b) Test Set $\hat{\sigma}^{(\text{TAGI-S})}$ Values

(c) Test Set $\hat{\xi}^{(\text{TAGI-S})}$ Values

(d) Predicted Output with $\mathcal{D}_{\text{obs}}$ in gray

Figure 4.5: 1,000 Simulated GEV values with $\Lambda = \{50, 1, -0.10\}$ with $\hat{\Lambda}^{(\text{TAGI-S})}$ and Predicted Outputs

We also record the Mean-Squared Error (MSE) of Λ with the test set,

$$\text{MSE} = \mathbb{E}\left(\left[\hat{\Lambda}_i^{(\text{TAGI-S})} - \Lambda_i\right]^2\right),$$

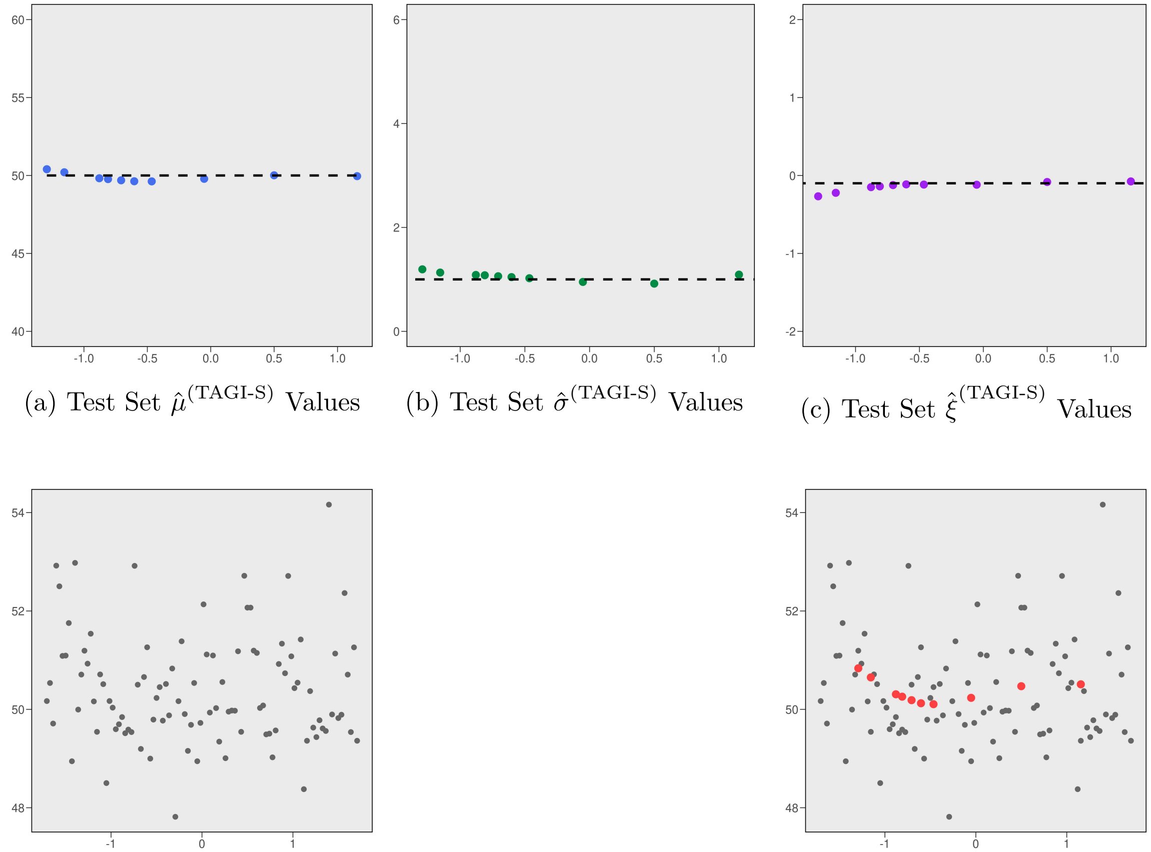
in Table 4.1, where we can effectively see that TAGI-S is able to recognise the correct GEV distribution parameters.

μ	σ	ξ
0.00287	0.0022	0.0006

Table 4.1: MSE of Test Values for First TAGI-S Example (First Simulation)

4.4.1.2 Second Simulation

We next consider the exact same simulation as Section 4.4.1.1, this time with $n = 100$ simulated values instead of a thousand. Using Algorithm 1 leads to the use of 19 epochs, which is 14 more epochs compared to five epochs needed in the previous example. This makes sense given the fact that since there is much less data than before, it takes more epochs to learn the embedded pattern. We show the corresponding verison of Figure 4.5 in Figure 4.6.



(a) Test Set $\hat{\mu}^{(\text{TAGI-S})}$ Values

(b) Test Set $\hat{\sigma}^{(\text{TAGI-S})}$ Values

(c) Test Set $\hat{\xi}^{(\text{TAGI-S})}$ Values

(d) Simulated Data

(e) Test Set Predicted Output

Figure 4.6: 100 Simulated GEV values with $\Lambda = \{50, 1, -0.10\}$ with $\hat{\Lambda}^{(\text{TAGI-S})}$ and Predicted Outputs

We also provide the MSE values of the test set in Table 4.2.

As expected, we see that given the smaller data set of a hundred simulated data, TAGI-S performs less well than its counterpart with a thousand simulated data. We however still see that TAGI-S is still able to adequately capture the true GEV parameters.

μ	σ	ξ
0.07137	0.00920	0.0049

Table 4.2: MSE of Test Values for First TAGI-S Example (Second Simulation)

4.4.2 Second Example

In this second example, we simulate GEV values in which the parameters are allowed to change with respect to the input. That is, we consider $\Lambda(x) = \{\mu(x), \sigma(x), \xi(x)\}$. The objective is to evaluate how well TAGI-S can infer non-constant parameters. Here, we take

$$\Lambda(x) = \begin{cases} \mu(x) = 50 + x, \\ \sigma(x) = 2 + 0.6x, \\ \xi(x) = -0.10 \end{cases}$$

for $x \in [-1.729, 1.729]$. We simulate 500 data points $Y(x) \sim \text{GEV}\,(\Lambda(x))$ and plot the simulated data in Figure 4.7. The observation set $\mathcal{D}_{\text{obs}}$ consists of 350 training values $\mathcal{D}_{\text{fit}}$, 100 validation points $\mathcal{D}_{\text{val}}$ and 50 test points $\mathcal{D}_{\text{test}}$.

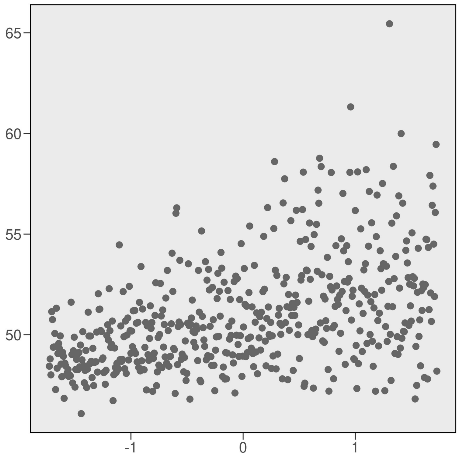


Figure 4.7: 500 Random Realisation of a GEV Distributon with $\mu = \mu(x)$, $\sigma = \sigma(x)$ and $\xi = -0.10$.

We see in Figure 4.7 that early values of $Y(x)$ clearly do not follow the same trend as later values of $Y(x)$. As such, instead of trying to fit one fixed set of parameters $\hat{\Lambda}$ for all inputs, being able to dynamically adapt to the data and infer non-constant GEV parameters

$\hat{\Lambda}(x)$ as a function of the inputs is of great benefit.

Applying TAGI-S to the above data set according to the methodology described in Section 4.4 leads to the following predicted outputs (from the test set):

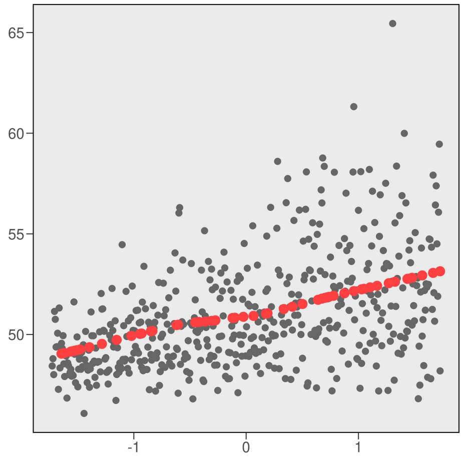
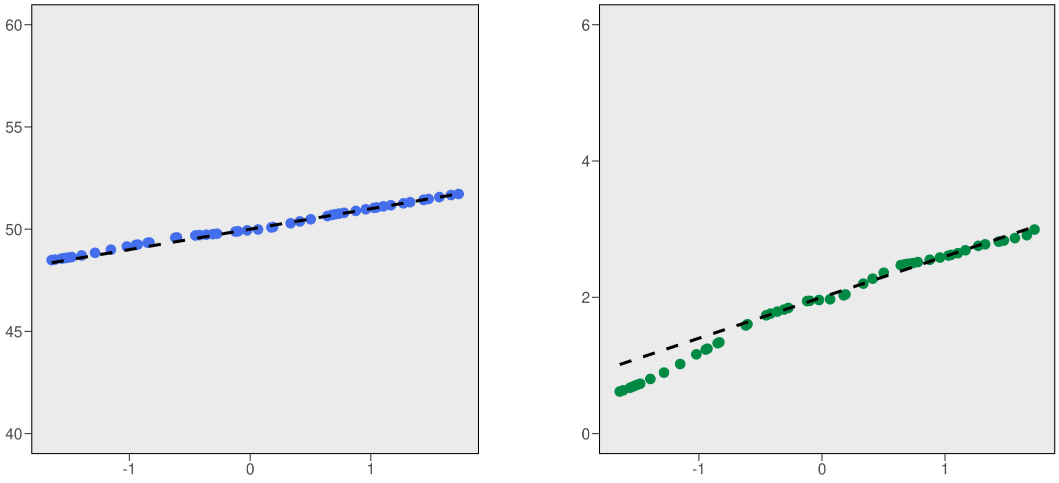


Figure 4.8: Predicted Values of the Test Set

In Figure 4.9 we plot the values of $\hat{\mu}^{(\text{TAGI-S})}$ and $\hat{\sigma}^{(\text{TAGI-S})}$, where the dashed lines represent the functions $\mu(x) = 50 + x$ and $\sigma(x) = 2 + 0.6x$ respectively.



(a) Test Set $\hat{\mu}^{(\text{TAGI-S})}$ Values

(b) Test Set $\hat{\sigma}^{(\text{TAGI-S})}$ Values

Figure 4.9: $\hat{\Lambda}^{(\text{TAGI-S})}$ and Expected Outputs of Simulated GEV Values

The test set MSE values for $\hat{\Lambda}^{(\text{TAGI-S})}$ are given in Table 4.3. We also add the MSE values when running the example with only $n = 100$ simulated data instead of $n = 500$.

From Table 4.3, we see that TAGI-S is able to infer dynamic values of Λ. We do notice

n	μ	σ	ξ
500	0.0096	0.0327	0.0233
100	0.2029	0.1430	0.0429

Table 4.3: MSE of Test Values for Second TAGI-S Example

that the quality of the modelling decreases when the number of available data decreases, which is to be expected when working with neural networks.

For the sake of argument, if we assume that the data set of Figure 4.7 represents annual maximal rainfall data for a given location across multiple years t, we see from Figure 4.9 that the parameter estimates obtained with TAGI-S are different at different times, that is $\hat{\Lambda}_{t_1}^{(\text{TAGI-S})} \neq \hat{\Lambda}_{t_2}^{(\text{TAGI-S})}$ for $t_1 \neq t_2$, since for example $\hat{\sigma}(t_1)$ will de different from $\hat{\sigma}(t_2)$.

Using a standard appoximation technique (maximum likelihood, method of moments, etc.) would lead to a fix set of parameters $\hat{\Lambda}$ for the whole data set, which would likely lead to erronous GEV parameters. As such, the above example showcases a distinct advantage of using TAGI-S: the parameter estimates $\hat{\Lambda}_t^{(\text{TAGI-S})}$ can adapt to the variable nature of the GEV parameter for a given year t.

In Appendix B.2, we consider more unorthodox forms of $\Lambda(x)$ and evaluate how TAGI-S performs.

Chapter 5

Application: Spatial Interpolation

In the previous chapter, we showed how TAGI-S can be applied to infer the parameters of a GEV distribution. The aim of this chapter is to apply TAGI-S to interpolate the GEV parameters Λ over a given surface $\mathcal{S}$.

In Section 5.1 we show how to perform spatial interpolation of GEV parameters with TAGI-S. Then, in Section 5.2 we show how TAGI-S performs on various simulated data sets. In Section 5.3 we consider annual maximal rainfall data in Eastern Canada, spatially interpolate with TAGI-S alongside two other methods and compare the goodness-of-fit of all methods across the interpolation region.

5.1 Spatial Interpolation with TAGI-S

In the following, let $\mathcal{S}$ be the total surface considered, with a set of n_{obs} observed locations $\{s_1, s_2, \ldots, s_{n_{\text{obs}}}\}$. To perform interpolation on the surface $\mathcal{S}$, we propose to feed the entire set of observed values from each of the n_{obs} locations along with the spatial location (lattitude, longitude) and other possible covariates. The training of TAGI-S then follows as in previous chapters, where we feed to TAGI-S an observation data set $\mathcal{D}_{\text{obs}} = \{\mathcal{D}_{\text{fit}}, \mathcal{D}_{\text{val}}\}$ of size n_{obs} consisting of n_{fit} training values $\mathcal{D}_{\text{fit}} = \{\mathbf{X}_{\text{fit}}, \mathbf{Y}_{\text{fit}}\}$ and n_{val} validation values $\mathcal{D}_{\text{val}} = \{\mathbf{X}_{\text{val}}, \mathbf{Y}_{\text{val}}\}$ of covariates and responses that span across all known locations of the surface $\mathcal{S}$. We set no prior assumption on the relationship between the stations and let TAGI-S infer itself how each station is related.

To help illustrate how we perform interpolation, assume that we have n_{fit} locations $s_1, \dots, s_{n_{\text{fit}}}$ with N data entries each, with longitude and lattitude coordinates $\texttt{long}_{s_i}$ $\texttt{lat}_{s_i}$ respectively, for $i \in \{1, 2, \dots, n_{\text{fit}}\}$. For each station s_i, denote by $y_{(i,j)}$ the j-th observed value.

Then, the training set $\mathcal{D}_{\text{fit}}$ for TAGI-S will be the matrices

$$\mathbf{X}_{\text{fit}} = \begin{bmatrix} \texttt{long}_1 & \texttt{lat}_1 \\ \texttt{long}_1 & \texttt{lat}_1 \\ \vdots & \vdots \\ \texttt{long}_1 & \texttt{lat}_1 \\ \texttt{long}_2 & \texttt{lat}_2 \\ \vdots & \vdots \\ \texttt{long}_n & \texttt{lat}_n \end{bmatrix} \quad \text{and} \quad \mathbf{Y}_{\text{fit}} = \begin{bmatrix} y_{(1,1)} \\ y_{(1,2)} \\ \vdots \\ y_{(1,N)} \\ y_{(2,1)} \\ \vdots \\ y_{(n,N)} \end{bmatrix},$$

and we likewise conisder $\mathcal{D}_{\text{val}} = \{\mathbf{X}_{\text{val}}, \mathbf{Y}_{\text{val}}\}$ for a given number n_{val} of validation stations. We train TAGI-S with TAGI-S $(\mathcal{D}_{\text{obs}})$. Once the network has been trained, we can input any valid input $\hat{\mathbf{X}}$ for the covariate and obtain the predicted output $\hat{Y} = \text{TAGI-S}\left(\hat{\mathbf{X}}\right)$.

Additionally, we note that TAGI-S can be fed with any number of covariates: we are not restricted to only input longitude and lattitude to the network to perform spatial interpolation. As such, if another covariate is believed to be of potential use, we can simply add said covariate to TAGI-S.

5.2 Simulations

For each of the simulation cases presented here, we consider a surface $\mathcal{S} = [-1, 1] \times [-1, 1]$ upon which we randomly generate 50 stations with coordinates $s_i = (x_{s_i}, y_{s_i})$ for $n \in \{1, 2, \dots, 50\}$. We take the x coordinates to be the longitude ($\texttt{long}$) and the y coordinates to be lattitude ($\texttt{lat}$) in Equation (5.14). We then simulate values from different GEV distributions at each of the stations. We feed the simulated values with their coordinates to TAGI-S for training.

To measure the performance of predictions, we take two sets of stations; an observation set for training and a testing set, which we denote with the indexes I_{obs} and I_{test} respectively. Out of the 50 stations, we take $n_{\text{obs}} = 40$ stations for training and $n_{\text{test}} = 10$ for testing.

To measure the goodness-of-fit, we will calculate the 95th quantile at each validation station $s \in I_{\text{test}}$ and compare its value to the 95th quantile value that is calculated with the true GEV parameter values Λ_s of that location. Since the choice of training/validation stations can influence the performance metrics, we run 100 combinations of training/validation sets and aggregate the goodness-of-fit metrics, which we now present.

For testing station $s \in I_{\text{test}}$, we denote q_s the true 95th quantile and $\tilde{q}_s$ its interpolated value. We set to measure how different q_s and $\tilde{q}_s$ are. To do so, we use the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE):

$$\text{RMSE} = \sqrt{\frac{1}{n_{\text{test}}} \sum_{i \in I_{\text{test}}} \left(q_{s_i} - \tilde{q}_{s_i}\right)^2}, \tag{5.1}$$

$$\text{MAE} = \frac{1}{n_{\text{test}}} \sum_{i \in I_{\text{test}}} \left|q_{s_i} - \tilde{q}_{s_i}\right|. \tag{5.2}$$

Next, To compare the performance of TAGI-S, we use two known interpolation methods, namely the Inverse Weighted Distance (IDW) and Polynomial Regression methods.

The IDW interpolation technique, which stems from Burrough (1986), is a simple and frequently used method. The rationale behind the IDW method is that a given point to be estimated is influenced the most by nearby points. As such, each observed location available is attributed a weight that is inversely proportional to the distance of the point to be interpolated.

For a location $s \in \mathcal{S}$ that is not part of the observed locations $\{s_1, s_2, \ldots, s_n\}$, we denote by d_{s_i} the euclidean distance (based on longitude and lattitude) between the interpolating station s_i and interpolated station s, $d_{s_i} = \sqrt{\left(x_s - x_{s_i}\right)^2 + \left(y_s - y_{s_i}\right)^2}$.

Then, denoting φ to be any one of GEV parameter values Λ,[1] for $\varphi \in \Lambda$ the interpolated value $\tilde{\varphi}(s)$ is given by

[1] That is, φ can be either the location parameter μ, scale parameter σ or the shape parameter ξ.

$$\tilde{\varphi}(s) = \sum_{i=1}^{n} w_i \cdot \varphi(s_i), \tag{5.3}$$

with weights defined as

$$w_i = \frac{d_{s_i}^{-1}}{\sum_{i=1}^{n} d_{s_i}^{-1}}.$$

As such, for a given location s we obtain $\tilde{\mu}(s), \tilde{\sigma}(s)$ and $\tilde{\xi}(s)$ by applying Equation (5.3) three times; once for each parameter.

The second method, Polynomial Regression, assumes that each parameter $\varphi \in \Lambda$ is linear with respect to the coordinates of the location. More precisely, we model any one of the GEV interpolated parameter as

$$\tilde{\varphi}(s) = \tilde{\beta}_0 + \tilde{\beta}_1 \cdot x_s + \tilde{\beta}_2 \cdot y_s + \tilde{\beta}_3 \cdot x_s^{\iota_x} \cdot y_s^{\iota_y}, \tag{5.4}$$

where $x_s^{\iota_x} \cdot y_s^{\iota_y}$ is the *interaction* term of degrees ι_x and ι_y. The values of $\tilde{\beta}_i$, $i \in \{0, 1, 2, 3\}$, are obtained through Ordinary Least Squares estimation (Rencher and Christensen (2012)). For each interpolated $\tilde{\varphi} \in \tilde{\Lambda}$, we consider all possible interactions up to degree three ($\iota_x, \iota_y \in \{0, 1, 2, 3\}$) and choose the model with the best AIC criterion (Akaike (1974)) to perform interpolation.

Thus, we compare TAGI-S to IDW and Polynomial Regression by means of Equation (5.1) and Equation (5.2). Since the choice of training/validation stations can impact the goodness-of-fit values, we choose 100 random combinations of fitting/validation stations and aggregate the RMSE and MAE scores obtained.

5.2.1 First Simulation

For the first simulation, we consider 50 randomly generated stations in the surface $\mathcal{S} = [-1, 1] \times [-1, 1]$. We generate GEV distributions which respect the following parameter function: for $s = (x_s, y_s) \in \mathcal{S}$ we let $s \equiv (x_s, y_s) \sim \text{GEV}\left(\Lambda(s)\right)$ with

$$\Lambda(s) \equiv \Lambda(x_s, y_s) = \begin{cases} \mu(x_s, y_s) & = \begin{cases} 60, & x_s, y_s < 0 \\ 45 & \text{o.w.} \end{cases} \\ \sigma(x_s, y_s) & = 1.5 \\ \xi(x_s, y_s) & = 0.20. \end{cases}$$

That is, all stations that are in the lower left quadrant have location parameter $\mu = 60$ and all other stations have location parameter $\mu = 45$. The 95th quantile q_{s_i} for each of the fifty stations s_i are graphed in Figure 5.1.

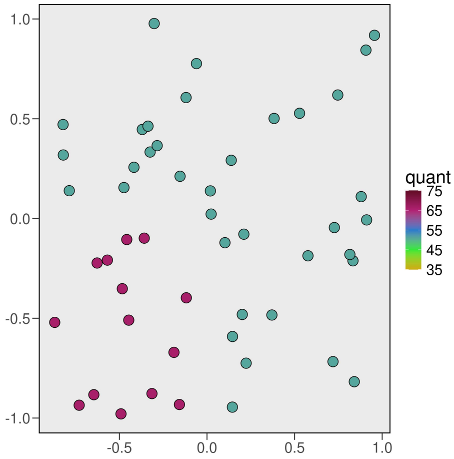


Figure 5.1: 95th Quantiles for the 50 Simulated Stations (First Simulation)

For each of the 50 stations, we generate $n_{\text{gen}} = 100$ random values of its respective GEV distribution. We randomly select 80% of the stations ($n_{\text{obs}} = 40$ stations) for fitting and 20% ($n_{\text{test}} = 10$ stations) for testing. We then feed TAGI-S all of the $n_{\text{obs}} \cdot n_{\text{gen}} = 40 \cdot 100 = 4,000$ observation GEV values with the x and y coordinate of each observed station. Within the the observed $4,000$ points fed to TAGI-S, we take 70% of these points for fitting purposes and 30% for validation. Thus, we let TAGI-S not only learn $\Lambda(s)$ for the fitting stations $s \in I_{\text{obs}}$, but also interpolate for any coordinate $(x_s, y_s) \in \mathcal{S}$.

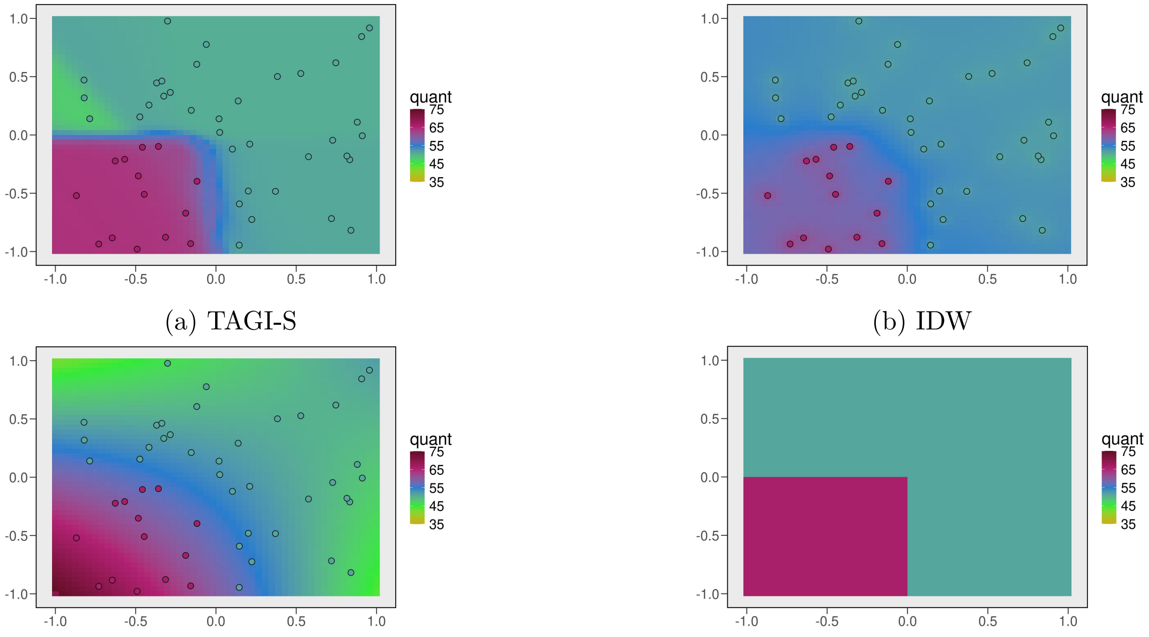
With the known values of Λ at each fitting station we perform the IDW and Regression procedures described by Equation (5.3) and Equation (5.4) respectively. We note that these methods take the values of Λ_s for granted at each known station s; they do not need to be inferred as with TAGI-S.

We repeat this process a hundred times and record the median (md.) and standard deviation (std.) values of the RMSE and MAE values in Table 5.1.

		TAGI-S	IDW	Regression
RMSE	md.	2.45	4.88	4.13
	std.	(0.98)	(1.00)	(0.91)
MAE	md.	2.02	3.77	3.18
	std.	(0.72)	(0.75)	(0.77)

Table 5.1: Error Measures of Testing Stations for the First Simulation

We also uniformly choose 2,500 points in $\mathcal{S}$ and apply interpolation according to each of the three methods and calculate the 95th quantile at each interpolation point. We then build the following quantile maps in Figure 5.2. We overlay the points of Figure 5.1 for an easy graphical reference and plot the true quantile map in Figure 5.2d.



(a) TAGI-S

(b) IDW

(c) Regression

(d) True Quantile

Figure 5.2: Calculated 95th Quantile Maps for First Simulation

By looking at Table 5.1 with the additional visual representation of Figure 5.2, we see that out of the three methods, TAGI-S is best able to detect the cluster in the lower-left quadrant. The IDW method in the lower-left quadrant is influenced too heavily by other stations, and the regression method forces a linear trend in the data.

5.2.2 Second Simulation

For the second simulation, we again consider 50 random locations. We let each location $s \in \mathcal{S}$ follow GEV distributions that decrease in mean as we get further away from the origin $(0,0)$. That is, we let $s \sim \text{GEV}\left(\Lambda(s)\right)$ with

$$\Lambda(s) \equiv \Lambda\left(x_s, y_s\right) = \begin{cases} \mu(x_s, y_s) & = 60 - 10|x_s| - 10|y_s|, \\ \sigma(x_s, y_s) & = 2, \\ \xi(x_s, y_s) & = -0.10, \end{cases} \tag{5.5}$$

for $x_s, y_s \in [-1, 1]$.

We plot the simulated stations and their 95th quantiles in Figure 5.3. As we can see, the higher quantile values are located in the center of the graph with maximal quantile value 63.63. The further we stray away from $(0,0)$, the lower the 95th quantiles become. The smallest quantile value is 46.51.

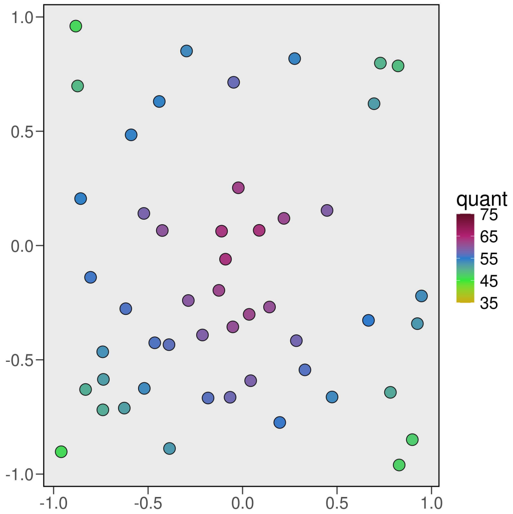


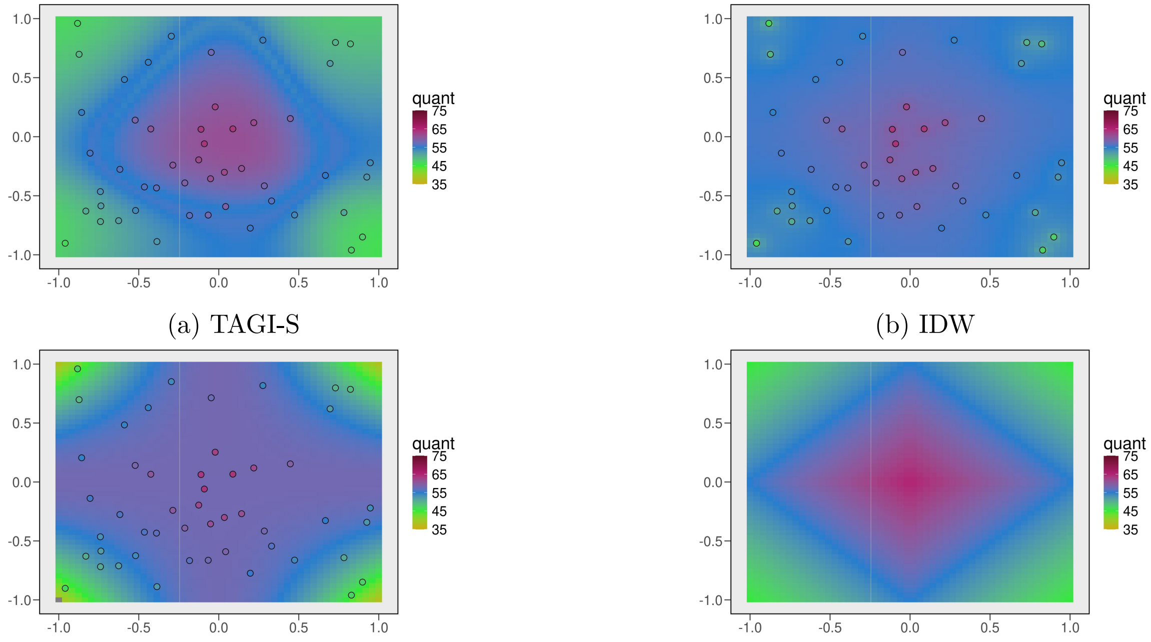
Figure 5.3: 95th Quantiles for the 50 Simulated Stations (Simulation Two)

As per the previous simulation of Section 5.2.1, for each of the 50 locations s we randomly generate $n_{\text{gen}} = 100$ values from the related GEV distribution with parameters $\Lambda(s)$ defined in Equation (5.5). We choose $n_{\text{fit}} = 40$ stations for fitting and $n_{\text{test}} = 10$ for testing and calulcate the RMSE and MAE values. We perform this 100 times and report the aggregated error measures in Table 5.2.

		TAGI-S	IDW	Regression
RMSE	md.	1.52	3.60	2.88
	std.	(0.28)	(0.80)	(0.57)
MAE	md.	0.82	3.04	2.51
	std.	(0.51)	(0.73)	(0.46)

Table 5.2: Error Measures of Testing Stations for the Second Simulation

We also plot the 95th quantile map for each method in Figure 5.4, where we consider 2,500 uniformly distributed points in $[-1, 1] \times [-1, 1]$. We overlay the generated quantile points of Figure 5.3 as well for visual reference and depict the true quantile map in Figure 5.4d.



(a) TAGI-S

(b) IDW

(c) Regression

(d) True Quantile

Figure 5.4: 95th Quantile Maps for Second Simulation

From Table 5.2, we see that TAGI-S performs the best, followed by the Polynomial Regression model and then the IDW method. Looking at Figure 5.4, we see that the regression models creates a smooth/linear map, whereas TAGI-S seems to be more flexible in how it interpolates over the surface.

Thus, we can conclude that TAGI-S is adequately able to interpolate GEV parameters across a given surface for the given parameter setups $\Lambda(s)$ we considered.

5.3 Spatial Interpolation in Eastern Canada

In this chapter, we model the joint distribution of annual maximal rainfall of 197 considered stations across the Canadian provinces of Ontario, Quebec, New Brunswick, Nova Scotia, Newfoundland and Prince Edward Island through spatial interpolation with various methods.

In Section 5.3.1 we present the dataset and the processing manipulations applied to it. Then, in Section 5.3.2, we present how we model extreme precipiation with TAGI-S and other methods used for comparison. In Section 5.3.3 we apply TAGI-S to each station individually to obtain GEV parameter estimates at each station. Then, in Section 5.3.4 we interpolate the GEV parameters across Eastern Canada using TAGI-S and two other methods. We present the results in Section 5.3.5.

5.3.1 Dataset

The raw dataset consists of yearly annual maximal rainfall data, measured in millimeters (mm), collected at 334 different stations across the eastern part of Canada spanning the years 1905 to 2017. The data is provided publicly by Canada (2019). The 334 stations cover longitudes -94.23 to -52.54 and lattitudes 41.57 to 61.30. The available information at each station s consists of the longitude ($\texttt{long}_s$), lattitude ($\texttt{lat}_s$), altitude ($\texttt{alt}_s$) and yearly maximal rainfall which we denote $y_{(s,t)}$, where t represents the year. We also include the 75% quantile precipiations at each station, which we denote $\texttt{Q75}_s$.

We desire to focus on stations that are of actuarial interest, that is the ones that present prominent risk exposure. As such, we narrow down the original 334 stations to the 197 stations that are covered by the longitude range $[-79.23, -54.57]$ and the lattitude range $[43.71, 50.24]$. These stations are plotted in Figure 5.5, labeled by the abbreviations for the name of each province: NB for New Brunswick, NL for Newfoundland, NS for Nova Scotia, ON for Ontario, PE for Prince Edward Island and QC for Quebec.

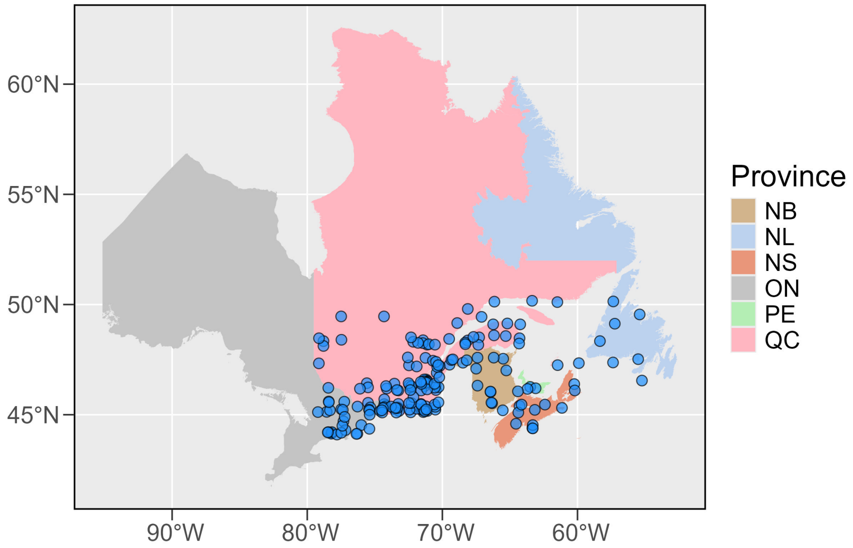


Figure 5.5: 197 Considered Stations in Eastern Canada

In Figure 5.6, we showcase for each year (1905 to 2017) the number of stations with available data. The dashed line represents the reference point of 197 stations, from which we can see that no station has data points for every year and face the issue of having an incomplete data set. In Section 5.3.2 we will use data augmentation to remedy this issue and be able to consider full data series.

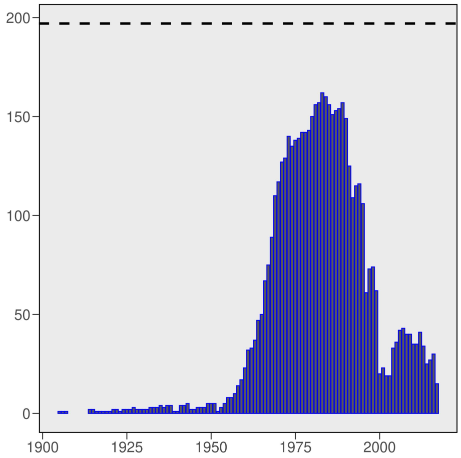


Figure 5.6: Number of Stations with Data per Year

With further analysis, we observe that for the years 1905 to 1952, only eight stations provide observed data, two of which contain no data point for years after 1953. The `Quebec` station has available data from years 1914 to 1943 and the `Saint John` station contains years 1924 to 1950[2]. The `Ottawa Cda Rcs` station has 6 points before 1953 and 50 after, `Halifax` has 12 data points before 1953 and 34 after, `Moncton` has 6 before 1953 and 43 after, `Kingston Pumping Station` has 17 points before 1953 and 46 after, `La Cave` has 1 point before 1953 and 9 after, and finally `Montreal Pierre Elliot Trudeau Intl.` airport station has 10 points before and 51 after 1953.

To minimise the number of pseudo observations we will have to generate in Section 5.3.2, we remove the `Quebec` and `Saint John` stations and truncate the other six stations by removing data observed prior to 1953; we remove all data from years 1905 to 1952. This creates an official data set of 195 stations with 65 years spanning from 1953 to 2017.

After this data processing, we are left with the following number of stations per province, shown in Table 5.3.

Province	NB	NL	NS	ON	PE	QC
Number of Stations	12	8	12	29	3	131

Table 5.3: Number of Stations Per Province

Before explaining the methodology, we lastly note that since the data is presented in yearly maximal blocks, the block maxima approach is a clear choice for the estimation of extremes in this situation.

5.3.2 Methodology

In this section, we first cover some notation then tackle the methodologies used to model the maximal annual rainfall across the considered region. We apply TAGI-S to each station individually in Section 5.3.3, then proceed to describe all the interpolation methods used in Section 5.3.4. We present the results of the spatial interpolation in Section 5.3.5.

As discussed in the previous section, the original dataset is not complete; for any given station there are multiple years that do not have data entries. To have a concordant data set and to be able to interpret the results we obtain across the region of interest, we employ

[2] We note that there are secondary stations named “Quebec” and “Saint John” that contain valid data post-1953 and which presumably replace these stations.

data-augmentation.

For a given station $s_i \in \{1, 2, \dots, n\}$ for $n = 195$ and year $t \in \{1, 2, \dots, T\}$ with $T = 65$, we let

$$\mathbf{y}_{s_i} = \left\{ y_{(s_i,1)}, y_{(s_i,2)}, \dots, y_{(s_i,T)} \right\}^\intercal$$

be the time series of annual maximal rainfall. The entirety of the data set can be writen as a $n \times T$ matrix

$$\mathbf{Y} = \begin{pmatrix} y_{(s_1,1)} & y_{(s_1,2)} & \cdots & y_{(s_1,T)} \\ y_{(s_2,1)} & y_{(s_2,2)} & \cdots & y_{(s_2,T)} \\ \vdots & \vdots & \ddots & \vdots \\ y_{(s_n,1)} & y_{(s_n,2)} & \cdots & y_{(s_n,T)} \end{pmatrix},$$

where the entry $y_{(s_i,t)}$ is the annual maximal rainfall of station i at time t, for $i \in \{1, 2, \dots, n\}$ and $t \in \{1, 2, \dots, T\}$. Given the incompleteness of our dataset, multiple elements of $\mathbf{Y}$ are empty.

We propose to fill the missing years of a given station s_i by using the L-moments described in Section 2.3.2. Since we have a low amount of observed data for many stations (see Figure 5.6), the use of L-moments is adequate given its efficiency for parameter estimation with small sample sizes (Hosking and Wallis (1997)). Thus, for every station s_i, we use the available data points in $\mathbf{y}_{s_i}$ to obtain the L-moments estimator for the GEV parameters $\hat{\Lambda}_{s_i}^{(\mathrm{LM})}$, as given by Equation (2.24).

Given the L-moments estimate of a station $\hat{\Lambda}_{s_i}^{(\mathrm{LM})}$, we build pseudo observations for the missing years using the quantile function Equation (2.14) with $\hat{\Lambda}_{s_i}^{(\mathrm{LM})}$ and obtain the data augmented vector $\mathbf{y}^*_{s_i}$ (we denote $\mathbf{Y}^*$ the data augmented matrix version of $\mathbf{Y}$). This provides us with a complete data set that we can now work on.

5.3.3 Estimation of Marginal Stations

With the data-augmented dataset, we can use TAGI-S exactly as described in Section 4.4 and obtain GEV parameter estimates for a given station s_i, which we denote $\hat{\Lambda}_{s_i}^{(\text{TAGI-S})}$. We refer to the present method as the *pointwise TAGI-S* method, since we estimate GEV parameters for each station individually. We will use pointwise TAGI-S as a lower bound for the error metrics in Section 5.3.4.

To run pointwise TAGI-S, we can specifiy any legitimate network setting that we want (number of layers L, nodes per layer A, activation function $\phi(\cdot)$, etc.) Here, we consider the same network setup for each station, in which we take two layers ($\mathsf{L} = 2$) of 100 nodes each ($\mathsf{A} = 100$) with tanh activation function $\phi(\cdot) = \tanh(\cdot)$, patience criteria $\eta = 10$ and difference parameters $\delta = 0.005$ (see Algorithm 1). At each station, we take 80% of the data to train and the rest to validate.

To measure the performance of TAGI-S, for each station we compare the predicted quantiles from TAGI-S to the empirical quantiles of the data-augmented station at multiple levels. For $M = 30$, we let $p_m = {}^{m-0.5}/_{M}$ for $m \in \{1, 2, \ldots, M\}$ be the associated probability. Let $q_{s_i}^{(m)}$ be the $(1 - p_m)$ empirical quantile of station i and $\hat{q}_{s_i}^{(m)}$ its predicted value according to the output of the neural network. We collect the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Bias, defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n \cdot M} \sum_{i=1}^{n} \sum_{m=1}^{M} \left(q_{s_i}^{(m)} - \hat{q}_{s_i}^{(m)} \right)^2}, \tag{5.6}$$

$$\text{MAE} = \frac{1}{n \cdot M} \sum_{i=1}^{n} \sum_{m=1}^{M} \left| q_{s_i}^{(m)} - \hat{q}_{s_i}^{(m)} \right|, \tag{5.7}$$

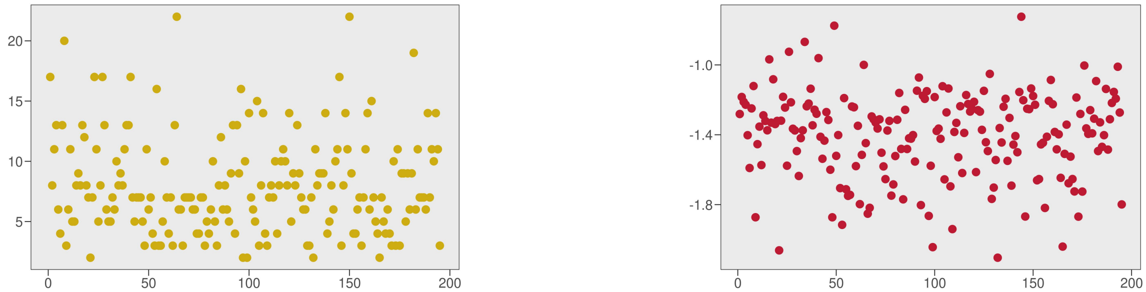
$$\text{Bias} = \frac{1}{n \cdot M} \sum_{i=1}^{n} \sum_{m=1}^{M} q_{s_i}^{(m)} - \hat{q}_{s_i}^{(m)}. \tag{5.8}$$

We thus obtain the median (md.) and standard deviation (std.) of the aggregated error measures for the 195 stations in Table 5.4.

	RMSE	MAE	Bias
md.	3.25	2.29	-0.13
std.	(0.83)	(0.52)	(0.14)

Table 5.4: Goodness-of-Fit Measure for Pointwise TAGI-S

Next, we plot the number of epochs and validation likelihood value for each of the 195 stations in Figure 5.7 that we obtain by running pointwise TAGI-S. The average number of epochs is 7.99, with smallest epoch number being 2, the highest 22, and average validation log-likelihood of -1.402.



(a) Number of Considered Epochs at Each Station

(b) Optimal Likelihood Value at Each Station

Figure 5.7: Epochs and Likelihood Values for Each Station

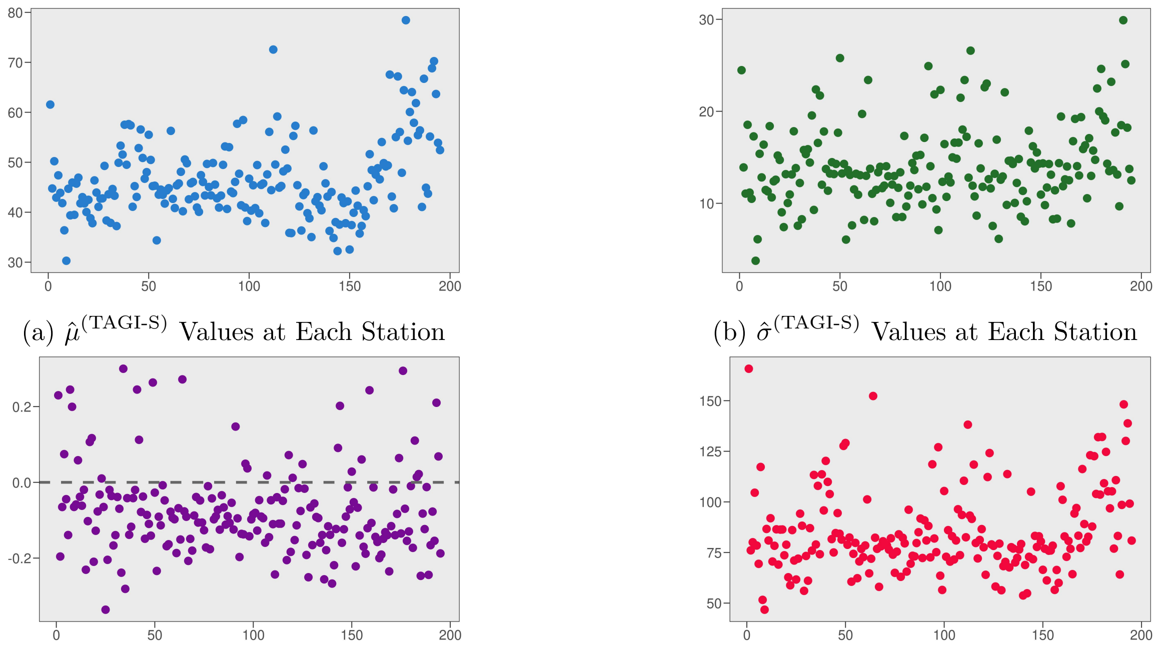
Next, in Figure 5.8 we plot the obtained parameter estimates $\hat{\Lambda}_{s_i}^{(\text{TAGI-S})}$ for each station s_i along with the predicted 95th quantile. We record an average location value of $\hat{\mu}^{(\text{TAGI-S})} = 46.62$, average scale value $\hat{\sigma}^{(\text{TAGI-S})} = 14.13$ and average shape value $\hat{\xi}^{(\text{TAGI-S})} = -0.076$. We notice that the values of $\hat{\xi}^{(\text{TAGI-S})}$, although for the majority being negative, do have a certain discrepancy present. The average predicted 95th quantile is 85.24.

Although pointwise TAGI-S is not an interpolation method, it is still of interest as it represents the lower bound of the error we can obtain when doing spatial interpolation with TAGI-S in Section 5.3.4. That is, assuming perfect spatial modelling across the interpolation surface with TAGI-S, it is impossible to obtain better measures of errors than the ones in Table 5.4.

5.3.4 Spatial Interpolation

Now, we turn our attention to spatial interpolation with TAGI-S and how it performs compared to other known methods. Given the 195 stations, our objective is to model the GEV parameters $\Lambda(s)$ given a location s. The methods that we will consider are *Pointwise L-moments* (Pointwise LM), *Polynomial Regression*, a modified version of *IDW* and *TAGI-S*. Before describing each method, we first define how we will measure performance.

The notation used here will be very similar to the notation of Section 5.2. Assume any of the previously named models is being used. Out of the 195 total stations, we use 80%



(a) $\hat{\mu}^{(\text{TAGI-S})}$ Values at Each Station

(b) $\hat{\sigma}^{(\text{TAGI-S})}$ Values at Each Station

(c) $\hat{\xi}^{(\text{TAGI-S})}$ Values at Each Station

(d) Predicted 95th
Quantile at Each Station

Figure 5.8: $\hat{\Lambda}_{s_i}^{(\text{TAGI-S})}$ Estimates and Expected Outputs for Each Station

($n_{\text{obs}} = 156$) stations for training and the remaining 20% ($n_{\text{test}} = 39$) stations for testing. The variables $I_{\text{obs}}, n_{\text{obs}}, I_{\text{test}}$ and n_{test} are used to denote in order the set of observation stations, the number of observation stations, the set of testing stations and finally the number of stations used for testing.

We perform error measurement on both the training and testing set. Let $q_{s_i}^{(1)}, q_{s_i}^{(2)}, \ldots, q_{s_i}^{(M)}$ be the $M = 30$ empirical quantiles associated at location s_i, with associated probability

$$p_m = \frac{m - 0.5}{M}.$$

Let $\tilde{q}_{s_i}^{(M)}$ be the *interpolated* $(1 - p_m)$ quantile at location s_i. Next, we write $n^* = n_{\text{obs}}$ when analysing training stations or $n^* = n_{\text{test}}$ when dealing with the testing stations, and I^* to be either I_{obs} or I_{test} likewise. We then consider the three following goodness-of-fit measures, which are analogous to Equation (5.6), Equation (5.7) and Equation (5.8) used earlier.

$$\text{RMSE} = \sqrt{\frac{1}{n^* \cdot M} \sum_{s_i \in I^*} \sum_{m=1}^{M} \left(q_i^{(m)} - \tilde{q}_i^{(m)} \right)^2}, \tag{5.9}$$

$$\text{MAE} = \frac{1}{n^* \cdot M} \sum_{s_i \in I^*} \sum_{m=1}^{M} \left| q_i^{(m)} - \tilde{q}_i^{(m)} \right|, \tag{5.10}$$

$$\text{Bias} = \frac{1}{n^* \cdot M} \sum_{s_i \in I^*} \sum_{m=1}^{M} q_i^{(m)} - \tilde{q}_i^{(m)}. \tag{5.11}$$

Since the choice of fitting/testing stations can influence the measures above, we will consider 100 randomly chosen combinations of I_{obs} and I_{test} and aggregate the RMSE, MAE and Bias scores.

To better understand the influence of additional information on the results that we obtain, we shall consider two sets of covariates that will be used with each of the proposed methods. The first set of covariates consists of the longitude (`long`), lattitude (`lat`) and altitude (`alt`) of each station. The second set of covariates will also include longitude, lattitude and altitude, but with and an additional covariate being the 75% quantile of precipitation, which we denote `Q75`. Thus, each method will be applied to both $\{\texttt{long}, \texttt{lat}, \texttt{alt}\}$ and $\{\texttt{long}, \texttt{lat}, \texttt{alt}, \texttt{Q75}\}$. We will be able to see if the addition of `Q75` brings meaningful improvements to predictions or not.

Pointwise LM. The first method we use is in fact quite simple. In describing the data augmentation procedure in Section 5.3.2, we use the L-moments estimates $\hat{\Lambda}_{s_i}^{(\text{LM})}$ for each station s_i. We can thus measure how well the L-moments fit the data by comparing the empirical quantiles of each station with the ones driven by the L-moments, where the predicted quantiles are obtained directly through Equation (2.14) by using parameters $\hat{\Lambda}_{s_i}^{(\text{LM})}$ at each station. We refer to this as the pointwise LM model, as the predictions stem from these said L-moments.

Like the pointwise TAGI-S method, the pointwise LM method is not a spatial interpolation method. Since the interpolation frameworks that follow (Regression, IDW) stem from the augmented dataset built with the L-moments, the pointwise LM model represents the lower bound of the errors we obtain: the Regression and IDW methods will never record better goodness-of-fit scores. The pointwise LM framework is thus used as a benchmark for other models instead of a direct comparison method with TAGI-S.

Polynomial Regression. The second method, and the first with which we will concretely compare TAGI-S interpolation with, is the Polynomial Regression (or just Regression) model. Recalling the notation from Section 5.2, we let φ denote either one of μ, σ or ξ. For a location s, we model the interpolated parmater $\tilde{\varphi}(s)$ as

$$\tilde{\varphi}(s) = \tilde{\beta}_0 + \sum_{i=1}^{r} \tilde{\beta}_i \cdot x_s^{(r)}, \tag{5.12}$$

where $x_s^{(r)}$ are the r covariates we described earlier (`lat`, `long`, `lat` and `Q75` when applicable). This modelling follows a classical regression setting without the error term that is usually included. For any of the two sets of covariates, we take all possible combinations between the specific covariates with highest degree of interaction of three. We then select the model with the best AIC value as the model to be used for interpolation.

IDW. The third method used is a modified version of the Inverse Distance Weighted method presented in Section 5.2, which adds a gradient correction (Nalder and Wein (1998)). Using this modification permits us to include other covariates that are not longitude and lattitude. The IDW technique provides good results when there are many points disitributed uniformly across the interpolating surface. However, if the surface is too large with not enough observations, the smoothing will be taken to too far of an extreme and remove local variations. On the flip side, too small a surface will produce closely similar interpolated values with little to no gain in information (Burrough and McDonnell (1998)).

Denoting by $x_s^{(1)}, \ldots, x_s^{(r)}$ the r recorded covariates for each station s, the interpolated parameter $\tilde{\varphi}(s)$ is defined as

$$\tilde{\varphi}(s) = \sum_{i=1}^{n} w_{s_i} \left[\hat{\varphi}(s_i) + \beta_1 \left(x_s^{(1)} - x_{s_i}^{(1)} \right) + \cdots + \beta_r \left(x_s^{(r)} - x_{s_i}^{(r)} \right) \right], \tag{5.13}$$

where $\hat{\varphi}(s_i)$ is the L-moments estimate of φ at station s_i and where

$$w_{s_i} = \frac{d_{s_i}^{-1}}{\sum_{i=1}^{n} d_{s_i}^{-1}},$$

with d_{s_i} being the euclidean distance between the interpolated station s with the interpolating station s_i. We still keep the general philosophy of IDW by which we give more importance to stations that are nearby, but add additional information ($\texttt{alt}$, $\texttt{Q75}$) to the interpolation procedure. The values of $\beta_1, \ldots, \beta_r$ in Equation (5.13) are obtained by minimizing the quantity

$$\sum_{i=1}^{n} \left(\hat{\varphi}(s_i) - \tilde{\varphi}(s_{-i})\right)^2 ,$$

where $\tilde{\varphi}(s_{-i})$ is the interpolated value of φ at s_i when it is not considered in Equation (5.13).

In our context, when working with the set of three covariates $\{\texttt{long}, \texttt{lat}, \texttt{alt}\}$ we will write Equation (5.13) as

$$\tilde{\varphi}(s) = \sum_{i=1}^{n} w_{s_i} \left[\hat{\varphi}(s_i) + \beta_{\texttt{alt}} \left(\texttt{alt}_s - \texttt{alt}_{s_i}\right)\right]$$

and when working with $\{\texttt{long}, \texttt{lat}, \texttt{alt}, \texttt{Q75}\}$ we will write Equation (5.13) as

$$\tilde{\varphi}(s) = \sum_{i=1}^{n} w_{s_i} \left[\hat{\varphi}(s_i) + \beta_{\texttt{alt}} \left(\texttt{alt}_s - \texttt{alt}_{s_i}\right) + \beta_{\texttt{Q75}} \left(\texttt{Q75}_s - \texttt{Q75}_{s_i}\right)\right] .$$

TAGI-S. Finally, to interpolate at a station s with TAGI-S, the methodology is the same as in Section 5.2, where we also include the altitude and 75th quantile when applicable. As such, we will train TAGI-S by feeding the entirity of the training stations' covariates. Assuming the usage of the second set of covariates, we feed to TAGI-S the matrices of covariates and data-augmented maximal rainfall data

$$\mathbf{X}_{\text{obs}} = \begin{bmatrix} \texttt{long}_{s_1} & \texttt{lat}_{s_1} & \texttt{alt}_{s_1} & \texttt{Q75}_{s_1} \\ \vdots & \vdots & \vdots & \vdots \\ \texttt{long}_{s_1} & \texttt{lat}_{s_1} & \texttt{alt}_{s_1} & \texttt{Q75}_{s_1} \\ \texttt{long}_{s_2} & \texttt{lat}_{s_2} & \texttt{alt}_{s_2} & \texttt{Q75}_{s_2} \\ \vdots & \vdots & \vdots & \vdots \\ \texttt{long}_{s_{n_{\text{obs}}}} & \texttt{lat}_{s_{n_{\text{obs}}}} & \texttt{alt}_{s_{n_{\text{obs}}}} & \texttt{Q75}_{s_{n_{\text{obs}}}} \end{bmatrix} \quad \text{and} \quad \mathbf{Y}_{\text{obs}} = \begin{bmatrix} y^*_{(s_1,1)} \\ y^*_{(s_1,2)} \\ \vdots \\ y^*_{(s_1,65)} \\ y^*_{(s_2,1)} \\ \vdots \\ y^*_{(s_{n_{\text{obs}}},65)} \end{bmatrix} \tag{5.14}$$

to obtain the trained network TAGI-S($\mathcal{D}_{\text{obs}}$), where $\mathcal{D}_{\text{obs}} = \{\mathbf{X}_{\text{obs}}, \mathbf{Y}_{\text{obs}}\}$[3]. Then, for a given location s that we desire to interpolate, we provide $\{\texttt{long}_s, \texttt{lat}_s, \texttt{alt}_s, \texttt{Q75}_s\}$ to the trained network to obtain the interpolated set of parameters $\tilde{\Lambda}(s) = \left\{\tilde{\mu}(s), \tilde{\sigma}(s), \tilde{\xi}(s)\right\}$. We note that unlike the other methods, where we need to repeat the given procedure three times (once for μ, σ and ξ), TAGI-S will directly give us the interpolated set of parameters.

The pointwise TAGI-S method shown in Section 5.3.3 represent the "best case" scenario when interpolating across the whole surface with TAGI-S, in which we are able to *exactly* recuperate each marginal parameter set $\hat{\Lambda}_{s_i}^{(\text{TAGI-S})}$ obtained by considering each station individually instead of all stations together. Thus, both the pointwise LM method and pointwise TAGI-S methods present lower bounds of errors for each interpolation method (pointwise LM bounds the regression and IDW methods, pointwse TAGI-S bounds the TAGI-S method).

5.3.5 Results

We now present the results of applying the described methodologies. In Table 5.5, we show the median (md.) and standard deviation (std.) of the aggregated goodness-of-fit scores Equation (5.9), Equation (5.10) and Equation (5.11) for the models that consider the first set of covarites $\{\texttt{long}, \texttt{lat}, \texttt{alt}\}$. In Table 5.6 we present the analogous performance metrics using the second set of covariates, with `Q75` added. The pointwise methods are seperated since they represent the lower bound of errors we can obtain: pointwise TAGI-S bounds the TAGI-S models and pointwise LM bounds the Regression and IDW methods.

We provide boxplots for each of the RMSE, MAE and Bias metrics of Table 5.5 and Table 5.6 in Figure 5.9, Figure 5.10 and Figure 5.11 respectively. Each model is represented by a distinct color (red for TAGI-S, blue for Regression and purple for IDW). The lighter tone is used when modelling with three covarites and a darker tone is used when considering four covariates.

From Table 5.5 and Table 5.6, we immediately notice that adding `Q75` increases performance overall. The best *training* scores come from the IDW method, which is expected

[3] Withing the n_{obs} stations fed to TAGI-S for training, we use 70% of data points for training and 30% for validation.

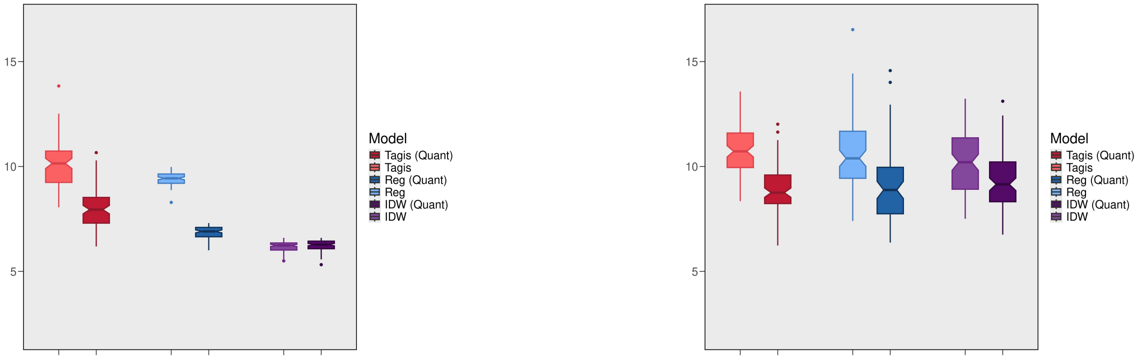
		Pointwise		Training			Testing		
Metric		TAGI-S	LM	TAGI-S	Reg	IDW	TAGI-S	Reg	IDW
RMSE	md.	3.25	3.37	10.14	9.43	6.23	10.72	10.55	10.20
	std.	(0.83)	(0.20)	(1.05)	(0.30)	(0.25)	(1.29)	(4.70)	(1.44)
MAE	md.	2.29	2.24	6.50	5.74	2.69	6.84	6.77	6.49
	std.	(0.52)	(0.04)	(0.75)	(0.17)	(0.05)	(0.89)	(1.14)	(0.79)
Bias	md.	-0.13	0.18	0.49	0.67	0.12	0.60	0.04	0.49
	std.	(0.14)	(0.03)	(0.52)	(0.07)	(0.08)	(1.59)	(1.40)	(1.30)

Table 5.5: Goodness-of-Fit Measures of Models With Three Covariates `long`, `lat` and `alt` for 100 Combinations of Fitting/Testing Stations

		Pointwise		Training			Testing		
Metric		TAGI-S	LM	TAGI-S	Reg	IDW	TAGI-S	Reg	IDW
RMSE	md.	3.25	3.37	7.94	6.92	6.27	8.76	9.22	9.15
	std.	(0.83)	(0.20)	(1.06)	(0.31)	(0.25)	(1.14)	(4.10)	(1.29)
MAE	md.	2.29	2.24	4.98	3.79	2.70	5.64	5.92	5.90
	std.	(0.52)	(0.04)	(0.73)	(0.10)	(0.05)	(0.81)	(0.99)	(0.65)
Bias	md.	-0.13	0.18	0.41	0.50	0.12	0.55	0.51	0.62
	std.	(0.14)	(0.03)	(0.49)	(0.08)	(0.08)	(1.21)	(1.26)	(1.27)

Table 5.6: Goodness-of-Fit Measures of Models With Four Covariates `long`, `lat`, `alt` and `Q75` for 100 Combinations of Fitting/Testing Stations

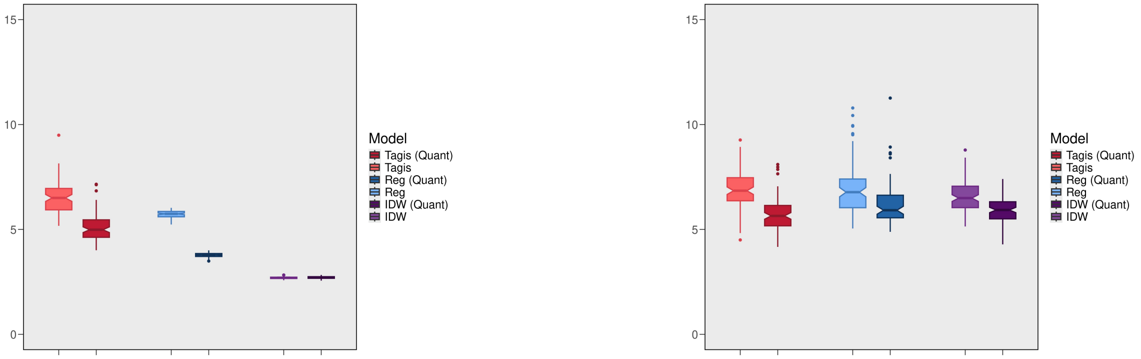
since it is an exact interpolation method (that is, it $\tilde{\varphi}(s_i) = \hat{\varphi}(s_i)$) up to the gradient correction. We observe fitting errors since we add the gradient correction (comparing Equation (5.3) and Equation (5.13)). For *testing* stations, we see that when considering three covariates, IDW performs the best overall. When looking at models with `Q75` added, TAGI-S performs the best and presents the lowest RMSE variance. Additionally, TAGI-S is the most stable across fitting and testing scores, as the Regression and IDW models show quick deterioration with respect to goodness-of-fit scores when going from fitting to testing data sets.



(a) Fitting RMSE

(b) Testing RMSE

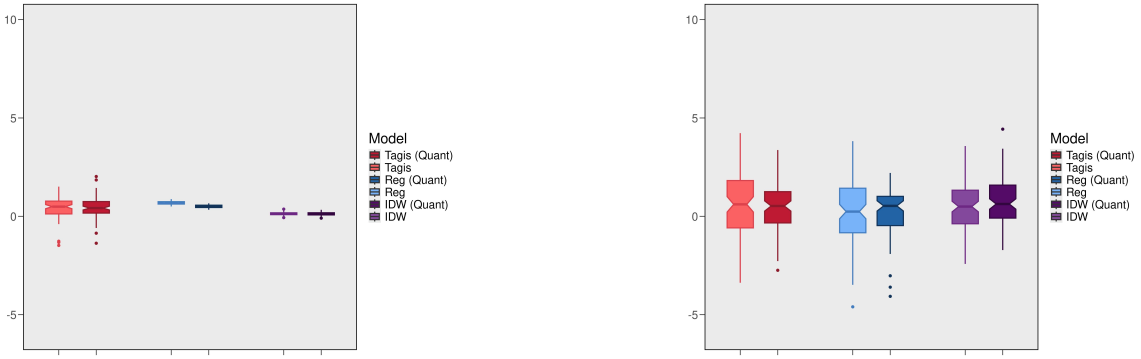
Figure 5.9: RMSE Measures of All Considered Models



(a) Fitting MAE

(b) Testing MAE

Figure 5.10: MAE Measures of All Considered Models



(a) Fitting Bias

(b) Testing Bias

Figure 5.11: Bias Measures of All Considered Models

Chapter 6

Conclusion

This thesis shows how BNNs can be effectively integrated with EVT. We built an extension of the TAGI neural network to be able to obtain the first three moments of a predicted value, with which we were able to obtain the parameters of the GEV distribution. We then applied our developped framework to successfully model and interpolate extreme rainfall in Eastern Canada, obtaining better performance than the comparing methodologies.

An advantage of the methodology developped here is the fact that it is data-driven. Indeed, TAGI-S is able to discern nonlinear trends (see Section 5.2) purely from the data itself, without any imposed assumption. Another advantage is the ease with which we can add/remove covariates to the network: the model is very flexible in terms of considered covariates.

Furthermore, this thesis highlights the importance of high-quality data and the need for continuous improvement in data collection. The GEV distribution being derived from an asymptotic assumption, compounded with the fact that neural networks desire large amounts of data, presents a downside to the present methodology. Having to use data augmentation to build complete time series not only affects the marginal estimation of parameters but also how the joint modelling occurs.

It is evident that advancements in meteorological data collection technologies will undoubtedly enhance the performance of such models in the future. Had the data set available been complete and with more available years, no data augmentation would have been needed, TAGI-S would have more data per station to perform inference and the issues raised in the previous paragraph would be greatly diminished.

With regards to possible extensions to the present work, a promising future path is the addition of other covariates to modelling with TAGI-S. For example, an ongoing possibility we are currently studying is the inclusion of a time component in TAGI-S. As such, being able to include time could permit us to extend spatial interpolations of Section 5.1 to *spatiotemporal* interpolation.

Another possible area of interest is with the estimation the parameters of the GPD of Equation (2.17). We can describe the distribution in terms of a location, scale and shape parameter. From there, we can follow the same setup as was done for estimating the GEV distribution: obtain the first thee moments with TAGI-S and use the method of moments to obtain the distribution parameters.

By the same line of reasoning, one is not constrained to keep TAGI-S to the branch of EVT. Theoretically, any distribution which has its first three moments defined can be estimated using TAGI-S. This drastically opens the horizons of how and where TAGI-S can be applied.

References

Akaike, H. (1974). Methods of multivariate analysis. *IEEE Transactions on Automatic Control*, 19:716–723.

Balkema, A. A. and de Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, 2(5):792–804.

Barbería, L., Amaro, J., Aran, M., and Llasat, M. C. (2014). The role of different factors related to social impact of heavy rain events: considerations about the intensity thresholds in densely populated areas. *Natural Hazards and Earth System Sciences*, 14(7):1843–1852.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural network. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France. PMLR.

Burrough, P. and McDonnell, R. (1998). Principle of geographic information systems.

Burrough, P. A. (1986). *Principle of Geographic Information Systems for Land Resources Assessement*. Oxford University Press.

Canada, G. (2019). Engineering clime datasets. https://climate.weather.gc.ca/prods_servs/engineering_e.html.

Canada, G. (2024). Canada's record-breaking wildfires in 2023: A fiery wake-up call. https://natural-resources.canada.ca/simply-science/canadas-record-breaking-wildfires-2023-fiery-wake-call/25303.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values.* Springer London, London.

de Haan, L. and Ferreira, A. (2010). *Extreme Value Theory: An Introduction (Springer Series in Operations Research and Financial Engineering).* Springer, 1st edition. edition.

Deka, B. (2022). *Analytical Bayesian Parameter Inference for Probabilistic Models with Engineering Applications.* Phd thesis, Polytechique Montéal, Montéal, QC.

Deka, B., Nguyen, L. H., and Goulet, J.-A. (2024). Analytically tractable heteroscedastic uncertainty quantification in bayesian neural networks for regression tasks. *Neurocomputing*, 572:127183.

Embrechts, P., Kluppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance.* Springer-Verlag.

Fisher, R. A. and Tippett, L. H. (1928). Limiting forms of the frequency distributionof the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24:180–290.

Fortunato, M., Blundell, C., and Vinyals, O. (2019). Bayesian recurrent neural networks.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, 44(3):423–453.

Goulet, J.-A. (2020). *Probabilistic Machine Learning for Civil Engineers.* MIT Press.

Goulet, J. A., Nguyen, L. H., and Amiri, S. (2021). Tractable approximate gaussian inference for bayesian neural networks.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R. (1979). Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, 15(5):1049–1054.

Gumbel, E. J. (1958). *Statistics of Extremes.* Columbia University Press, New York Chichester, West Sussex.

Hernández-Lobato, J. M. and Adams, R. P. (2015). Probabilistic backpropagation for scalable learning of bayesian neural networks.

Hosking, J. R. M. (1986). The theory of probability weighted moments. *Research Report RC12210, IBM Research.*

Hosking, J. R. M. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society*, 52(1):105–124.

Hosking, J. R. M. and Wallis, J. R. (1997). Regional frequency analysis: An approach based on l-moments. *Cambridge University Press.*

Hosking, J. R. M., Wallis, J. R., and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27(3).

Johnson, R. A. and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis.* Prentice Hall, 6. ed edition.

Katz, R. W. and Brown, B. G. (1992). Extreme events in a changing climate: Variability is more important than averages. *Climatic Change*, 21(3):289–302.

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources*, 25(8):1287–1304.

Leadbetter, M. R., Lindgren, G., and Rootzen, H. (1983). *Extremes and Related Properties of Random Sequences and Processes.* Springer.

Leibig, C., Allken, V., Ayhan, M., Berens, P., and Wahl, S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7(1):17816.

McNeil, A. J. (1997). Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin*, 27(1):117–137.

Muraleedharan, G., Soares, C., and Lucas, C. (2009). Characteristic and moment generating functions of generalised extreme value distribution (gev). *Nova Science Publishers.*

Nalder, I. A. and Wein, R. W. (1998). Spatial interpolation of climatic normals: test of a new method in the canadian boreal forest. *Agricultural and Forest Meteorology*, 92(4):211–225.

Neal, R. (1996). *Bayesian Learning for Neural Networks.* Springer New York, NY.

Perkins, S. E., Alexander, L. V., and Nairn, J. R. (2012). Increasing frequency, intensity and duration of observed global heatwaves and warm spells. *Geophysical Research Letters*, 39(20).

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1):119–131.

Poon, S.-H., Rockinger, M., and Tawn, J. (2004). Extreme value dependence in financial markets: Diagnostics, models, and financial implications. *The Review of Financial Studies*, 17(2):581–610.

Rauch, H. E., Striebel, C. T., and Tung, F. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA*, pages 1445–1450.

Rencher, A. C. and Christensen, W. F. (2012). *Methods of Multivariate Analysis.* Wiley Series in Probability and Statistics.

Sharma, S., Sharma, S., and Athaiya, A. (2020). Activation functions in neural networks. *International Journal of Engineering Applied Sciences and Technology*, 4(12):310–316.

Silva Lomba, J. and Fraga Alves, M. (2020). L-moments for automatic threshold selection in extreme value analysis. *Stochastic Environmental Research and Risk Assessment*, 34:465–491.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9:1135–1151.

Tabari, H. (2020). Climate change impact on flood and extreme precipitation increases with water availability. *Scientific Reports*, 10(1):2045–2322.

Wu, A., Nowozin, S., Meeds, E., Turner, R. E., Hernández-Lobato, J. M., and Gaunt, A. L. (2019). Deterministic variational inference for robust bayesian neural networks.

Appendix A

Feedforward Neural Network

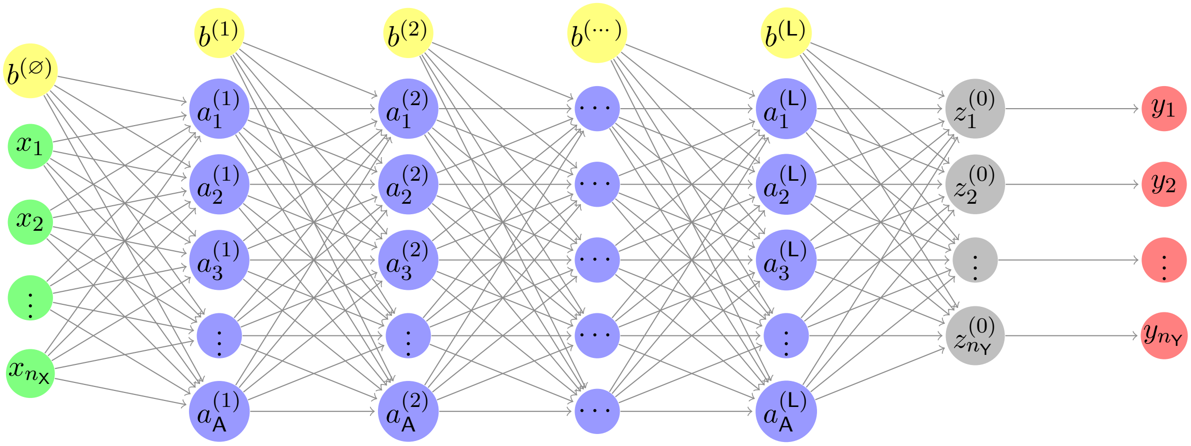


Figure A.1: Expanded Graphical Representation of the Feedforward Neural Network

The above expanded representation can also be written in a compact version,

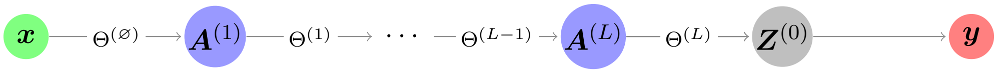


Figure A.2: Compact Graphical Representation of the Feedforward Neural Network

Appendix B

TAGI-S

B.1 Recursive Formula for Moments of the Gaussian Distribution

Let $X \sim \text{Gaussian}\big(\mu, \sigma^2\big)$. Here we show that for $k \in \mathbb{N}$,

$$\mathbb{E}\Big(X^k\Big) = \sum_{i=0}^{k} {}_kC_i \, \mu^i \sigma^{k-i} \mathbb{E}\Big(Z^{k-i}\Big), \tag{B.1}$$

where ${}_nC_k = n! \big/ (n-k)!k!$ and $Z \sim \text{Gaussian}(0,1)$.

First, we show that

$$\mathbb{E}\Big(Z^k\Big) = 0, \quad k \text{ odd} \tag{B.2}$$

$$\mathbb{E}\Big(Z^k\Big) = (k-1)!!, \quad k \text{ even} \tag{B.3}$$

where

$$n!! = \prod_{k=0}^{\lceil n/2 \rceil - 1} (n-2k) = n \cdot (n-2) \cdot (n-4) \ldots$$

and $1!! = 1$.

We note that we have $\mathbb{E}(Z) = 0$ and $\mathbb{E}\big(Z^2\big) = 1$. Then, we use Stein's Lemma (see Stein (1981)), which states that for a Gaussian distribution X with mean μ and variance σ^2 along with a differentiable function g, we have that

$$\mathbb{E}(g(X)\,(X-\mu)) = \sigma^2 \cdot \mathbb{E}(g'(X))\,. \tag{B.4}$$

Taking a standard normal variable with $\mu = 0$ and $\sigma^2 = 1$, along with the function $g(x) = x^k$ for $k \in \mathbb{N}$, we can re-write Equation (B.4) as

$$\mathbb{E}\Big(Z^{k+1}\Big) = k \cdot \mathbb{E}\Big(Z^{k-1}\Big)\,. \tag{B.5}$$

For example, $\mathbb{E}(Z^3) = 2 \cdot \mathbb{E}(Z)\,0$ and $\mathbb{E}(Z^4) = 3 \cdot \mathbb{E}(Z^2) = 3$. It is then apparent from Equation (B.5) that any odd moment of Z will be zero. For k being even, induction shows that $\mathbb{E}\Big(Z^k\Big) = (k-1)!!$. For the base case ($k = 2$), $\mathbb{E}(Z^2) = 1!! = 1$. If we assume that $\mathbb{E}\Big(Z^k\Big) = (k-1)!!$, then we use Stein's lemma Equation (B.4)) to obtain

$$\begin{aligned}
\mathbb{E}\Big(Z^{k+2}\Big) &= (k+1) \cdot \mathbb{E}\Big(Z^k\Big) \\
&= (k+1) \cdot (k-1)!! \\
&= (k+1) \cdot (k-1) \cdot (k-3) \cdots \\
&= (k+1)!!.
\end{aligned}$$

Thus, Equation (B.2) and Equation (B.3) are proved.

Next, we can write X in terms of standard normal variables as $X = \mu + \sigma \cdot Z$, where $X \sim \text{Gaussian}(\mu, \sigma^2)$ and $Z \sim \text{Gaussian}(0, 1)$. We can then use the Binomial Theorem to write for $k \in \mathbb{N}$:

$$\begin{aligned}
\mathbb{E}\Big(X^k\Big) &= \mathbb{E}\Big((\mu + \sigma Z)^k\Big) \\
&= \mathbb{E}\left(\sum_{i=0}^{k} {}_kC_i\, \mu^i\, (\sigma Z)^{k-i}\right) \\
&= \sum_{i=0}^{k} {}_kC_i\, \mu^i \sigma^{k-i} \mathbb{E}\Big(Z^{k-i}\Big)\,.
\end{aligned}$$

As such, Equation (B.1) with Equation (B.2) and Equation (B.3) let us calculate the moments of powers of the the Gaussian distribution.

As an example, if we take $V \sim \text{Gaussian}(0, \sigma_V^2)$, we can calculate $\mathbb{E}(V^6)$ as

$$\begin{aligned}
\mathbb{E}\Big(V^6\Big) &= \sum_{i=0}^{6} {}_6C_i\, \mu^i \sigma^{6-i} \mathbb{E}\Big(Z^{6-i}\Big) \\
&= {}_6C_0\, \sigma_V^6 \cdot \mathbb{E}\Big(Z^6\Big) + {}_6C_2\, \mu_V^2 \sigma_V^4 \cdot \mathbb{E}\Big(Z^4\Big) \\
&+ {}_6C_4\, \mu_V^4 \sigma_V^2 \cdot \mathbb{E}\Big(Z^2\Big) + {}_6C_4\, \mu_V^6 && \text{since odd moments of } Z \text{ are null} \\
&= {}_6C_0\, \sigma_V^6 \cdot \mathbb{E}\Big(Z^6\Big) && \text{since } \mu_V = 0 \\
&= 1 \cdot \Big(\sigma_V^2\Big)^3 \cdot 5!! && \text{from Equation (B.3)} \\
&= 15 \cdot \Big(\sigma_V^2\Big)^3 .
\end{aligned}$$

B.2 More Numerical Examples

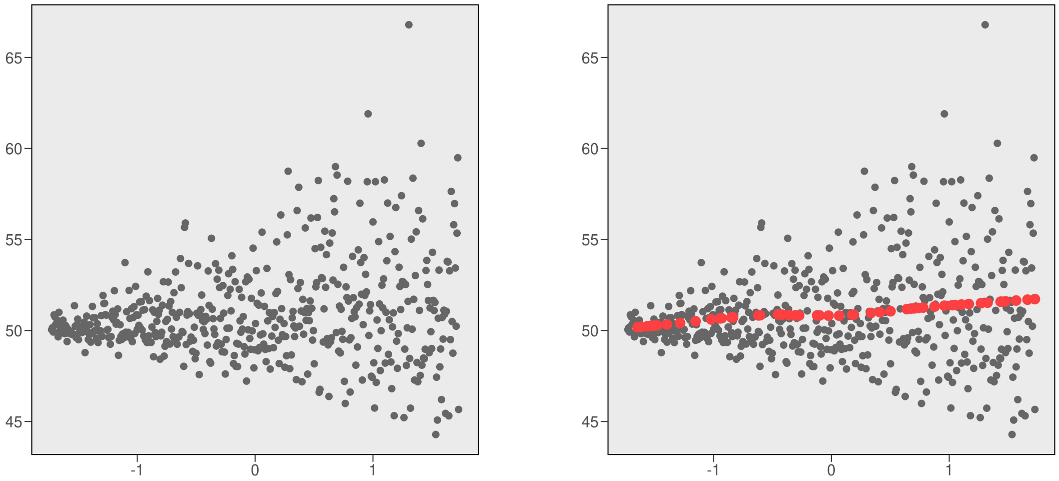
We consider more unusual formulations of $\Lambda(x)$, in the sense that in the context of the modelling of extreme rainfall, one would not expect $\Lambda(x)$ to follow the functions we shall now consider.

B.2.1 Third Example

We first take $\Lambda(x)$ to be defined as

$$\Lambda(x) = \begin{cases} \mu(x) = 50, \\ \sigma(x) = x + 2, \\ \xi(x) = -0.10 \end{cases}$$

for $x \in [-1.729, 1.729]$. The $n = 500$ simulated data is plotted in the left panel (Figure B.1a) and the predicted output in the right panel (Figure B.1b), where we see that the increase in scale parameter as the input values increase is not a realistic situation with respect to rainfall data. We note that we follow the same methodology as explained in Section 4.4 when running TAGI-S.



(a) Simulated Values

(b) Predicted Values

Figure B.1: $\hat{\Lambda}^{\text{(TAGI-S)}}$ and Expected Outputs of Simulated GEV Values

We register MSE values of μ, σ, ξ in Table B.1.

μ	σ	ξ
0.0163	0.0257	0.0363

Table B.1: MSE of Test Values (First Additional TAGI-S Example)

B.2.2 Fourth Example

We also run TAGI-S on simulated where $\Lambda(x)$ is given as:

$$\Lambda(x) = \begin{cases} \mu(x) = 50 = 0.9x^3, \\ \sigma(x) = x + 2, \\ \xi(x) = -0.10 \end{cases}$$

for $x \in [-1.729, 1.729]$. As we can see in Figure B.2, this data set is clearly not realistic (in terms of what one would expect when modelling extreme rainfall). We evaluate how well TAGI-S can infer non-linear patterns in this given context.

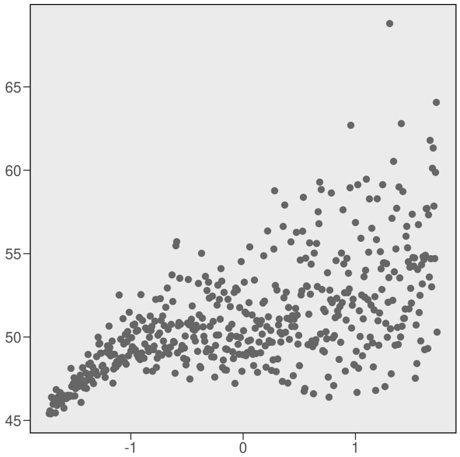
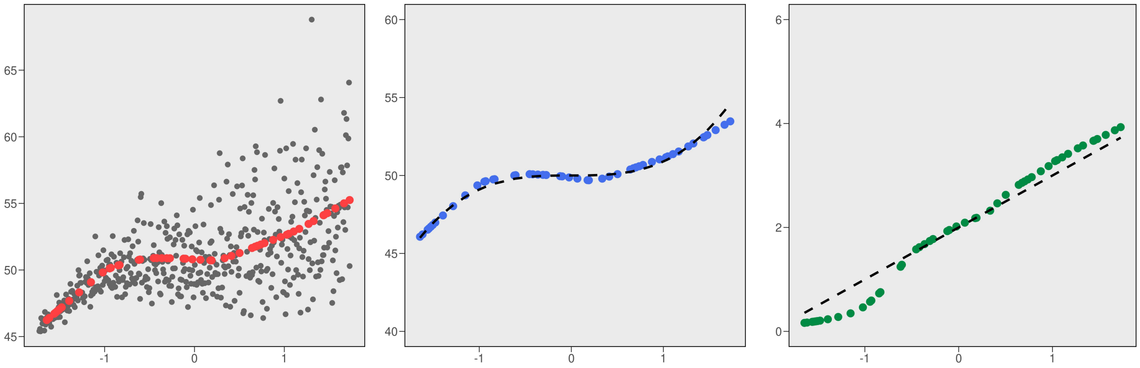


Figure B.2: 500 Random Realisation of a GEV Distributon with $\mu = \mu(x)$, $\sigma = \sigma(x)$ and $\xi = -0.10$.

Running TAGI-S with 14 epochs leads to the following test set values of $\hat{Y}, \hat{\mu}$ and $\hat{\sigma}$ in Figure B.3[1]. The dashed lines in Figure B.3b and Figure B.3c are the functions $\mu(x) = 50 = 0.9x^3$ and $\sigma(x) = 2 + x$ respectively.



(a) Test Set $\hat{Y}$ Values

(b) Test Set $\hat{\mu}^{(\text{TAGI-S})}$ Values

(c) Test Set $\hat{\sigma}^{(\text{TAGI-S})}$ Values

Figure B.3: Predicted $\hat{Y}$, $\hat{\mu}^{(\text{TAGI-S})}$ and $\hat{\sigma}^{(\text{TAGI-S})}$ Values for the Test Set

The MSE values are provided in Table B.2.

μ	σ	ξ
0.0842	0.0627	0.0499

Table B.2: MSE of Test Values (Second Additional TAGI-S Example)

[1] We omit the plot of $\hat{\xi}$ since it is constant and we use the MSE to evaluate the quality of the predicted shape parameter.

The examples here highlight a key strength of TAGI-S: how it can dynamically adapt to extreme data trends. Although the data sets of Figure B.1a and Figure B.2 are very unorthodox situations and are not very realistic scenarios one would expect to see happen in the context of rainfall precipitation, we see how TAGI-S adapts to such extreme data sets.