Faculty of Science
University of Helsinki

FROM DATA PROCESSING TO DISTRIBUTIONAL MODELLING OF TRAFFIC MEASUREMENTS

Jorma Kilpi

DOCTORAL DISSERTATION

To be presented for public discussion with the permission of the Faculty of Science of the University of Helsinki, in Auditorium B123, Exactum, on the 1st of April, 2022 at 13 o'clock.

Helsinki 2022

Department of Mathematics and Statistics
Doctoral School in Mathematics and Statistics (Domast)

ISBN 978-951-51-7916-6 (pbk.)
ISBN 978-951-51-7917-3 (PDF)
Cover image: ©Alma Kilpi 2022
Unigrafia
Helsinki
2022

Abstract

This thesis is motivated by the need to analyse measured traffic data from networks. It develops and applies statistical methods to characterize and to model such data. The application areas are related to *teletraffic* and *telecommunication networks*, *vehicular traffic* and *road/street networks*, and *Internet of Things* applications. The research is based on four scientific publications, augmented with the statistical framework and theoretical development included in this summary. From the applications' point of view, the addressed research problems diverge on the types of the engineering problems, while from the statistical point of view, they share common theoretical methods.

The application problems are: i) to study whether a Gaussian process is a feasible model for aggregated Internet traffic, ii) to obtain aggregated flow level models for flow sizes, flow durations and their bivariate joint distribution, iii) to deduce vehicular traffic routes from correlated counts of vehicles that are observed at different locations of a street network, and iv) to develop a data reduction algorithm that works with limited computational capacity and can be deployed by Internet of Things applications.

This summary provides the statistical framework that combines the developed and applied methodologies and emphasizes their common features. Rigorous mathematical proofs are given for certain less-known, possibly novel, results about mutual information of pairs of order statistics, and a convergence result related to simultaneous estimation of several quantiles. These were used in the publications or, alternatively, bring new statistical insight to the methods that were used in the publications.

Publications

[1] Jorma Kilpi and Ilkka Norros. Testing the Gaussian approximation of aggregate traffic. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurment*, IMW '02, page 49–61, New York, NY, USA, 2002. Association for Computing Machinery.

[2] Natalia M. Markovich and Jorma Kilpi. Bivariate statistical analysis of TCP-flow sizes and durations. *Annals of Operations Research*, 170(1):199–216, 2009.

[3] Jorma Kilpi, Ilkka Norros, Pirkko Kuusela, Fanny Malin, and Tomi Räty. Robust methods and conditional expectations for vehicular traffic count analysis. *European Transport Research Review*, 12(10), 2 2020.

[4] Jorma Kilpi, Timo Kyntäjä, and Tomi Räty. Online percentile estimation (OPE). *Journal of Signal Prosessing Systems*, 93:1085–1100, 2021.

Author contributions

All publications are joint research work with the co-authors of the publication.

Publication [1] The research topic of [1] was suggested by Prof. Norros, who also wrote the motivation for the research. The present author is responsible for the statistical content of the article. All content topics were discussed together several times during the research work and writing of the manuscript.

Publication [2] The research topic was chosen from a couple of topics that the present author was working on before this collaboration. The theoretical approach in [2] was proposed by Prof. Markovich, who mastered the mathematical theory of heavy-tailed distributions. The writing of the manuscript was divided in the following way. The present author is responsible for the contents of Section 1, Section 2.2, Section 3.2, Section 4.2 and its subsections 4.2.1 and 4.2.2 of [2].

Publication [3] The research topic was chosen by the present author as a part of a wider research project lead by Prof. Norros. The complete author contributions of [3] are listed at the end of the article in the Supplementary information section. The present author is responsible for the final mathematical and statistical approach of [3] and did the data analysis for it. The present author also wrote the first draft of the manuscript, all authors then contributed in the writing and revising of the journal manuscript.

Publication [4] The research topic was suggested by the present author, and originally funded as a part of an internal IoT project portfolio at VTT. The complete author contributions of [4] are listed at the end of the article in the Supplementary information section. The present author is responsible for the algorithm development and the statistical foundations of the control loop. The present author also programmed the proof-of-concept implementation with valuable help from Timo Kyntäjä. The first draft of the manuscript was written by the present author as a project report, all authors then contributed in the writing and revising of the journal manuscript.

Acknowledgements

First of all, I want to thank Docent Ilkka Norros. In addition to being my supervisor and co-author, Ilkka has contributed to this thesis as a Research Professor and as a project manager in several projects where the research for publications were done.

I had the luxury of three supervisors and my other two supervisors, Prof. Sangita Kulathinal and Prof. Eero Saksman, also deserve my great gratitude. I want to thank the pre-examiners, Prof. Esa Hyytiä and Prof. Rob van der Mei, for careful reading of the manuscript of this thesis.

I want to thank all other co-authors of the publications: Prof. Natalia Markovich, Pirkko Kuusela, Fanny Malin, Tomi Räty and Timo Kyntäjä. In addition to my co-authors, Kari Seppänen had a significant role in data anonymization that was required for research done in publications [1] and [2].

I want to thank also my Thesis Committee members Pasi Lassila, Dario Gasbarra and Olli Saarela.

I want to thank all my former and present collegues at VTT, the Technical Research Centre of Finland, and I also thank VTT as my employer and for funding the writing of the summary part of this dissertation.

I want to thank my wife Sirpa, and my children Feeliks and Alma for luxurious and loving family life that provides a complete break with the research work. Alma designed the cover image, which is related to publication [3].

Finally, I dedicate this thesis to the memory of my mother, Phil.Lic Sisko Kilpi (1928-2018). Her enthusiasm towards microbiology has been an example to me.

Contents

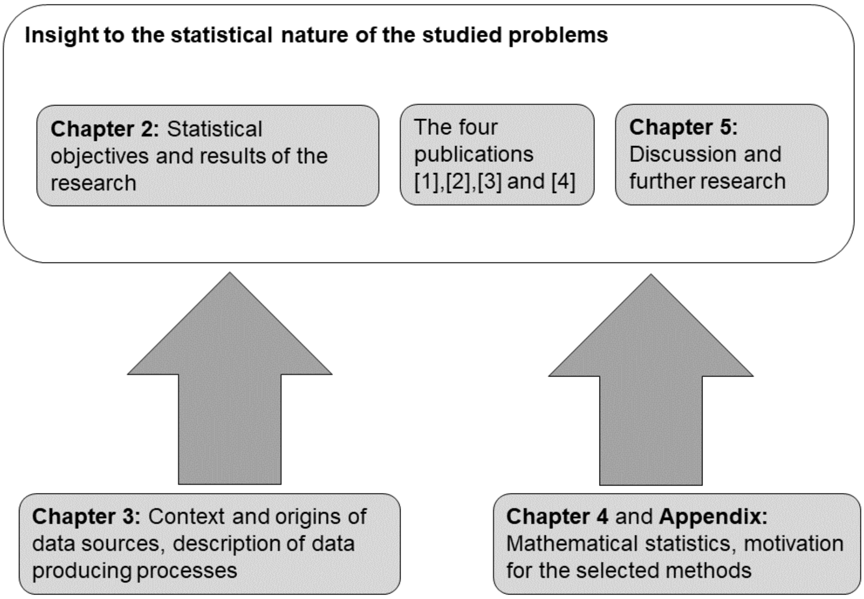
Chapter 1

Introduction

In May 2021, while we had already started to write this summary, we had the privilege of participating in a statistics course led by Neil Sheldon. The aim of the course was to sharpen statistical thinking, the use of statistical concepts and the language used in presenting results of statistical analysis. Sheldon forced us, the students, to think about *why* we use statistical concepts, *when* we should or should not use statistical language and *what* we should communicate to other people, colleagues and society, about our findings. We fully agree with the opinion of Neil Sheldon, which he shared in his lectures, that the purpose of statistics is insight and not numbers. First, a statistician should gain enough insight about the nature of the statistical problem for oneself and then communicate one's insight to other people, from project collaborators to the scientific community, so clearly that they can agree or disagree with the insight. The possibility to disagree is, of course, the more crucial option as it allows us to find better insight in the future. The insight that a statistician can gain on a statistical problem is based on two corner stones, theory and data. In Figure 1.1, the structure of this summary is drawn based on these corner stones. The publications of this thesis contain the major insight to the statistical nature of the studied problems, which combines background knowledge, inference and interpretation.

This research is based on the four scientific publications [1], [2], [3], and [4]. These articles develop and apply statistical methods that are used to characterize, to analyze and to model observed or measured traffic data from a network. The application areas of [1] and [2] are related to *teletraffic* and *telecommunication networks* while [3] is related to *vehicular traffic* and *road/street networks*. Publication [4] describes an algorithm, originally designed for Internet of Things applications, which we expect to have a wider range of applications. For example, the algorithm of [4] has already been applied in intensive network delay measurements of a test network, therefore the application area of [4] includes telecommunications. Table 1.1 offers an overview to the engineering research topics and to the different sources of data that are used in the publications. However, the focus of the publications is more on the developed and applied methods than on the data itself.

From the applications point of view, the addressed research problems diverge on



Insight to the statistical nature of the studied problems
Chapter 2: Statistical objectives and results of the research
The four publications [1],[2],[3] and [4]
Chapter 5: Discussion and further research
Chapter 3: Context and origins of data sources, description of data producing processes
Chapter 4 and **Appendix:** Mathematical statistics, motivation for the selected methods

Figure 1.1: The structure of this thesis.

Table 1.1: Quick reference to the application research topics and data sources.

Publication	**Research topic**	**Data source**
[1]	Feasibility of a Gaussian process as a model for aggregated TCP/IP traffic	TCP/IP packet headers
[2]	TCP flow level models for flow sizes, flow durations and flow rates	TCP/IP packet headers
[3]	Deduce vehicular traffic routes from correlated traffic counts observed at different locations of a street network	Vehicular traffic counts from loop detectors
[4]	Perform data reduction with limited computational capacity	Internet of Things applications

the types of engineering problems, while from the statistical point of view, they share common theoretical basis. Our main objective in this thesis summary is to emphasize the shared common statistical methods and to provide necessary details of them. These include order statistics and quantiles, multivariate analysis, and statistical time dependence models in the context of data traffic or vehicular traffic. We hope to bring a fresh view to order statistics by exploring the mutual information between two order statistics. The statistical framework also contains rigorous proofs of some less-known, possibly novel, results about mutual information of certain pairs of order statistics, and simultaneous estimation of several quantiles. These were used in the publications or bring new insight to the methods that were used. Therefore the scope of the statistical framework of this summary extends beyond the publications.

The wider motivation background and context of publications [1] and [2] is the observation of long-range dependence and heavy-tailed distributions in Internet traffic starting from Leland's Ethernet measurements [Leland et al., 1994], in which features of self-similar processes were noticed. One observed characteristic feature was that the bursty nature of Ethernet traffic does not get smoother when the time scale and the level of traffic aggregation are increased [Leland et al., 1994, Fig.4]. These topics had been rather rare in telecommunications, when they suddenly became the object of wide engineering and mathematical interest.

A common motivation in publications [3] and [4] is the attempt to achieve purely algorithmic solutions to distill statistical information from data, that is, solutions that can be programmed as a single piece of code. In the era of continuous measurements and constantly growing data sets, it seems necessary to process and reduce data online before it is forwarded for applications. In many engineering applications, the end users need the information content of the data rather than the raw data.

This summary is structured as follows. In Chapter 2, we describe the objectives and the results of the research. In Chapter 3, we provide application-specific background knowledge that describes what the data represents and how the data collection was done. In Chapter 4, we provide the statistical framework, which describes the major theoretical issues that were used in the publications, and we provide proofs of some relevant results that were not included in the publications but were used in them. In Chapter 5, we discuss further the insight that we have after the research work has been done and the next steps. In Appendix A, we compute a formula that we use in Chapter 4.

Chapter 2

Statistical objectives and results of the research

In this Chapter, we describe the objectives and the results of the research of each of the the four publications.

2.1 Publication [1]: Feasibility of a Gaussian traffic model

A Gaussian process (X_t) has the property that all of its finite dimensional marginals are multinormally distributed and it is completely determined by its mean value function $\mathbb{E}X_t$ and covariance function $Cov(X_s, X_t)$ ([Parzen, 1962],[Priestley, 1982]). The main objective in publication [1] was to study the possibility to model the increments of the aggregate TCP/IP traffic flow with a Gaussian process in different time scales when the data have long-range-dependent (LRD) features. We focused on 1-dimensional marginal distribution of increments of real traffic and on the question of how well a normal distribution approximation describes the data. No assumption about the covariance structure was done. During the research work, a statistical description of required aggregation types, vertical and horizontal, arose as a research objective. Vertical aggregation means the amount of users per time slot of width Δ and horizontal aggregation means the width of the time slot.

We studied a TCP/IP packet data trace which was aggregated into different scales Δ, from 1 millisecond (ms) up to 4 seconds (s). The scales increased in doubled manner: 1 ms, 2 ms, 4 ms, and so on until 4096 ms = 4 s. We showed that the known method, which is based on computing the linear correlation coefficient r_n^2 from normal-quantile plots, is able to distinguish between a relatively good fit and a bad fit. We used this method to study different scales and considered the behavior of $r_n^2(\Delta)$ when the sample size n was increased. Because the method is simple to compute it was meaningful to study increasing sample sizes. The possibility to consider increasing sample sizes was important since LRD means that a sample of size n contains less information about the possible model parameters than an independent sample of the same size n would contain. In practice, it means that the convergence towards stable parameter values is slower.

Increasing the sample size shows whether the normal distribution fit improves or seems to stop improving at some point. Both negative and positive cases were shown based on the data. We also compared the empirical tail $1 - F_n(x) = \mathbb{P}_n\{X > x\}$ against the model tail $1 - \Phi_{\mu,\sigma}(x)$. The same method also applies to lognormal models and comparison between the normal and lognormal fit is fruitful since the lognormal distribution is known to belong to the class of subexponential distributions and, therefore, lognormality indicates that the assumed Central Limit Theorem (CLT) based assumption does not hold. The main result can be formulated by saying that the CLT assumption must hold between those sources that contribute in the largest magnitude to the aggregate traffic rate. If the largest magnitude contributors are rare, then even the sum of all smaller magnitude contributors is not comparable to a single large contribution.

2.2 Publication [2]: TCP data flow sizes, durations and rates

The first objective in publication [2] was to characterize the univariate heavy-tail properties of TCP flow sizes S and flow durations D. The second objective was to model the bivariate dependence structure of the joint distribution of (S, D). The third objective was to obtain the distribution of average flow rates, defined as the ratio $R = S/D$. There was also some interest in dependency between the pair (S, R).

The results were based on analysis of TCP connection data of mobile Internet users. The data of flow sizes and durations were highly variable and had subexponential features. First, we applied several known methods to study the heaviness of the tails. Then we approximated the distribution of the TCP flow rate by deriving it from the joint bivariate extreme value distribution of the maxima of flow sizes and flow durations. Due to the heavy tailed nature of flow sizes and durations, the joint distribution was represented by a bivariate extreme value distribution using the Pickand's dependence function $A(t)$, $0 \leq t \leq 1$. We estimated the A function with known non-parametric estimators to measure the dependencies of random pairs: (S, D), (S, R), and (D, R). In [2, Section 4.2.1], we provided a generally applicable method to test that the achieved estimate of the A function is good. This method is based on the observation that the Pickand's A function allows to write the distribution function of the ratio of the two variables in terms of A and its derivative A' [2, formula (14)]. We also demonstrated the use of this method by selecting a parametric model, the logistic model, for A. The selection of the logistic model was based on the non-parametric estimates of A. In this way, we obtained a computable distribution model for the flow rates of large flows with $S \geq 200$ kB [2, Section 4.2.2].

2.3 Publication [3]: Multinormal models for vehicular traffic

The major research work of publication [3] was done in the Finnish Academy-funded project called Stomograph. The broad objective in the Stomograph project was to recover the vehicular traffic routes from traffic count data that was collected from different, mutually relevant and correlated locations from a central area of the city of Tampere. These locations bounded the area and there were also measurements from locations inside the area. In publication [3], we restricted the study to the boundary locations of the area. The objective of the research in publication [3] was limited to utilize the information from correlated counts of vehicles in two (or three) mutually relevant locations and to deduce smaller spatial scale conclusions about traffic dynamics from these correlations.

The result is the following algorithmic framework that can be used to extract information about traffic dynamics from counts of vehicles in the case where the traffic counts are available in the opposite directions and in two or more locations. Denote two such locations as 1 and 2 and also the two directions by 1 and 2. The data was counts of vehicles per 15-minute time slots X_{ij}, with $i = 1, 2$ as a location and $j = 1, 2$ as a direction at the location. Our algorithmic framework was built on several basic ideas. First, we selected and named the locations and directions. Second, we performed a linear transformation for the data in order to change the focus to the difference and to the sum of the counts of vehicles in the opposite directions in every location. The difference $Z_i = X_{i1} - X_{i2}$ was called asymmetry and the sum $V_i = X_{i1} + X_{i2}$ was called volume at location i. Mutually relevant means locations where it was justified to assume that the two asymmetries (Z_1, Z_2) correlated due to detecting a proportion of the same vehicles at these locations. Then we used multinormal distributions as a baseline model for the asymmetries (Z_1, Z_2) at different locations. Third, we estimated the parameters of the multinormal distributions, including correlation, using robust methods. The fourth idea was the sample version of conditional expectation, which is supported by the model-based estimates with confidence intervals.

Using these steps we presented three applications of the framework. The first application was to recognize that the correlation matrices of the asymmetries can be used to restrict the solution space of the more general origin-destination matrix estimation problem. In the second application, we computed conditional expectations of type $\mathbb{E}(Z_2|Z_1 > a)$ and deduced results about traffic dynamics from these. That is, we quantified how much asymmetry in one location affects the asymmetry in another location. The third application was to reconstruct missing data in one location given the traffic dynamics in nearby locations.

2.4 Publication [4]: Online percentile estimation (OPE)

Our objective was to develop a statistical data reduction algorithm for IoT applications. It was assumed that an IoT application produce univariate data and, because of low latency requirements, the computation of the algorithm should be performed near the origin of the data. Therefore, the algorithm should be applicable in situations where the computational resources, CPU power and memory, are limited.

The result is an algorithm called OPE, which is best described as a control loop over an existing sequential quantile estimation algorithm. We used the extended P^2 algorithm of [Raatikainen, 1987], but the control loop can be based on any sequential quantile estimation algorithm. The control loop either computes the sample quantiles by ordering a small buffer or uses the sequential algorithm to estimate quantiles of any univariate input data sequence. It transforms an univariate input sequence into output sequence of (variable bin width) histograms without storing or sorting (ordering) all of the observations. The algorithm is designed to continuously test whether the input data appears stationary, and to react to events that do not appear to fit in the stationary model. By using meta-data, OPE indicates how to interpret the histograms since their information content varies according to whether the quantiles, which define the histogram, were computed or estimated. It works with parameter-defined, fixed-size small memory and limited CPU power. The control loop algorithm has a built-in feasibility metric called *applicability*. The applicability metric is based on the meta-data that OPE produces and it indicates whether the use of the algorithm is justified for a data source: OPE is designed to work for an arbitrary univariate numerical input, but it is statistically feasible only if the data source is in a stationary state more often than in a non-stationary state. The theoretical part includes an extension of a known mathematical theorem about the convergence of a single quantile estimate. The extended theorem covers the case of simultaneous estimation of several quantiles. We also presented the results of a performance study done for the algorithm with positively autocorrelated simulated data from the moving average model.

Chapter 3

Background knowledge of applications

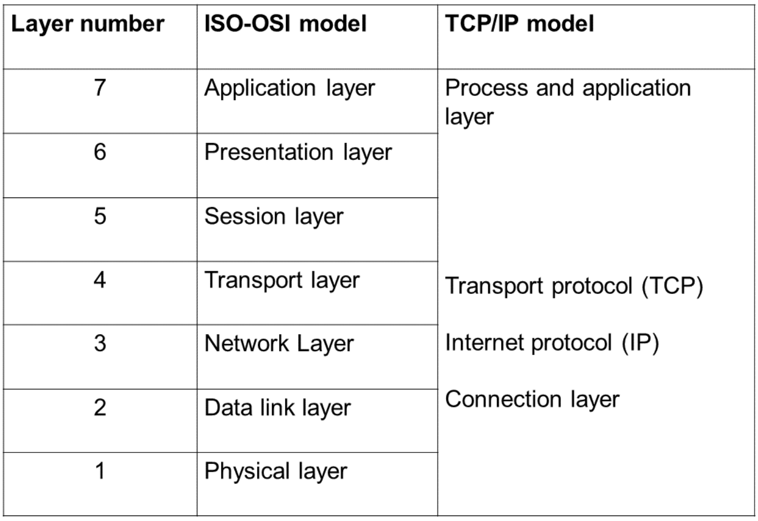
The beginning of this chapter is aimed for those readers who are not yet familiar with the *Transport Control Protocol (TCP)*, the *Internet protocol (IP)*, and TCP/IP networks. A reader who is already familiar with TCP/IP should at least have a look at the figures and tables as they will be referred to later on. In addition to TCP/IP, we aim to familiarize the reader with some relevant issues about the *Internet of Things (IoT)* and *vehicular traffic.*

3.1 A brief overview of TCP/IP networks

A fast way to understand the functionality of a communication network is to consider *the layered architecture* description of it ([Stevens, 1994], [Medhi and Ramasamy, 2007], [Heckmann, 2006], [Lin et al., 2012]). Figure 3.1 shows the layers. The layered architecture means that the functionality of a layer is based on the functionality of the lower layer. In this context, there are two different descriptions to consider. The first is the *Reference Model for Open Systems Interconnection (OSI model)* made by the *International Standardization Organization* and the other model is the TCP/IP model.

The OSI model is an abstraction and the TCP/IP model is the one that is actually needed in the context of this thesis. However, together they give an overview of the different functions that a communication network has and how a TCP/IP network functions.

A protocol is an explicit set of messages and associated rules, which two or more devices must obey so that they can communicate with each other. The notation TCP/IP is read as "TCP over IP" and it essentially has the meaning that the functionality of the TCP layer is based on the functionality of the IP layer. Each node of an IP network has the same TCP/IP protocol stack implemented, and the corresponding layers at each node communicate using the protocol of the layer. An example of the connection layer protocol is the Ethernet, which was the measurement interface of the data used



Layer number	ISO-OSI model	TCP/IP model
7	Application layer	Process and application layer
6	Presentation layer	
5	Session layer	
4	Transport layer	Transport protocol (TCP)
3	Network Layer	Internet protocol (IP)
2	Data link layer	Connection layer
1	Physical layer	

Figure 3.1: Layered architecture models.

in publications [1] and [2], and also in the early motivation study [Leland et al., 1994] discussed in Introduction.

The TCP/IP networks are *packet switched* networks. It means that the source node segments a data file into payloads of *TCP packets* and transmits them over the network by *IP packets* that carry the TCP packets as payload. At the destination node, the small segments are collected and reassembled back to the original file. The TCP layer of the source node takes care of the segmentation and the TCP layer of the destination node takes care of reassembling the data. The TCP packet header contains the information that the destination node needs to reassemble the original file. The IP layer functions take care that every IP packet that is sent from the source node eventually finds it way to the destination node and these functions use the header information of the IP packet.

A *TCP connection* is a connection between the end points and traffic in the connection can flow in both directions. The TCP protocol takes care that possibly lost packets are re-transmitted and that packets that arrive out-of-sequence are reordered at the receiving end node. In publication [2], we call the sequence of TCP/IP packets that traverse from the source node to the destination node *a TCP flow*. Thus, connection is a bidirectional concept and flow is a unidirectional concept; a TCP flow is a part of a TCP connection.

A characteristic property of IP networks is that the individual packets of a single TCP flow may traverse different routes between the source and the destination. It is customary to draw an IP network as a cloud to indicate that all routes inside the cloud are possible.

The TCP layer of the destination node sends *acknowledgment (ACK)* packets back to the TCP layer of the source node and with the information from the ACK packets the source node knows to re-transmit a lost segment, send packets faster or send them at a slower pace. TCP has sliding window-based *flow control mechanisms* that allow a slow receiver to slow down the sending rate of the data sender. The time it takes for a piece of information to traverse from the source node to the destination node and back to the source node is called the *round-trip-time (RTT)*, see Figure 3.2. The RTT and possible *packet loss* determine the performance of a TCP flow. A packet loss is considered to be a sign of congestion in the network and the sender TCP node reacts to the packet loss by decreasing its sending window size. This is called *a congestion control mechanism.*

The dynamics of a single TCP connection can vary a lot. The source slides the sending window over the segmented data file and, when the earliest segment of the window is acknowledged, it moves the window onward over the segmented file and sends a new segment. The sliding window controls the number of unacknowledged segments that can exist at any time. However, the consecutive ACK packets may have an improper spacing due to the cross-traffic, the other traffic in the network, which affects by mixing the ACK packets with the cross-traffic in the queues along the reverse path(s) from destination to the source. The spacing between consecutive ACK packets may be diminished so that the ACK packets arrive to the source in clusters. In [Lin et al., 2012] this is called the ACK-compression problem. It has the consequence that the source node sends a burst of packets and waits for the feedback from the destination before it sends the next burst. The burst sizes vary according to the flow control and congestion control mechanisms. The TCP protocol tries to maximize the throughput and, occasionally, the source node may send larger bursts than what the destination node receives. The protocol measures the burst size in segments or packets and it is usually measured in bytes in the data analysis. The concepts of a burst and the burst size variability, and RTT are important in publication [1] and it is the message to remember from Figure 3.2.

Every TCP/IP packet contains *an IP header* to perform the functionality needed at the IP layer and *a TCP header* to perform the described functions at the TCP layer. These headers contain necessary information that these protocols need to perform the above described data transfer functions, including flow control and congestion control. The size of the IP header is 20 bytes and the size of the TCP header is 20 bytes, making it 40 bytes in total. The size of an ACK packet, which carries information from the destination back to the source is 40 bytes. If this header information is available afterwards, then it is possible to reconstruct the protocol events that occurred during the file transfer process. Usually, this header information is unavailable since there are several possible routes and, excluding the source and the destination nodes, there is no such place where the header information of every packet of the TCP connection could be monitored.

In the case of data we used in publications [1] and [2], the situation was as illustrated in Figure 3.3 below. All traffic traversed through a gateway node between two IP clouds, therefore it was possible to monitor and, in case of the data used in [2], also to reconstruct the events of TCP connections from the TCP/IP packet header data collected at the gateway.

One important aspect in the reconstruction of the events of a TCP connection is the

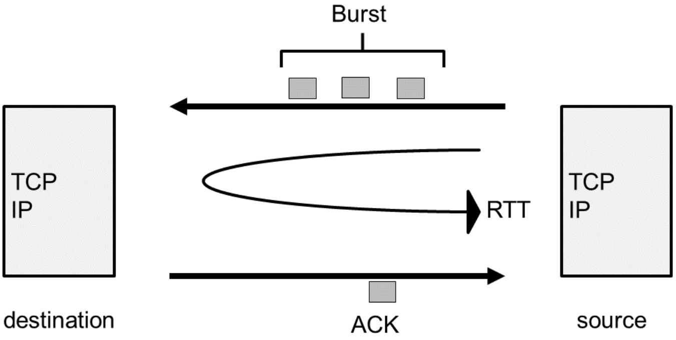


Figure 3.2: The TCP source transmits data packets in bursts.

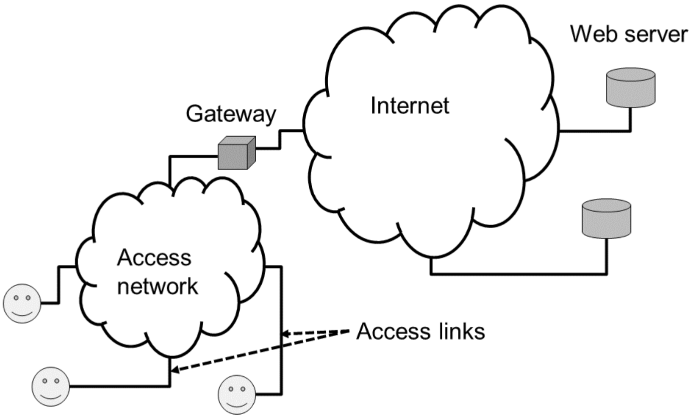


Figure 3.3: A gateway node between two IP clouds.

ability to detect the beginning and the end of it. A TCP connection begins with *the three-way handshake* procedure which is recognized from the *synchronization bits (SYN)* at the TCP headers. The end of the connection is recognized with *the finish bit (FIN)* at the TCP header.

Data *transfer rates* are of interest both at the level of an individual TCP flow of a user and at the level of the aggregate of TCP flows of several users. The aggregate of TCP flows is called *a traffic flow.* The concept of a flow always has a direction associated to it and, in publications [1] and [2], we call the direction either *an upstream* or *a downstream* direction. In the context of web browsing traffic (HTTP or HTTPS protocol), the downstream direction means from a web server on the Internet towards an end user, and the upstream direction means from an end user towards a web server on the Internet.

In publications [1] and [2], we express the transfer rates with the unit *bits per second (b/s)* and its derivatives when multiplied by 10^3, as explained in Table 3.1. The *file sizes* are commonly expressed in the unit *byte (B)* and its derivatives when multiplied by $2^{10} = 1024 \neq 10^3$. The prefixes "kilo", "Mega" and "Giga" have different meanings that depend on the context. Also, $1\mathrm{B} = 8\mathrm{bits} = 8\,\mathrm{b}$. As an example, with constant data transfer rate 384 kb/s, it takes *at least* 14 minutes and 34 seconds to transmit 40 MB of data: $40 \times 2^{20} \times 8\,\mathrm{b}/384 \times 10^3\,\mathrm{b/s} \approx 874s$. A more realistic estimate takes into account the overhead due to TCP/IP packet headers and the performance of the TCP protocol that is affected by the RTT and packet loss. One such estimate rate is given in [Heckmann, 2006] by the formula for the average rate r of a long-lived TCP flow

$$r = 1.22 \frac{\mathrm{MTU}}{\mathrm{RTT}\sqrt{\frac{2}{3}p}}, \tag{3.1}$$

in which p is the packet loss probability and MTU is *the maximum transmission unit*, that is, the maximum TCP/IP packet size. For an individual TCP flow, the MTU information may be available in the header information at the beginning of the connection since it can be negotiated. If it is not negotiated, a default value is used. Even in the case of a single TCP connection, the RTT should be considered as a random variable with ideally a small variance so that, for example in (3.1), the 'RTT' is interpreted as the expected value of the RTT. However, for the TCP protocol RTT is just a parameter that is estimated at the beginning of the connection. In the context of an aggregate TCP traffic, the RTT refers to the distribution of the RTTs that each contributing TCP connection uses as its parameter.

A reconstruction of the events of the TCP connection from the header information allows to estimate the RTT and p. The same data that is analyzed in publication [2] were also used in an earlier study [Kilpi and Lassila, 2006] where we analyzed the RTTs.

The *aggregate traffic process* $A(0,t)$ represents a cumulative amount of all traffic in one direction during a time period $[0,t]$. The difference $A(0,t_2) - A(0,t_1)$ is *an increment* of traffic during the interval $]t_1,t_2]$, $0 < t_1 < t_2 < t$. The time interval $]t_1,t_2]$ is *a slot* and the width $(t_2 - t_1)$ of the time slot is *a scale.* In publication [1] we use the word 'resolution'

Table 3.1: Different magnitudes of the basic units.

Type	**Value**	**Unit**	**Description**
Transfer rate	1	b/s	bits per second
	10^3	kb/s	kilobits per second
	10^6	Mb/s	Megabits per second
	10^9	Gb/s	Gigabits per second
File size	1	B	Byte
	2^{10}	kB	kilobytes
	2^{20}	MB	Megabytes
	2^{30}	GB	Gigabytes

for scale. If the scale is finer than the RTT of a TCP connection, then the connection may not be able to contribute to the amount of traffic in consecutive slots. On the other hand, if the scale is larger than the RTT of the connection then the source may contribute to the consecutive time slots. If a scale is chosen afterwards and if the scale is smaller than many of the individual RTTs of all contributing sources, then we also perform unintentional selection (possible *selection bias*) as some of the connections may not even be able to contribute to the aggregate traffic at every consecutive slot. Therefore, RTT is an important factor also in publication [1] even though it is not emphasized there. In publication [1] the measurement was layer 3 level with some information about layer 4, like protocol, port numbers and the size of the payload.

The results of publication [1] are based on data analysis of an IP-packet level traffic trace that was measured from a gateway node which connected two communication networks as illustrated in Figure 3.3. The other network was a dial-up network of a Finnish Internet Service Provider and the other network was the Internet. The monitoring location was such that the trace could be considered statistically representative in the following sense. A relatively large number of traffic sources contributed to the aggregate traffic at the measurement point. Moreover, users were home users with limited individual access link rates (typically less than 64 kb/s, at most 128 kb/s) compared to the aggregate traffic rate (> 1 Mb/s) at the measurement point. This meant that the largest bursts of packets that the TCP protocol of the source nodes injected should be limited in size. Therefore, if the aggregate traffic rate at the measurement point were large enough, no single traffic source should be able to dominate in the trace.

Both the measurements, of publication [1] and of publication [2], were done from a commercial network. To protect the end users' privacy, the IP addresses were anonymized before analysis and the business secrets of the operator were kept confidential.

3.2 Internet of Things

The Internet of Things (IoT) means the ability to connect all kinds of devices with the Internet so that they are accessible via an Internet connection. In the first place, this requires methods to address a device so that the network layer protocol can find the device. Version 6 of the IP (IPv6) has a very large address space but, actually, methods to reuse the existing address space of version 4 of the IP (IPv4) are also very efficient. As soon as a device is connected to the Internet and has an IP address, data communication becomes possible. There are, of course, a large number of privacy and security issues that need to be solved. IoT can sometimes be just an extension of TCP/IP, but other layer 4 protocols may be preferred instead of TCP. The main difference related to earlier TCP/IP discussion is that the machine-to-machine (M2M) communications dominate the IoT concept. M2M means that communication occurs because some algorithm detects a situation where the information exchange is needed and opens a connection for communication.

The typical devices that can be connected in the IoT framework include *meters* for measuring the energy, electricity or water consumption, *sensors* for measuring temperature, pressure, humidity, speed, vibration, or detection of the presence of a vehicle, and *controllers* that can detect working modes (on/off) or working states (high speed/low speed) of engines or devices. These devices provide measured information at some rate that can be fast or slow. In publication [4] we target the cases where the information rate is high. There is also a question of where the data processing is optimally done: cloud computing is a tempting solution due to huge memory and CPU capacity, but latency is then also large. Computation in the proximity where the data collection is done may be needed if small latency is required but then the memory and CPU capacity are limited and this is the context of publication [4].

3.3 Vehicular traffic

While one second is a long time in data communication, in vehicular traffic the time scales of interest start from 1 minute and include tens of minutes, hours, days, weeks, months and years. Vehicles are not dropped, duplicated, re-transmitted nor segmented. Queues, congestion and traffic flows are, however, concepts that are similar to packet data traffic.

In the case of vehicular traffic, a flow of all vehicles can be divided into sub-flows in many ways. The flow of buses can be considered as a separate flow from other traffic flow in the same direction. If there are more than one lane available, vehicles in different lanes can be considered to form different flows.

The routing functionality is basically in the head of the driver of a vehicle. However, if a driver uses a GPS navigator then this navigator system is analogous to the routing layer functionality of a communication network. Traffic lights, the lane and the road signs in street crossings, in turn, form a functionality that is analogous to the link layer *switching* functionality of communication networks.

In the data communication case, a single physical network can maintain several *logical* networks and the logical *network topology* need not be the same as the physical network topology. For example, the users of the communication network can be divided into several groups that use the same physical network but are unaware of other groups. An analogy in the vehicular traffic case could be a network of regular bus lines in a city, which traverse the same streets as other traffic, but does not traverse through all of the possible streets.

The vehicle count data is obtained by loop detectors. A loop detector is an example of an IoT application: it measures changes in the inductance of a wired loop that is embedded under the road surface and forwards this data onward. Changes in the inductance occur when a vehicle drives over the loop. An algorithm interprets the changes in the inductance as counts of vehicles and the computation of this algorithm can be done either in the loop detector, in some control unit of several loop detectors, or at the cloud server where all raw data is collected. Loop detectors are often located in street crossings near traffic lights. This is the context and origin of traffic count data that is used in publication [3].

Chapter 4

Statistical framework

In this chapter, we provide a statistical framework that, in a sense, combines the methods of the publications of this dissertation under a broad theoretical umbrella. The scope of the framework extends beyond the publications because we also provide new results which are closely related to the publications but are not included in them. Our aim is to facilitate the reading of the publications of this thesis by giving some background knowledge, motivation and intuition of the selected methods. However, we do not aim for a comprehensive coverage of all statistical issues that are used in the publications. Indeed, the methods in publications [1] and [3] are motivated by *the Central Limit Theorem (CLT)* and, therefore, they should be more common. The methods in publications [2] and [4] are less common and we will emphasize them more. We start with the order statistics and quantiles, and then we present some multivariate issues and, finally, some time dependence issues. At the end of this chapter there is a brief summary of the framework.

4.1 Order statistics and sample quantiles

We begin this section by introducing some theoretical results about *mutual information* of pairs of order statistics. Both mutual information and order statistics are well-known basic statistical concepts supported by a vast literature ([Cover and Thomas, 2006], [Casella and Berger, 2002], [Nevzorov, 2001]). In [Ebrahimi et al., 2004], mutual information of two *consecutive* order statistics was computed. We believe that the results of [Ebrahimi et al., 2004] about mutual information bring some insight to publication [2]. We will show how to compute the mutual information from the basic definitions of [Cover and Thomas, 2006], [Casella and Berger, 2002] and some further assumptions. Actual computations of the mutual information are done later in Appendix A.1.

4.1.1 Mutual information of pairs of order statistics

Let X and Y be two real-valued random variables with a joint density function $f(x, y)$ and let the marginal density functions be f_X and f_Y, respectively. The mutual information $I(X;Y)$ is defined as

$$I(X;Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\log\frac{f(x,y)}{f_X(x)f_Y(y)}dxdy \tag{4.1}$$

The main property of the mutual information is that $I(X;Y) \geq 0$ with equality if and only if X and Y are independent [Cover and Thomas, 2006]. In information theory, the uncertainty of a random variable means the complexity that is required to describe it, and this *descriptive complexity* is measured by the entropy of the random variable. The mutual information $I(X;Y)$ is interpreted as the reduction in the uncertainty of X after observing or knowing the value of Y. Since $I(X;Y) = I(Y;X)$, this interpretation is symmetric between X and Y.

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a distribution F. Ordering of the sample gives the order statistics $X_{(1)} \leq \cdots \leq X_{(n)}$. The new "name", the index (i), is given for each random variable and this index indicates that there are i values that are smaller than or equal to the i:th order statistic $X_{(i)}$ in the sample of size n. The actual ordering can be done only after the sample is observed, in the case of random variables the order statistic notation is thus conditional on the ordering to be done. It is also conditioned on the sample size n.

If X_i and X_j are independent, then $I(X_i;X_j) = 0$. In the case of the order statistics, however, we expect that $I(X_{(i)};X_{(j)}) > 0$ because, when $(i) \neq (j)$, they contain information about each other. This difference seems contradictory since $X_{(i)}$ and $X_{(j)}$ are members of the original sample $\{X_1, \ldots, X_n\}$. The explanation for this is that $X_{(i)} = T(i, X_1, \ldots, X_n)$, that is, each order statistic is a function T of the whole sample and *the parameters* i and n. Function T consists of two parts, *sorting* and *selection.* Whenever $n \geq 3$, there exists at least one variable X_k which affects both $X_{(i)}$ and $X_{(j)}$ but is not equal to either of them. The X_k acts as a common cause for $X_{(i)}$ and $X_{(j)}$ and makes them associated.

From now on, assume that the distribution function $F(x)$, $x \in \mathbb{R}$, is continuous and, following [Casella and Berger, 2002], the density of an order statistic $X_{(i)}$, denoted as $f_{(i)}$, can be written in terms of the distribution function $F(x)$ and the density function $f(x) = F'(x)$ as follows

$$f_{(i)}(x) = \frac{n!}{(i-1)!(n-i)!}f(x)F(x)^{i-1}[1-F(x)]^{n-i}, \qquad x \in \mathbb{R}. \tag{4.2}$$

The joint density $f_{(i)(j)}$of two order statistics $X_{(i)}$ and $X_{(j)}$ can also be expressed in terms of F and f as ([Casella and Berger, 2002])

$$f_{(i)(j)}(x,y) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}f(x)f(y)F(x)^{i-1}[F(y)-F(x)]^{j-i-1}[1-F(y)]^{n-j},$$

$$(x,y) \in \mathbb{R}^2,\ x < y \text{ and } 1 \leq i < j \leq n. \tag{4.3}$$

Together (4.2) and (4.3) suggest that by making some assumptions about the distribution function F or specifying it somehow, then it is possible to obtain some results of the mutual information (4.1) of any pair of the order statistics.

The definitions (4.2) and (4.3) hold when F is continuous. Assume further that F is strictly increasing. A consequence of this assumption is that the inverse function F^{-1} exists and, by making *the probability integral transformation (PIT)* $X \mapsto F(X) = U$, all computations can be transferred to the unit interval $[0, 1]$. The transformed variables $U_i = F(X_i)$ are uniformly distributed in the unit interval $[0, 1]$ with the order statistics $U_{(i)} = F(X_{(i)})$.

Generally, when making the componentwise one-to-one PIT

$$(X, Y) \mapsto (F_X(X), F_Y(Y)) = (U, V), \tag{4.4}$$

the Jacobian of $(x, y) \mapsto (F_X(x), F_Y(y)) = (u, v)$ is

$$J(u, v) = \frac{1}{f_X\left(F_X^{-1}(u)\right) f_Y\left(F_X^{-1}(u)\right)} = \frac{1}{f_X(x) f_Y(y)}. \tag{4.5}$$

Also, $du = dF_X(x) = f_X(x)\, dx$ and $dv = dF_Y(y) = f_Y(y)\, dy$. Next, let f be the joint density of the pair (X, Y) and compute

$$\begin{aligned} I(X; Y) &= \iint f(x, y) \log\left(\frac{f(x, y)}{f_X(x) f_Y(y)}\right) dxdy \\ &= \iint f\left(F_X^{-1}(u), F_Y^{-1}(v)\right) \log\left(\frac{f\left(F_X^{-1}(u), F_Y^{-1}(v)\right)}{f_X\left(F_X^{-1}(u)\right) f_Y\left(F_Y^{-1}(v)\right)}\right) J(u, v)\, dudv \\ &= I(F_X(X); F_Y(Y)) \\ &= I(U; V). \end{aligned} \tag{4.6}$$

This computation shows a known fact that the mutual information (4.1) is invariant under the component-wise one-to-one PITs. Since F is an increasing function, it is order preserving: $F(X)_{(i)} = F(X_{(i)})$. Therefore, a consequence of (4.6) is that

$$I(X_{(i)}; X_{(j)}) = I(U_{(i)}; U_{(j)}). \tag{4.7}$$

The integration range was dropped from the notation in (4.6) because it may change with PITs.

The formulae (4.2) and (4.3) become computationally easy since the distribution and the density functions of the uniform distribution are the simplest possible: $F_U(u) = u$ and $f_U(u) = 1$. Moreover, the $U_{(i)}$ follows *the Beta distribution* law $Beta(i, n-i+1)$. The Beta distribution and *the Beta function* have a fundamental role in the computations and in the interpretation so we will briefly list some of the properties of the Beta function for further reference.

The Beta function $B(\alpha, \beta)$ is defined as

$$B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du. \tag{4.8}$$

The value of the Beta function can be expressed in terms of the Gamma function Γ,

$$\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt, \tag{4.9}$$

as

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{4.10}$$

Using these facts it is possible to compute partial derivatives $\frac{\partial}{\partial\alpha}$ and $\frac{\partial}{\partial\beta}$ of the Beta function in two ways, both in (4.8) and in (4.10), and this gives the formulae

$$\frac{\partial}{\partial\alpha}B(\alpha,\beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\log u \, du = B(\alpha,\beta)[\psi(\alpha) - \psi(\alpha+\beta)] \tag{4.11}$$

and

$$\frac{\partial}{\partial\beta}B(\alpha,\beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\log(1-u)du = B(\alpha,\beta)[\psi(\beta) - \psi(\alpha+\beta)] \tag{4.12}$$

where the ψ-function

$$\psi(x) = \frac{d\log\Gamma(x)}{dx} = \frac{\Gamma'(x)}{\Gamma(x)} \tag{4.13}$$

is the logarithmic derivative of the Gamma function. It is also computationally feasible to express the factorials of (4.2) and (4.3) in terms of the Gamma function, which has the recursive property $\Gamma(x+1) = x\Gamma(x)$ and with positive integers this implies $\Gamma(n+1) = n!$. In addition to the Gamma function, we use the n:th *Harmonic number* $H_n = \sum_{k=1}^n 1/k$ that is related to the ψ -function at the integers $n \geq 1$ as follows [DLMF, , Eq. 5.4.14]

$$\psi(n+1) = H_n - \gamma_E, \tag{4.14}$$

where γ_E is *the Euler's constant*:

$$\gamma_E = \lim_{n\to\infty}(H_n - \log n) \approx 0.57721566490153286061\ldots \tag{4.15}$$

Also, $\psi(1) = -\gamma_E$ and we define $H_0 = 0$ so that (4.14) actually holds for all $n \geq 0$. In the literature, γ_E is also called *the Euler-Mascheroni constant.* We consider (4.14) important because the interpretation of the ψ-function from its definition (4.13) is difficult. The equality (4.14) states a close relationship to Harmonic numbers and it is widely known that $H_n \to \infty$ when $n \to \infty$. Moreover, the recursive property $H_n = H_{n-1} + 1/n$ is immediate to check from the definition of H_n.

4.1.2 Mutual information of two consecutive order statistics

In publication [2], the distribution of $X_{(n)}$ is of interest when we estimate the upper tail probabilities of X since

$$\{X_{(n)} > x\} = \{X_1 > x \text{ or } \ldots \text{ or } X_n > x\} = \bigcup_{i=1}^n \{X_i > x\}$$

and, when all the variables have identical distribution,

$$\mathbb{P}\{X > x\} \leq \mathbb{P}\{X_{(n)} > x\} = \mathbb{P}\left(\bigcup_{i=1}^{n}\{X_i > x\}\right) \leq \sum_{i=1}^{n}\mathbb{P}\{X_i > x\} = n\mathbb{P}\{X > x\}. \quad (4.16)$$

In publication [2], we have a large number $N = 610\,000$ of observations of TCP flow sizes which we treat in the order of arrival. We divide the data into m blocks of size n with $N = mn$ with different choices for n and m. By taking the largest value $x_{(n)}$ of each block we obtain data of the block maxima, which we use to estimate the distribution of the largest observation $X_{(n)}$. This classical *Extreme Value Theory (EVT)* approach is quite a waste of data! However, the second-largest variable $X_{(n-1)}$ also contains useful information about the upper tail since, after observing $X_{(n-1)} = x_{(n-1)}$, the probability of the event $\{X > x_{(n-1)}\}$ is always strictly positive and it has the empirical probability estimate equal to $1/n$. This is the simplest data-based prediction model we can think of that extends beyond data: we ignore the value $x_{(n)}$ and give probability $\frac{1}{n}$ to the infinite interval $]x_{(n-1)}, \infty)$. We interpret this to mean that observing the second-largest observation reduces the uncertainty of the largest observation in this way. By symmetry, $X_{(n)}$ reduces the uncertainty of $X_{(n-1)}$ the same amount. Next, we study $I\left(X_{(n-1)}; X_{(n)}\right)$ to see how much the uncertainty can reduce.

In [Ebrahimi et al., 2004], a more general closed form result of any two *consecutive* order statistics was claimed:

$$I\left(U_{(i)}; U_{(i+1)}\right) = \log B(i+1, n-i) + n\psi(n) - i\psi(i) - (n-i)\psi(n-i) - 1 \quad (4.17)$$

$$= nH_{n-1} - iH_{i-1} - (n-i)H_{n-i-1} - 1 - \log\binom{n}{i}. \quad (4.18)$$

The formula (4.17) is written in the same general format with the ψ function as in [Ebrahimi et al., 2004]. In (4.18) it is rewritten in terms of the binomial coefficient and Harmonic numbers because that format is easier to interpret. Fast conclusions include the following ([Ebrahimi et al., 2004]):

1. The right-hand side of (4.17) indicates that the mutual information between consecutive order statistics does not depend on the distribution F. It depends only on the sample size n and on i. We anticipated this since the positive mutual information is the consequence from the sorting and the selection phases of the definition of the order statistics.

2. The Beta function has the symmetry property $B(\alpha, \beta) = B(\beta, \alpha)$ and the binomial coefficient has the symmetry property $\binom{n}{i} = \binom{n}{n-i}$. Hence, from (4.17) it can be concluded that there is symmetry between i and $n - i$.

3. The mutual information of consecutive order statistics increases when n increases.

The interpretation of (4.18) is worth thinking over. Any process that computes the ordering of a sample can be represented by a decision tree, where each vertex of the tree represents an ordering comparison ('$<$' or '$\geq$'?) of two sample values. The decision tree that is needed to order a sample of size n is a binary tree with $n!$ leaves and height at least $\log_2(n!) = \mathcal{O}(n \log_2 n)$ ([Biggs, 1989, Chapter 9.2],[Cormen et al., 2009, Theorem 8.1]. Recall from (4.15) that $H_n \approx \log n + \gamma_E$. Hence,

$$nH_{n-1} \approx n\log(n-1) + n\gamma_E = \mathcal{O}(n \log n)$$

can be interpreted as the term reflecting the amount of information that is needed to describe the required number of ordering comparisons when the whole sample is ordered. But, when computing $I\left(U_{(i)}, U_{(i+1)}\right)$, it is not necessary to describe the computations of $U_{(j)}$ with $j < i$ or $j > i+1$. The terms '$-iH_{i-1}$' and '$-(n-i)H_{n-i-1}$' in (4.18) reflect this. Hence,

$$nH_{n-1} - iH_{i-1} - (n-i)H_{n-i-1} \approx \log \#[\text{common ordering comparisons}],$$

where the common ordering comparisons are those that are needed to determine both $U_{(i)}$ and $U_{(i+1)}$. The binomial coefficient $\binom{n}{i}$ is the number of ways that a subsample of size i can be selected from a sample of size n, without replacement, and it satisfies the 'Pascal's triangle' recursion formula $\binom{n}{i} = \binom{n-1}{i-1} + \binom{n-1}{i}$ [Biggs, 1989, Chapter 4]. Hence, the term '$-\log\binom{n}{i}$' in (4.18) can be interpreted as the amount of information required to describe the selection of $U_{(i)}$ and $U_{(i+1)}$. Thus, sorting and selection are transparently present in (4.18). In addition to this, the more there are common ordering comparisons done when the mutual information is determined, the more information there is about the location of $U_{(i+1)}$ given $U_{(i)}$ or *vice versa*.

Next, notation I_n is used because different sample sizes n are considered. For example, if $n = 2$ then $I_2\left(U_{(1)}; U_{(2)}\right) = 1 - \log 2 > 0$. This must be a baseline level since, when $n = 2$, sorting and selection are essentially the same process and there are no variables that could act as common causes for both $U_{(1)}$ and $U_{(2)}$. Generally, selection and sorting are not the same algorithmic processes. For example, it is possible to select $U_{(n)}$ and $U_{(n-1)}$ from a sample of size n without sorting all values and, in that algorithm, there are at most $n + \lceil \log n \rceil - 2$ ordering comparisons needed ([Cormen et al., 2009, Chapter 9]) so that the height of the corresponding decision tree should be $\approx \log(n + \lceil \log n \rceil - 2)$. Indeed, the general case can be computed from (4.18):

$$\begin{aligned} I_n\left(U_{(n-1)}; U_{(n)}\right) &= nH_{n-1} - (n-1)H_{n-2} - 1 - \log n \\ &= H_{n-1} - \log n > 0. \end{aligned} \tag{4.19}$$

From the definition of γ_E in (4.15) it follows that $I_n\left(U_{(n-1)}; U_{(n)}\right) \to \gamma_E$, when $n \to \infty$. This was an unexpected result! The next question is whether γ_E is a large or a small amount of information? The value of γ_E in natural units (nats) was already given in (4.15), the value in binary units is $\frac{\gamma_E}{\log 2} \approx 0.832746\ldots$ bits.

The cases $i = n-1, \ldots, n-5$ of $I_n\left(U_{(i)}; U_{(i+1)}\right)$ near the upper tail are collected in Table 4.1. Relative to these other cases in Table 4.1, γ_E is the smallest limiting amount of mutual information that two consecutive order statistics can have. However, restricting to consecutive pairs seems like a limitation and other pairs may also be of interest. In the next section, we continue this argument and ask how much uncertainty $X_{(n-k)}$ can reduce about $X_{(n)}$ when $1 < k < n$.

Table 4.1: Values of $I_n\left(U_{(i)}; U_{(i+1)}\right)$ with some $i < n$ from the upper tail.

Case	Formula	Asymptotic formula
$i = n-1$	$H_{n-1} - \log n$	γ_E
$i = n-2$	$2H_{n-3} + n(H_{n-1} - H_{n-3}) - 3 - \log\binom{n}{n-2}$	$\log 2 + 2\gamma_E - 1$
$i = n-3$	$3H_{n-4} + n(H_{n-1} - H_{n-4}) - \frac{11}{2} - \log\binom{n}{n-3}$	$\log 6 + 3\gamma_E - \frac{5}{2}$
$i = n-4$	$4H_{n-5} + n(H_{n-1} - H_{n-5}) - \frac{25}{3} - \log\binom{n}{n-4}$	$\log 24 + 4\gamma_E - \frac{13}{3}$
$i = n-5$	$5H_{n-6} + n(H_{n-1} - H_{n-6}) - \frac{137}{12} - \log\binom{n}{n-5}$	$\log 120 + 5\gamma_E - \frac{77}{12}$

4.1.3 Mutual information between $X_{(i)}$ and $X_{(n)}$

In publication [2], we use several estimators to estimate the *Extreme Value Index (EVI).* These include the Hill estimator $\mathsf{H}_{k,n}$. The Hill estimator utilizes the information of the k largest order statistics $X_{(n-k)}, \ldots, X_{(n)}$. It is defined as ([Hill, 1975],[Beirlant et al., 2004], [Németh and Zempléni, 2020])

$$\mathsf{H}_{k,n} = \frac{1}{k}\sum_{j=1}^{k} \log X_{(n-j+1)} - \log X_{(n-k)}. \tag{4.20}$$

It is known ([Beirlant et al., 2004],[Caers and Van Dyck, 1999]) that the Hill estimator is consistent and asymptotically normally distributed if $k = k(n)$ depends on n in such a way that

$$k(n) \to \infty \qquad \text{and} \qquad \frac{k(n)}{n} \to 0 \qquad \text{when} \qquad n \to \infty \tag{4.21}$$

A large class of possible forms of $k = k(n)$ that satisfy (4.21) is obtained by $k(n) = Cn^a$, $0 < a < 1$ and $C > 0$ a constant ([Caers and Van Dyck, 1999]). A possible value for k, that is usually searched for, is such that $\mathsf{H}_{k,n}$ is approximately a constant near k and (4.21) can be assumed to hold. It may be visually detected from the Hill plot, which is the plot of pairs

$$(k, \mathsf{H}_{k,n}), \qquad 1 \le k \le n-1. \tag{4.22}$$

Publication [2, Figure 1] contains two examples of the Hill plot, for observed TCP flow sizes and flow durations. The interpretation of the Hill plot can be difficult and it is far from easy to select k from the plot. In publication [2, Section 2.2], we used bootstrapping ([Hall, 1990], [Caers and Van Dyck, 1999]) to select one k for each block.

In Figure 4.1, $m = 61$ Hill plots for flow sizes S are shown in gray color, and the black curve is the average of the m Hill plots: $\frac{1}{m} \sum \mathsf{H}_{k,n}$. The block size $n = 10\,000$, but the horizontal axis is restricted to $k \leq 5\,000$.

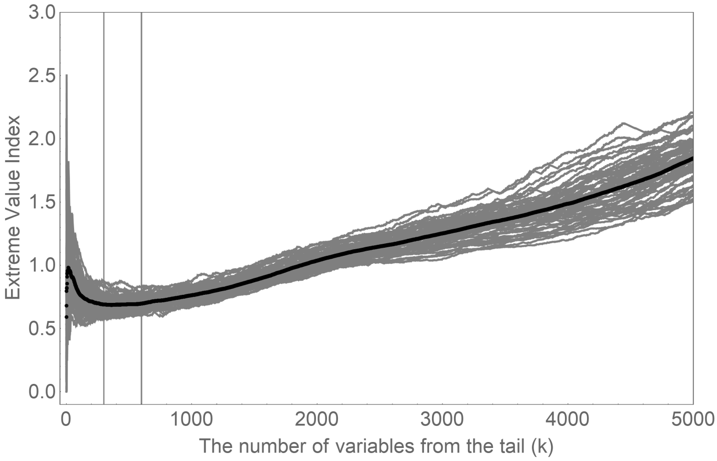


Figure 4.1: Hill plots of $m = 61$ blocks in gray. The black curve is the average of them. The red vertical lines are at $k = 300$ and $k = 600$.

The black average curve in Figure 4.1 is visually flat when $300 \leq k \leq 600$ indicating that in this interval $\mathsf{H}_{k,n}$ may have an average value independent of k. Numerically, the first 3 decimals of the average appear stable: $\frac{1}{m} \sum \mathsf{H}_{k,n} \approx 0.688\ldots$. If only one sample (block) were available, then finding such a flat portion visually would be hard in practice. Bootsrapping solves the problem giving a good candidate for k, but we still ask why such a k exists? The Hill estimator (4.20) is expressed in terms of order statistics and there is the log transformation which is order preserving. The asymptotic condition (4.21) is very general. There is no clue about which k should be selected. Hill [Hill, 1975] already discussed the selection of k as a problem of high threshold selection: $X_{(n-k)}$ is a data-based estimate of a high threshold and Pickands [Pickands, 1975] had shown that for values X that exceed a fixed high threshold x_0, the distribution of $X - x_0$ converges to the Generalized Pareto distribution. Hence, k should be such that $X_{(n-k)} > x_0$.

For this purpose, we compute $I_n\left(X_{(i)}; X_{(n)}\right) = I_n\left(U_{(i)}; U_{(n)}\right)$, $i < n$, to get more insight into why (4.21) is needed for the asymptotical results. The computation is given in Appendix A.1 and the result is

$$I_n\left(U_{(i)}; U_{(n)}\right) = H_{n-1} - \log n - [H_{n-i-1} - \log(n-i)]. \tag{4.23}$$

Substituing $i = n - k$ to (4.23) gives

$$I_n\left(U_{(n-k)}; U_{(n)}\right) = H_{n-1} - \log n - (H_{k-1} - \log k). \tag{4.24}$$

It seems that the easiest is to visualize this by plotting $k \mapsto I_n\left(U_{(n-k)}; U_{(n)}\right)$ as shown in Figure 4.2 with $n = 10\,000$ so that it could be compared with Figure 4.1. Note that the vertical axis in Figure 4.2 is in log scale. It seems that the flat region in Figure 4.1 occurs since, after $k > 300$, there is practically no mutual information left. After $k > 600$, the cumulative effect of rounding errors or noise start to distort the Hill plot curves in Figure 4.1.

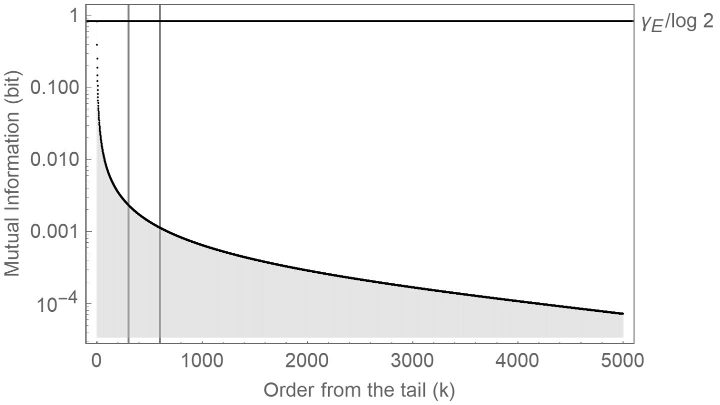


Figure 4.2: Plot of $k \mapsto I_n\left(U_{(n-k)}; U_{(n)}\right)$ with $n = 10\,000$. The vertical axis is in log scale and the vertical unit is bits.

In publication [2, Section 2.2], the bootstrapping provided values of k varying from 129 to 302 when the data was the same flow sizes S as in Figure 4.1. (The bootstrapping computation was saved in a notebook. This piece of information was not included in [2].) We conclude that, as long as the data is representative, bootstrapping is really able to squeeze out all relevant information from the data!

4.1.4 Order statistics are empirical quantiles

In the publications of this thesis, we utilize both *quantile-quantile plots* (*qq-plots*) and *probability-probability plots* (*pp-plots*). In the former, the quantiles of two distributions are visually compared. In the latter, the cumulative probabilites of the data are compared against the cumulative probabilities of a model of the data. Stuart Coles says in his book [Coles, 2001] that the pp-plot and the qq-plot, when done for the same data and model, contain the same information but it is expressed on a different scale. However, they are complementary methods, they add value to each other, rather than alternative tools.

Assume now that a continuous model candidate F is selected and $x_1, \dots, x_n$ are the data and $x_{(1)}, \dots, x_{(n)}$ are the ordered data. While the expectation $\mathbb{E}(X_{(i)})$ may be computed or approximated numerically from (4.2), the PIT version can be computed exactly and $\mathbb{E}\left(F(X_{(i)})\right) = \mathbb{E}(U_{(i)}) = \frac{i}{n+1}$, since $U_{(i)} \sim Beta(i, n-i+1)$. The pp-plot is the plot of pairs

$$\left(F(x_{(i)}), \frac{i}{n+1}\right), \qquad i = 1, \dots, n. \tag{4.25}$$

If model F is correct the linear shape is then expected in the pp-plot. The qq-plot is the plot of the pairs

$$\left(F^{-1}\left(\frac{i}{n+1}\right), x_{(i)}\right), \qquad i = 1, \ldots, n. \tag{4.26}$$

Two more definitions are needed to proceed. The *empirical distribution* can be written with the order statistics as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}} = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_{(i)} \leq x\}} = \begin{cases} 0, & x < X_{(1)} \\ \frac{i}{n}, & X_{(i)} \leq x < X_{(i+1)} \\ 1, & x \geq X_{(n)}. \end{cases} \tag{4.27}$$

It is immediate that $F_n(X_{(i)}) = \frac{i}{n}$. It is a piecewise constant function and not invertible but, if *the generalized inverse function* is used

$$F^{-1}(q) = \inf\{x \mid F(x) \geq q\} \tag{4.28}$$

then $F_n^{-1}(q) = X_{(i)}$ for all $\frac{i-1}{n} < q \leq \frac{i}{n}$. A special case of this is $F_n^{-1}\left(\frac{i}{n+1}\right) = X_{(i)}$ since $\frac{i-1}{n} < \frac{i}{n+1} < \frac{i}{n}$. Order statistics are the quantiles of the empirical distribution function computed from the data. Now the qq-plot (4.26) can be rewritten as the plot of the pairs

$$\left(F^{-1}\left(\frac{i}{n+1}\right), F_n^{-1}\left(\frac{i}{n+1}\right)\right), \qquad i = 1, \ldots, n, \tag{4.29}$$

from which it is easier to see that a linear shape in (4.26) indicates that the empirical quantiles are linearly related to the model quantiles. The linear relationship between the probabilities (4.25) is a different concept than the linear relationship between the quantiles (4.26), hence these methods complement each other.

The pp-plots and qq-plots were developed for small samples. The “small” sample means the case where there is not enough data to draw a histogram. A clear benefit of pp- and qq-plots is that a human eye catches clear deviation from linear shapes easily. An algorithmic method to quantify a lack of linearity is obtained by computing the linear correlation coefficient for the data in the plots (4.25) and (4.26). This is a well-known method ([Johnson and Wichern, 2007, Chapter 4.6],[Filliben, 1975]) that we also used in publication [1].

Brown and Hettmansberger [Brown and Hettmansperger, 1996] introduced *the plotting positions* that we used in *the normal-quantile plots (nq-plots)* of [1]. They were selected since they extend further into the tails and the possible lack of fit in the upper tail of the Gaussian model was of importance in publication [1]. *Normal-probability plots (np-plots)* were also made during the research work of [1], actually we started with them, although the np-plots were not included in the article.

In publication [1] we also made qq-plots of *ordered data against ordered data.* This is discussed more in Section 4.3 since that issue is related to time dependence. However, if the data sets have different sizes $n_1 \neq n_2$, then the quantiles need to be reconsidered and this is discussed next.

4.1.5 Estimating the q:th quantile

Assume $0 < q < 1$ and let x_q be the q:th quantile of a distribution F, $F(x_q) = q$. A quantile need not be unique, hence assume that F is strictly increasing at least in the neighborhood of x_q so that it is unique. Let $\lceil x \rceil = \min\{k \in \mathbb{Z} \mid k \geq x\}$ be *the ceiling function.* The estimator $F_n^{-1}(q)$ of the quantile x_q can be written as the $\lceil qn \rceil$:th order statistic $X_{(\lceil qn \rceil)}$:

$$F_n^{-1}(q) = X_{(\lceil qn \rceil)}. \tag{4.30}$$

Assume that F has a density f in a neighborhood of x_q, that $f(x_q) > 0$ and f is continuous at x_q, then it is well known ([Serfling, 1980]) that $X_{(\lceil qn \rceil)}$ is asymptotically normally distributed with parameters

$$N\left(x_q, \frac{q(1-q)}{f(x_q)^2 n}\right). \tag{4.31}$$

This result is useful unless $f(x_q) = F'(x_q) \approx 0$, which is typical in the tail area.

One of the simplest statistical ways to reduce univariate data is to first select m, then select $0 < q_1 < \ldots < q_m < 1$, compute indexes $\lceil q_j n \rceil$, and then select the values $x_{(\lceil q_j n \rceil)}$ from the sorted data. For example, qq-plots can be made at predefined probabilities q_j, $j = 1, \ldots, m$, and this solves the problem of different sample sizes n_1 and n_2 when $m < \min\{n_1, n_2\}$. In publication [4, formula (15) and Section 8.3], we show that estimates of the sample mean and the sample variance can be computed from the m quantile estimators and the minimum and maximum observations of the sample.

4.1.6 Simultaneous estimation of several quantiles

In publication [4, Section 8], we state and apply Theorem 1 below but without proof. In this section we provide the proof of Theorem 1.

Theorem 1. *Let $0 < q_1 < \ldots < q_m < 1$. If $X_1, \ldots, X_n$ are i.i.d., $X_i \sim F$ where the q_j quantiles of F are x_{q_j} and F is strictly increasing in the neighborhoods of each x_{q_j}, then for all $\varepsilon > 0$*

$$\mathbb{P}\left\{\left|\frac{1}{m}\sum_{j=1}^{m}\left(X_{(\lceil q_j n \rceil)} - x_{q_j}\right)\right| > \varepsilon\right\} \leq 2\sum_{j=1}^{m} e^{-2n\delta_{\varepsilon,j}^2} \tag{4.32}$$

where

$$\delta_{\varepsilon,j} = \min\left\{F(x_{q_j} + \varepsilon) - q_j, q_j - F(x_{q_j} - \varepsilon)\right\}. \tag{4.33}$$

The proof of Theorem 1 is based on the following Theorem 2, which in Serfling's book [Serfling, 1980] is attributed to Wassily Hoeffding [Hoeffding, 1963].

Theorem 2. *Assume that $Y_1, \ldots, Y_n$ are i.i.d. random variables with $\mathbb{P}\{0 \leq Y_i \leq 1\} = 1$ and $\mathbb{E}(Y_i) = \mu$ for all $i = 1, \ldots, n$. Then, for all $t > 0$, we have the estimates*

$$\mathbb{P}\left\{\sum_{i=1}^{n} Y_i - n\mu \geq nt\right\} \leq e^{-2nt^2}, \tag{4.34}$$

$$\mathbb{P}\left\{\sum_{i=1}^{n} Y_i - n\mu \leq -nt\right\} \leq e^{-2nt^2}. \tag{4.35}$$

For the proof of Theorem 2, we refer to Hoeffding's original article [Hoeffding, 1963] or to the book [Devroye et al., 1996].

Proof of Theorem 1. Let $\varepsilon > 0$. If $\left|(X_{(\lceil q_j n\rceil)} - x_{q_j}\right| \leq \varepsilon$ for every $j = 1, \dots, m$, then

$$\left|\frac{1}{m}\sum_{j=1}^{m} X_{(\lceil q_j n\rceil)} - x_{q_j}\right| \leq \frac{1}{m}\sum_{j=1}^{m}\left|(X_{(\lceil q_j n\rceil)} - x_{q_j}\right| \leq \frac{m\varepsilon}{m} = \varepsilon.$$

Changing the focus on the complements, there is the following inclusion of the events

$$\left\{\left|\frac{1}{m}\sum_{j=1}^{m} X_{(\lceil q_j n\rceil)} - x_{q_j}\right| > \varepsilon\right\} \subseteq \bigcup_{j=1}^{m}\left\{|X_{(\lceil q_j n\rceil)} - x_{q_j}| > \varepsilon\right\},$$

and an inequality of the probabilities of the events as follows

$$\mathbb{P}\left\{\left|\frac{1}{m}\sum_{j=1}^{m} X_{(\lceil q_j n\rceil)} - x_{q_j}\right| > \varepsilon\right\} \leq \sum_{j=1}^{m}\mathbb{P}\left\{|X_{(\lceil q_j n\rceil)} - x_{q_j}| > \varepsilon\right\}. \tag{4.36}$$

After this observation we follow the proof from Chapter 2.3.2 in Serfling's book [Serfling, 1980]. (Serfling attributes the proof technique to Smirnov [Smirnov, 1949].) First, we choose an arbitrary $j \in \{1, \dots, m\}$. Then we remove the absolute values to get two mutually exclusive cases

$$|X_{(\lceil q_j n\rceil)} - x_q| > \varepsilon \quad \Leftrightarrow \quad X_{(\lceil q_j n\rceil)} > x_q + \varepsilon \quad \text{or} \quad X_{(\lceil q_j n\rceil)} < x_q - \varepsilon.$$

We go through the first case in detail. Apply F_n to both sides of the inequality $X_{(\lceil q_j n\rceil)} > x_{q_j} + \varepsilon$ and compute

$$\begin{aligned}\frac{\lceil q_j n\rceil}{n} = F_n(X_{(\lceil q_j n\rceil)}) > F_n(x_{q_j} + \varepsilon) &= \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x_{q_j}+\varepsilon\}} = \frac{1}{n}\sum_{i=1}^{n}\left(1 - 1_{\{X_i > x_{q_j}+\varepsilon\}}\right) \\ &= 1 - \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i > x_{q_j}+\varepsilon\}}.\end{aligned}$$

The strict inequality is maintained since F_n has a jump at $X_{(\lceil q_j n\rceil)}$ and, since the jump size is $\frac{1}{n} \geq \frac{\lceil q_j n\rceil}{n} - q_j$, equivalently $q_j \geq \frac{\lceil q_j n\rceil}{n} - \frac{1}{n}$, there is the following inclusion of the events

$$\left\{X_{(\lceil q_j n\rceil)} > x_{q_j} + \varepsilon\right\} \subseteq \left\{q_j > 1 - \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i > x_{q_j}+\varepsilon\}}\right\}. \tag{4.37}$$

Multiplying boths sides with $-n < 0$ in the larger event above allows to rewrite its condition as

$$\sum_{i=1}^{n} 1_{\{X_i > x_{q_j} + \varepsilon\}} > n - q_j n = n(1 - q_j). \tag{4.38}$$

Since $X_i \sim F$,

$$\mathbb{E}\left(1_{\{X_i > x_{q_j} + \varepsilon\}}\right) = \mathbb{P}\left\{X_i > x_{q_j} + \varepsilon\right\} = 1 - F\left(x_{q_j} + \varepsilon\right).$$

We add the term $-n(1 - F(x_{q_j} + \varepsilon))$ to both sides of the inequality (4.38) and then we rewrite it as

$$\sum_{i=1}^{n} 1_{\{X_i > x_{q_j} + \varepsilon\}} - n\left(1 - F\left(x_{q_j} + \varepsilon\right)\right) \geq n\left(F\left(x_{q_j} + \varepsilon\right) - q_j\right). \tag{4.39}$$

Next, we apply the inequality (4.34) of Hoeffding's Theorem 2 with $Y_i = 1_{\{X_i > x_{q_j} + \varepsilon\}}$ and $t = F\left(x_{q_j} + \varepsilon\right) - q_j$. Then

$$\mathbb{P}\left\{X_{(\lceil q_j n \rceil)} > x_{q_j} + \varepsilon\right\} \leq e^{-2n(F(x_{q_j}+\varepsilon) - q_j)^2}$$

Repeating the same reasoning with the case $X_{(\lceil q_j n \rceil)} < x_{q_j} - \varepsilon$ gives

$$\sum_{i=1}^{n} 1_{\{X_i < x_{q_j} - \varepsilon\}} - nF\left(x_{q_j} - \varepsilon\right) \leq \left(q_j - F\left(x_{q_j} - \varepsilon\right)\right) n, \tag{4.40}$$

and then we use the inequality (4.35) of Theorem 2.

Next, combining the two cases gives

$$\begin{aligned} \mathbb{P}\left\{|X_{(\lceil q_j n \rceil)} - x_{q_j}| > \varepsilon\right\} &\leq \mathbb{P}\left\{X_{(\lceil q_j n \rceil)} < x_{q_j} - \varepsilon\right\} + \mathbb{P}\left\{X_{(\lceil q_j n \rceil)} > x_{q_j} + \varepsilon\right\} \\ &\leq e^{-2n(q_j - F(x_{q_j} - \varepsilon))^2} + e^{-2n(F(x_{q_j}+\varepsilon) - q_j)^2} \\ &\leq 2e^{-2n\delta_{\varepsilon,j}^2}, \end{aligned} \tag{4.41}$$

in which

$$\delta_{\varepsilon,j} = \min\left\{F\left(x_{q_j} + \varepsilon\right) - q_j, q_j - F\left(x_{q_j} - \varepsilon\right)\right\}. \tag{4.42}$$

In the last step, we apply the above reasoning for all $j = 1, \ldots, m$ to obtain (4.32). The proof of Theorem 1 is now finished. □

Theorem 1 is the core element of publication [4]. This extension to multiple simultaneous quantiles required only (4.36). There is a separate $\delta_{\varepsilon,j}$ for all j, a common value $\delta_\varepsilon = \min_j \delta_{\varepsilon,j}$ does not make sense. The interpretation of $\delta_{\varepsilon,j}$ is that the smaller it is, the more uncertain the estimation of x_{q_j} is. The $\delta_{\varepsilon,j}$ is small if the distribution function F is almost flat near x_{q_j}. Theorem 1 states that, in this sense, the most uncertain quantile affect the convergence of all simultaneous quantile estimates. If the distribution is multimodal, then between the modes there can be regions where F is almost flat. If the distribution has long or heavy tails, then in the tail area the distribution function F is almost flat.

4.1.7 The subexponential class of distributions

In publication [2], we do not introduce any distribution class to model heavy-tailed data. However, in the introductory section of [2] we used the subexponential class as a motivating example. Next, we discuss this model class more specifically.

Let F be a distribution function in $]0,\infty[$ and $F*F = F^{*2}$ be the convolution product defined as

$$F * F(x) = \int_0^x F(x-y)dF(y).$$

The class of *subexponential distribution functions* satisfy

$$\lim_{x\to\infty} \frac{1-F^{*2}(x)}{1-F(x)} = 2 \tag{4.43}$$

or, equivalently,

$$2(1-F(x)) \sim 1-F^{*2}(x) \qquad \text{when} \qquad x \to \infty.$$

If this holds for $n = 2$, then it holds for all $n \geq 2$ ([Chistyakov, 1964])

$$n(1-F(x)) \sim 1-F^{*n}(x) \qquad \text{when} \qquad x \to \infty. \tag{4.44}$$

Chistyakov [Chistyakov, 1964] appears to be the first who proved that the property (4.43) is equivalent to (4.44). A more advanced theory about the class of subexponential distribution functions is given in [Kluppelberg, 1988], [Kluppelberg, 1989]. The Pareto and the lognormal distributions belong to the class of subexponential distributions and, for the lognormal distribution, the proof of this fact seems to require the general theory developed in [Kluppelberg, 1988] or by a sufficient criterion given in [Pitman, 1980] and used in [Samorodnitsky, 2002].

The class of the subexponential distribution functions is not mathematically convenient since, for example, it is not closed under summation [Leslie, 1989]. The class of distributions with regularly varying right tails with exponent $\theta > 0$ is generally used if some properties need to be proved. For example, a direct proof that the Pareto distribution $F(x) = 1 - \left(\frac{k}{x}\right)^{\alpha}$ is subexponential by using (4.43) is surprisingly difficult, but an easy way is obtained via reqularly varying tails. The Pareto distribution has a regularly varying right tail with exponent α and, for example, [Samorodnitsky, 2002] contains a short proof sketch that a distribution with a regularly varying right tail is subexponential.

Assume $X_1, \ldots, X_n$ are i.i.d. random variables with a common subexponential distribution function F. The asymptotic condition (4.44) can be interpreted as

$$\mathbb{P}\{X_{(n)} > x\} \sim \mathbb{P}\{X_1 + \ldots + X_n > x\} \qquad \text{when} \qquad x \to \infty.$$

Since $\{X_{(n)} > x\} \subset \{X_1 + \ldots + X_n > x\}$, the asymptotic equality has an important interpretation that the sum is going to exceed a large threshold x because one of the summands is going to exceed the threshold. This is the opposite to what the CLT states and, in practise, it occurs in our study [2] because the flow sizes S can have vastly different magnitudes, from 1 kB to 42 MB. It occurs in our study [1] in the upstream direction since the control traffic of a TCP flow has a small magnitude contribution to the aggregate rate (an ACK packet contributes $40\mathrm{B}/\Delta$) compared to the rate magnitudes of data bursts that can be thousands of bytes (a single 1500B packet contributes $1500\mathrm{B}/\Delta$).

4.2 Multivariate analysis and models

In publication [2], we analyse the bivariate distribution of TCP flow sizes and flow durations. In publication [3], we model the joint distribution of vehicular traffic volumes or asymmetry in two mutually relevant locations by a binormal distribution. In publication [4], we estimate several quantiles simultaneously and, in the background of this approach, there is an asymptotic multinormal joint distribution of the estimators. Next, we will discuss each of these topics more.

4.2.1 Bivariate and heavy-tailed distributions

If the TCP flow size S is large and the access link rate is limited, then the sending window of the TCP source cannot be large and the flow duration D must be large but the opposite need not hold: for various reasons D can be relatively large even when S is relatively small. Therefore, we expected a rather strong dependence for the pair (S, D) beforehand.

An asymptotic bivariate distribution G of normalized maxima (S^*, D^*) can be represented by its margins G_1 and G_2 by

$$G(x,y) = \exp\left(\log\left[G_1(x)G_2(y)\right] A\left(\frac{\log G_2(y)}{\log\left[G_1(x)G_2(y)\right]}\right)\right) \tag{4.45}$$

where $A(t)$, $t \in [0, 1]$ is called *the Pickands dependence function* ([Beirlant et al., 2004]). The normalization (S^*, D^*) mentioned above is implicit since the generalized extreme value (GEV) distribution has location and scale parameters and, assuming that the GEV margins are good approximations, the normalization is included in the estimated parameters [Coles, 2001, Theorem 3.1.1]. The book [Beirlant et al., 2004] is a good source of further information about Pickand's dependence function. The function $A(t)$ satisfies $A(0) = A(1) = 1$, it is convex and lies inside the triangle determined by points $(0, 1)$, $(1, 1)$ and $(0.5, 0.5)$. This triangle is shown in publication [2, Fig. 7] as dashed lines. Cases $A(t) \equiv 1$ and $A(t) = \max\{1 - t, t\}$ correspond to independence and total dependence between S^* and D^*, respectively.

The main reason to choose this approach was that there exists at least two non-parametric estimators for A. We refer to publication [2] or to the original references [Capéraà et al., 1997] and [Hall and Tajvidi, 2000] for information about these estimators. Non-parametric estimators were crucial since we did not know beforehand what A could or should look like. Once we obtained a good guess of the shape of A from the nonparametric estimates, we could find a parametric model of A with a similar shape.

This was done in publication [2] where, for that specific data, we found *the logistic model* $A_r(t) = ((1-t)^r + t^r)^{1/r}$ with the parameter value $r = 2$ as a possible model.

4.2.2 Binormal and trinormal distributions

In publication [3], we apply binormal and trinormal distribution models [Kotz et al., 2000] for pairs or triples of asymmetries of two or three locations. Here we provide reasons why in publication [3] we employed $\mathbb{E}(Z_2|Z_1 > a)$ instead of customary $\mathbb{E}(Z_2|Z_1 = a)$. We had three reasons for this.

First, vehicular traffic counts have a non-linear interpretation: the counts of vehicles are highest when the traffic is fluent but if the count is low it can be either due to a small amount of traffic or due to so large an amount of traffic that congestion slows down the traffic flow over the loop detector. The interpretation requires additional information from other sources.

Second, conditioning by an event with positive probability should be more robust and an event of type $\{Z_1 > a\}$ usually has positive probability. For such values of a that make sense, both the model probability and the empirical probability of the event $\{Z_1 > a\}$ are typically positive.

Third, this formulation is already a simple prediction: Consider a 15 minute time slot. If after 5 minutes $Z_1 = a$ is observed, then $Z_1 > a$ is likely to be true at the end of the time slot and the conditional expectation gives an immediate and directly interpretable prediction of Z_2 at the end of the slot. The prediction is provided with confidence intervals.

After the volume-asymmetry transformation in publication [3], the binormal distribution is a natural model for the asymmetry at two different but mutually relevant locations since the correlation is due to the same vehicles observed in the two locations and the directions are meant to be chosen in such a way that there can be a causal explanation for the correlation. Here is actually also a connection to the mutual information: if ρ is the correlation between two normally distributed variables Z_1 and Z_2, then it is well-known ([Kullback, 1968],[Cover and Thomas, 2006, Example 8.5.1]) and straightforward to compute that $I(Z_1; Z_2) = -\frac{1}{2}\log(1-\rho^2)$.

4.2.3 Asymptotic joint distribution of quantile estimates

In publication [4], the OPE algorithm enters regularly into *the model building phase* [4, Figure 1] and a new model building always starts by first trying to get a good estimate for the median and then, once the median estimate appears satisfactory, it proceeds to check that the estimates of the upper and the lower quartiles are sufficiently good. There is a justification for this that we explain next.

Suppose F has a continuous density f in neighborhoods of x_{q_i} and x_{q_j}, $f(x_{q_i}) > 0$ and $f(x_{q_j}) > 0$. Then the asymptotic covariance is $Cov\left(X_{(\lceil q_i n \rceil)}, X_{(\lceil q_j n \rceil)}\right) = \sigma_{ij}/n$, where, for $i \leq j$

$$\sigma_{ij} = \frac{q_i(1-q_j)}{f(x_{q_i})f(x_{q_j})}. \tag{4.46}$$

and $\sigma_{ij} = \sigma_{ji}$ for $i > j$ [Serfling, 1980].

Choose $m = 3$ and $0 < q_1 < q_2 < q_3 < 1$. According to (4.46), the asymptotic covariance matrix Σ of the estimators $X_{(\lceil q_1 n \rceil)}$, $X_{(\lceil q_2 n \rceil)}$ and $X_{(\lceil q_3 n \rceil)}$ is

$$\Sigma = \frac{1}{n} \begin{pmatrix} \frac{q_1(1-q_1)}{f(x_{q_1})^2} & \frac{q_1(1-q_2)}{f(x_{q_1})f(x_{q_2})} & \frac{q_1(1-q_3)}{f(x_{q_1})f(x_{q_3})} \\ \frac{q_1(1-q2)}{f(x_{q_1})f(x_{q_2})} & \frac{(1-q_2)q_2}{f(x_{q_2})^2} & \frac{q_2(1-q_3)}{f(x_{q_2})f(x_{q_3})} \\ \frac{q_1(1-q_3)}{f(x_{q_1})f(x_{q_3})} & \frac{q_2(1-q_3)}{f(x_{q_2})f(x_{q_3})} & \frac{(1-q_3)q_3}{f(x_{q_3})^2} \end{pmatrix}. \tag{4.47}$$

By (4.31) and (4.46), asymptotically

$$\left(X_{(\lceil q_1 n \rceil)}, X_{(\lceil q_2 n \rceil)}, X_{(\lceil q_3 n \rceil)}\right) \sim N_3(\mu, \Sigma), \tag{4.48}$$

where $\mu = (x_{q_1}, x_{q_2}, x_{q_3})$. Assume now that the asymptotic trinormal distribution is the true distribution. The inverse of the covariance matrix, sometimes called *the concentration matrix* $K = \Sigma^{-1}$ is

$$K = n \begin{pmatrix} \frac{q_2 f(x_{q_1})^2}{q_1 q_2 - {q_1}^2} & -\frac{f(x_{q_1})f(x_{q_2})}{q_2 - q_1} & 0 \\ -\frac{f(x_{q_1})f(x_{q_2})}{q_2 - q_1} & \frac{f(x_{q_2})^2(q_3-q_1)}{(q_2-q_1)(q_3-q_2)} & -\frac{f(x_{q_2})f(x_{q_3})}{q_3-q_2} \\ 0 & -\frac{f(x_{q_2})f(x_{q_3})}{q_3-q_2} & \frac{(1-q_2)f(x_{q_3})^2}{(1-q_3)(q_3-q_2)} \end{pmatrix}. \tag{4.49}$$

The important thing in this concentration matrix $K = (k_{ij})$ is that $k_{13} = k_{31} = 0$. This has the well-known effect ([Højsgaard et al., 2012]) that the joint trinormal density ϕ,

$$\phi(\mathbf{x}) = \frac{1}{(2\pi)^{3/2}\sqrt{det(\Sigma)}} e^{\frac{1}{2}(\mathbf{x}-\mu)^T K (\mathbf{x}-\mu)}, \tag{4.50}$$

splits into the product $\phi(x_1, x_2, x_3) = g(x_1, x_2)h(x_2, x_3)$ at every point $\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$, where g and h are functions. By Dawid's work [Dawid, 1979], (4.49) means that $X_{(\lceil q_1 n \rceil)}$ and $X_{(\lceil q_3 n \rceil)}$ are *conditionally independent* given the estimator $X_{(\lceil q_2 n \rceil)}$ which is in the middle of them,

$$X_{(\lceil q_1 n \rceil)} \perp\!\!\!\perp X_{(\lceil q_3 n \rceil)} \Big| X_{(\lceil q_2 n \rceil)}. \tag{4.51}$$

If it is assumed that the asymptotic trinormal distribution is the true distribution of the triple (4.48), then the above is actually the proof of the conditional independence. Conditional independence means that knowing the value of the middle quantile estimator blocks all information flow between the the two other quantile estimators that are on the opposite sides of the middle estimator. The quartiles are an example of such conditional independence and hence, the model building process in [4] always starts with the median estimate.

Indeed, the conditional independence (4.51), when $q_1 < q_2 < q_3$, is very intuitive and may be true more generally. The uniqueness of the quantiles could be a sufficient criterion since a counterexample may exist if x_{q_2} is not unique.

4.3 Time dependence

Next, we will discuss some statistical time dependence issues that were relevant to us when writing the publications.

There are two possibilities for the statistical time dependence: 1) non-stationary data with recurrent patterns, and 2) autocorrelated stationary data. In publications [1] and [3] we consider slotted time and the variable of interest is the amount of traffic per time slot.

Vehicular traffic data in publication [3] is non-stationary in the sense that if we want to predict the amount of traffic in the next slot, given the amount of traffic in the current slot, we might get the best prediction by just looking at the time when the next slot begins rather that the amount of traffic in the current slot. The 15-minute length of a time slot makes sense for the vehicular traffic but it is too coarse for a stationary model as abrupt time dependent changes can occur in this granularity level.

In publication [1], we have a stationary but positively autocorrelated data. Then, if we want to predict the amount of traffic in the next time slot given the current and some history of previous slots, the time does not matter. Instead, the amount of traffic in the current and the previous slots may contain all information that is useful for prediction. Prediction under a LRD correlated stationary input in the teletraffic context has been studied in [Mannersalo, 2002].

In real data, both cases need to be considered and the choice of the length of the time slot matters. Daily, weekly, monthly, and yearly profiles are always present in human activity. The daily traffic profile, and other recurrent patterns in the data, may be estimated and the estimated profiles may be included in the model. In shorter intervals, the assumption of a stationary model makes things so much easier that it is worth doing unless the data clearly does not support it.

In publication [1], we used a well-known method, in which we compared the empirical distributions of two disjoint time intervals by qq-plots to see if the assumption of a stationary data is plausible. More precisely, we had more than $2n$ obsevations and we took two data samples $(x_1, \ldots, x_n)$ and $(x_{1+h}, \ldots, x_{n+h})$ of consecutive observations, where $1 + h > n$, and made a qq-plot with ordered data against ordered data [1, Section V.A]. This method is one of the many choices to detect deviations from the strict or complete form of stationary time series [Priestley, 1982, Chapter 3]. It is known that if a Gaussian process has time invariant mean and variance functions, then the process is completely stationary [Priestley, 1982, Chapter 3.4.1].

Deviations from strict stationary assumption may thus be detected with an elementary qq-plot method. However, traffic modelling research usually assumes covariance stationary (also called weak, or second order stationary) time series since this concept is mathematically more convenient. The sample autocorrelation function (ACF) may indicate deviations from any assumed covariance stationary model for data but there can be a multitude of reasons for deviations. In this way, the sample ACF may serve as a diagnostic tool. However, if the sample ACF is used as an inference tool about the autocorrelation, then one needs to first assure oneself by some other methods that the assumption of a stationary time series is feasible.

In publication [2, Section 3.2], we used a modified sample ACFs as a diagnostic tool. The ‘heavy-tail’ modification taken from [Resnick, 1997] aims at diagnosing heavy-tailed time series data. Our major argument is based on comparing the heavy-tailed time series data against the same data in random order [2, Fig. 3].

Time dependence affects the estimation of the parameters of a model. In the stationary case, a sample of size n positively correlated observations contain *less* information about the parameters of the model distribution for the data than an independent sample of the same size n would contain [Beran, 1994, Chapter 1]. Estimates that are computed from negatively correlated observations may converge most rapidly to the corresponding model values. Truly negative autocorrelation at some lag strongly suggests a periodic structure in the data. In a non-stationary case, we should first understand all relevant periodic or recurrent patterns that are in the data and, after this, we should collect n observations of each of the patterns. This is vastly more data than what is needed in the stationary case.

4.3.1 Vertical and horizontal traffic aggregation

The possibility of Gaussian traffic models for data traffic was discussed in Section 13 of the book [Roberts et al., 1996]. The research work reported in [Roberts et al., 1996] focused on *Asynchronous Traffic Mode (ATM)* technology, which divided data traffic into ATM cells, bursts of cells and calls. From the TCP/IP traffic modelling point of view, these hierarchical concepts can be re-interpreted as follows. ATM cells can be replaced by TCP/IP-packets, bursts of ATM cells replaced by TCP flows where the source injects a number of packets and waits for the time correponding to the RTT before it injects a new burst of packets, and an ATM call can be interpreted as a TCP connection correspondingly. Therefore, the discussion in Section 13 of [Roberts et al., 1996] about traffic models is still valid after this re-interpretation. Only the scale of the amount of bytes is different.

In Section 13.3.4 of [Roberts et al., 1996] the concepts “aggregation in time” and “aggregation in space” were mentioned. These are the “horizontal” and the “vertical” aggregations of publication [1], respectively. Vertical aggregation is basically CLT, but horizontal aggregation is more complicated. We already indicated in Chapter 3 that RTT, or the distribution of RTTs of TCP connections, may affect the horizontal aggregation if the time scale of interest is smaller than typical RTTs. In publication [1, Section V.C] we made a data-based synthetic vertical aggregation to study how much traffic aggregation would be needed in the smaller time scales, less than $\Delta = 128$ ms, where normal distribution approximation could be ruled out even before testing. In this argument we assumed ‘infinite capacity’, that is, that the variability of the RTTs does not increase, but it is very likely that the variability of RTTs increases when TCP traffic is aggregated more.

4.3.2 Long range dependence and self-similarity

Willinger and Park [Park and Willinger, 2000] provided a telecommunication oriented discussion about (asymptotic) self-similarity and long range dependence (LRD). These concepts belong to the background knowledge of publication [1] so we will briefly discuss them here.

A continuous time stochastic process $Y(t)$ is considered as a model for the cumulative aggregate traffic and such a model is self-similar with self-similarity parameter H if for all $a > 0$ and $t \geq 0$ the scaled and normalized version $a^{-H}Y(at)$ has the same distribution as $Y(t)$, $Y(t) =_d a^{-H}Y(at)$. A self-similar process cannot be stationary but it can have stationary increments. If the cumulative process $Y(t)$ has stationary increments, finite variance and second order stationarity is assumed, then the covariance is

$$\begin{aligned} Cov[Y(t), Y(s)] &= \frac{1}{2}\left[\mathbb{E}Y(t)^2 + \mathbb{E}Y(s)^2 - \mathbb{E}(Y(t) - Y(s))^2\right] \\ &= \frac{\sigma^2}{2}\left[|t|^{2H} - |t-s|^{2H} + |s|^{2H}\right]. \end{aligned}$$

For $\frac{1}{2} < H < 1$, the autocorrelation at lag k is then $\rho(k) \approx ck^{2-2H}$ which means that

$$\sum_{i=-\infty}^{\infty} \rho(i) = \infty. \tag{4.52}$$

Condition (4.52) is called a long memory effect in [Beran, 1994], since it means that events in the faraway past may affect the current events. Condition (4.52) is also used as the definition of LRD. It is difficult to distinguish a stationary LRD process from a nonstationary process in general and, especially for short time series, it may be practically impossible [Beran, 1994, Chapter 7.4]. In practice, the condition (4.52) cannot be studied from the sample ACFs since there is never enough stationary data for large lags and stronger methods to detect the possible scaling laws is required.

Perhaps the best tool for all kinds of scaling phenomena in traffic data is *wavelets* [Abry et al., 2000]. Longitudinal studies ([Fontugne et al., 2017], [Borgnat et al., 2009]) show that the existence of many scaling laws have been a persistent feature of Internet traffic and there is a different scaling law for scales smaller than RTT versus scales larger than (typical) RTT. Above the scale of RTTs, a single scaling law may be possible. This suggests that, before trying to estimate the self-similarity parameter H, one should use wavelets to study the scaling laws of the time scale of interest to decide whether a single scaling exponent H makes sense.

4.4 Summary

Table 4.2 summarizes the concepts that were discussed in this framework. The result (4.23) that gave $I_n\left(U_{(n-k)}; U_{(n)}\right)$ was computed for the purposes of this summary and, to our best knowledge, is not found in literature. The proof of Theorem 1 combined

(4.36) to an existing proof and this combination, to our best knowledge, is not found in the literature.

Table 4.2: Summary of the discussed topics in the statistical framework

	Publication			
Topic	[1]	[2]	[3]	[4]
Order statistics and quantiles	qq-plots nq-plots r_n^2	Hill estimator $\mathsf{H}_{k,n}$ $I_n\left(U_{(n-k)};U_{(n)}\right)$	(**Note 1**)	order statistics in initialization, quantiles in estimation, Theorem 1
Multivariate models	(**Note 2**)	bivariate with tail dependence: Pickands A	binormal: $\mathbb{E}(Z_2 \vert Z_1 > a)$ where Z_1 and Z_2 are asymmetries in two locations	trinormal: conditional independence of quantile estimators
Time dependence	stationary LRD	stationary not LRD	non-stationary daily profile	arbitrary: the algorithm attempts to detect if stationary

In addition to the discussed topics, the following two notes in Table 4.2 complete the framework:

Note 1 In publication [3, Section 4.3], we use robust estimates of the parameters of the normal distribution and they are computed from the sample quartiles.

Note 2 All finite dimensional marginal distributions of a Gaussian process are multinormally distributed ([Parzen, 1962],[Priestley, 1982]). The 2-dimensional case of the same data as in publication [1] was studied afterwards in an unpublished study. The further insight of that study was that the binormal fit of the 2-dimensional marginals was adequate when the 1-dimensional marginal fit was good. This already suggests that some covariance structure may have existed for the data of publication [1].

Chapter 5

Discussion

A result of statistical analysis or modelling is never an end to research. It is possible to improve the insight about the statistical nature of the studied problems by looking at the results of the publications afterwards, especially when some feedback is available.

5.1 Traffic data

The application research topic of publication [1] has been active ([De Meent et al., 2006], [d. O. Schmidt et al., 2013], [d. O. Schmidt et al., 2014]) and, because Internet traffic evolves, it is still active [Alasmar et al., 2021].

In hindsight, we should have been more careful when formulating the conclusion in publication [1]. Publication [1] is methodological in nature, but sometimes it has been cited as if the results of data analysis could be generalized. Normal-quantile plots are not meant to be used to prove a distributional assumption, instead they form a visually fast method to check if the normal distribution assumption does not hold. The use of the linear correlation coefficient, which we denoted as r_n^2 in publication [1], makes this method algorithmic. With the data in publication [1], we were able to show both a good fit to normal distribution *and* examples of a poor fit using the method. A value of $r_n^2(\Delta) \approx 1$ alone is not an argument, the argument comes with *both* vertical and horizontal aggregation.

With the data in publication [1] we showed the following: In the downstream direction the assumptions of the vertical aggregation could be analyzed and verified to hold. In the upstream direction the analysis of the traffic characteristics indicated that the assumptions of the vertical aggregation did not hold sufficiently well in the sense that the quantitative differences between control traffic flows and rare data uploading flows was too large. There were at least one magnitude differences within the sources in their contributions per time slot.

Another issue is the horizontal aggregation argument and the effect of RTT to the possible scaling laws as discussed in [Fontugne et al., 2017]. Even if the burst sizes were commonly bounded for all TCP connections of an aggregate, but if the RTTs have different magnitudes and some TCP connections have a very small RTT, they may contribute to a fixed time slot Δ with two or even more bursts. However, other TCP connections may only contribute to every second slot or even more rarely. The conclusion is that the data-based synthetic vertical aggregation study in publication [1, Section V.C] and the observed reality [Fontugne et al., 2017] indicate that vertical aggregation *alone* is insufficient for a Gaussian traffic model in small time scales.

The qq-plot-based methodology of publication [1] has been used in various other studies, We mention [De Meent et al., 2006], [Juva et al., 2007], [d. O. Schmidt et al., 2013] and [d. O. Schmidt et al., 2014] because they have also found examples of both roughly Gaussian and non-Gaussian traffic. Non-Gaussian examples are informative because their analysis suggest possible causes for non-Gaussianity. In [Alasmar et al., 2021] the methods also include qq-plots but they base their main conclusions on the log-likelihood ratio methodology of [Clauset et al., 2009]. Likelihood methods are generally based on the assumption of independent observations. We do not know how well the log-likelihood ratio methodology of [Clauset et al., 2009] works for LRD data or data with possible multi-fractal scaling laws [Fontugne et al., 2017]. However, according to [Alasmar et al., 2021], qq-plots support their conclusions of lognormal distribution as the best model between those models that they compared. This is another piece of evidence that the horizontal aggregation argument fails for time scales that are smaller than typical RTTs.

Regarding the qq-plot methodology, numerical problems with $r_n^2(\Delta)$ may be an issue when n is very large, which is the case if Δ is small. In addition to this, plotting too much data on the nq-plot (or np-plot) may not be a good thing as the amount of *new* information may not grow as expected. For example, if the sample size is $n = 2k$ and the mutual information of the consecutive order statistics next to the median $U_{(k)}$ is computed, then

$$I_{2k}\left(U_{(k)}, U_{(k+1)}\right) = 2kH_{2k-1} - 2kH_{k-1} - 1 - \log\binom{2k}{k} \to \infty, \qquad n \to \infty. \tag{5.1}$$

When an order statistic next to the median is considered then, asymptotically, it shares the same information as the median. Therefore, it does not bring much new information to the nq-plot or to the np-plot.

There are at least two alternatives to study the behavior of r_n^2 with increasing sample sizes n. The first alternative is to fix m and then select probabilities $0 < q_1 < \ldots < q_m < 1$. The nq-plot could be done by plotting only m points

$$\left(\Phi^{-1}(q_j), x_{(\lceil q_j n \rceil)}\right), \qquad j = 1, \ldots, m, \tag{5.2}$$

and the linear correlation coefficient computed from these m points. For any value $n \geq m$, the nq-plot then has only m points. To distinguish this from r_n^2 of [1], let $r_{n,m}^2$ be the notation for the linear correlation coefficient computed in this way. This requires only slightly different computation due to selection of $\lceil q_j n \rceil$:th members of the ordered data. With $r_{n,m}^2$ the effect of increasing the sample sizes n can still be compared as in [1]: for non-normal data $r_{n,m}^2$ stops improving at some point when n increases, at least when m is large enough. The problem with (5.2) is that the fit to the central part is only taken into account, not the tails.

Another fast alternative, which keeps the tails included, is to include only every h:th pair in the nq-plot. For example, including only every second pair in (4.26), that is, $i = 1, 3, 5, \ldots, n$ in steps of $h = 2$, reduces the number of points down to $\approx n/h = 50\%$ of n in the plot. No data is dropped, only the number of points in the qq-plot is reduced. The corresponding correlation coefficient $r^2_{n,n/h}$ still has the same behavior when the sample size n is increased, at least when h is small.

The fact (5.1) suggests that the probabilities q_j in (5.2) or h in the above discussed thinned alternative should be chosen so that

$$I_n\left(X_{(\lceil q_j n\rceil)}; X_{(\lceil q_{j+1} n\rceil)}\right) \qquad \text{or} \qquad I_n\left(X_{(i)}; X_{(i+h)}\right) \tag{5.3}$$

are small so that the amount of new information is large in the nq-plot. This criterion together with the criteria that m is large enough or that h is small suggest that a compromise may exist. This is a topic of further research.

5.2 Online, sequential or real-time analysis

The more there is data, the more there is a need to reduce the dimensionality of the data retaining nearly the same amount of information. We have worked several years with data analysis of different types of data from engineering applications/field. The amount of data in typical traces has grown from Megabytes to Gigabytes and even more and this is a thousandfold increase, as Table 3.1 indicated.

The increased amount of data also demands careful preprocessing and algorithmic methods for analyses. However, the quality of the data is far more important than the amount of data. The most important is to understand the process that produces the data. The data may be collected under specific design, sampling scheme or selection processes, and may give rise to missing data. Such aspects have to be accounted for in the analysis.

The teletraffic speeds and volumes are already huge and they are still growing. The consequence of the approximate self-similarity, that the bursty nature of the data traffic does not get smoother when aggregated [Leland et al., 1994], is a persistent feature [Fontugne et al., 2017]. Some of the network operator staff personnel have told us that, as a rule of thumb, whenever the traffic average load exceeds 50% of the available capacity they start to invest in new hardware that will increase the capacity. Increasing the network capacity tends to be cheaper than fine-tuning the existing network resources.

If a plausible traffic model has been found, the computation of its parameters is usually fast while checking if the model describes the data is usually a slow process. An easier and faster way is to detect deviations from a given model and then ranking them. However, the main problem in the approach of taking snapshots of data traffic and spending the effort to analyze them is that teletraffic data out-dates rapidly. Changes in the network configurations tend to be frequent and they may affect the aggregate traffic profiles instantly. This calls for online, sequential or real-time algorithms that are capable of providing relevant non-trivial information about the existing traffic in real time. We believe that there are a lot of fruitful research possibilities in that area.

Bibliography

[Abry et al., 2000] Abry, P., Flandrin, P., Taqqu, M., and Veitch, D. (2000). *Wavelets for the analysis, estimation, and synthesis of scaling data*, chapter 2, pages 39–88. John Wiley & Sons, Ltd.

[Alasmar et al., 2021] Alasmar, M., Clegg, R., Zakhleniuk, N., and Parisis, G. (2021). Internet Traffic Volumes are Not Gaussian—They are Log-Normal: An 18-Year Longitudinal Study With Implications for Modelling and Prediction. *IEEE/ACM Transactions on Networking*, 29(3).

[Beirlant et al., 2004] Beirlant, J., Goegebeur, Y., Teugels, J., and Segers, J. (2004). *Statistics of Extremes.* J. Wiley & Sons.

[Beran, 1994] Beran, J. (1994). *Statistics for Long-Memory Processes.* Chapman & Hall, Routledge., 1st edition.

[Biggs, 1989] Biggs, N. L. (1989). *Discrete Mathematics.* Clarendon Press, Oxford, Walton Street, Oxford, revised edition edition.

[Borgnat et al., 2009] Borgnat, P., Dewaele, G., Fukuda, K., Abry, P., and Cho, K. (2009). Seven years and one day: Sketching the evolution of internet traffic. In *IEEE INFOCOM 2009*, pages 711–719.

[Brown and Hettmansperger, 1996] Brown, B. M. and Hettmansperger, T. P. (1996). Normal scores, normal plots, and tests for normality. *Journal of the American Statistical Association*, 91(436):1668–1675.

[Caers and Van Dyck, 1999] Caers, J. and Van Dyck, J. (1999). Nonparametric tail estimation using a double bootstrap method. *Computational Statistics & Data Analysis*, (29):191–211.

[Capéraà et al., 1997] Capéraà, P., Fougères, A.-L., and Genest, C. (1997). Estimation of bivariate extreme value copulas. *Biometrika*, 84:567–577.

[Casella and Berger, 2002] Casella, G. and Berger, R. L. (2002). *Statistical Inference.* Duxbury, 2nd edition.

[Chistyakov, 1964] Chistyakov, V. (1964). A theorem on sums of independent positive random variables and its application to branching random processes. *Theory Probab. Appl.*, (9):640–648.

[Clauset et al., 2009] Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

[Coles, 2001] Coles, S. (2001). *An introduction to statistical modeling of extreme values.* Springer Series in Statistics. Springer-Verlag, London.

[Cormen et al., 2009] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to Algorithms.* The MIT Press, Cambridge, Massachusetts, third edition.

[Cover and Thomas, 2006] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing).* Wiley-Interscience, New York, NY, USA.

[d. O. Schmidt et al., 2014] d. O. Schmidt, R., Sadre, R., Melnikov, N., Schönwälder, J., and Pras, A. (2014). Linking network usage patterns to traffic gaussianity fit. In *2014 IFIP Networking Conference*, pages 1–9.

[d. O. Schmidt et al., 2013] d. O. Schmidt, R., Sadre, R., and Pras, A. (2013). Gaussian traffic revisited. In *2013 IFIP Networking Conference*, pages 1–9.

[Dawid, 1979] Dawid, A. (1979). Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society*, 41:1–31.

[De Meent et al., 2006] De Meent, R. V., Mandjes, M., and Pras, A. (2006). Gaussian traffic everywhere? In *2006 IEEE International Conference on Communications*, volume 2, pages 573–578.

[Devroye et al., 1996] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Applications of Mathematics. Springer.

[DLMF,] DLMF. *NIST Digital Library of Mathematical Functions.* http://dlmf.nist.gov/, Release 1.1.1 of 2021-03-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

[Ebrahimi et al., 2004] Ebrahimi, N., Soofi, E. S., and Zahedi, H. (2004). Information properties of order statistics and spacings. *IEEE Transactions on Information Theory*, 50(1):177–183.

[Filliben, 1975] Filliben, J. J. (1975). The probability plot correlation coefficient test for normality. *Technometrics*, 17(1).

[Fontugne et al., 2017] Fontugne, R., Abry, P., Fukuda, K., Veitch, D., Cho, K., Borgnat, P., and Wendt, H. (2017). Scaling in internet traffic: A 14 year and 3 day longitudinal study, with multiscale analyses and random projections. *IEEE/ACM Transactions on Networking*, 25(4):2152–2165.

[Hall, 1990] Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of Multivariate Analysis*, 32(2):177–203.

[Hall and Tajvidi, 2000] Hall, P. and Tajvidi, N. (2000). Distribution and dependence-function estimation for bivariate extreme-value distributions. *Bernoulli*, 6:835–844.

[Heckmann, 2006] Heckmann, O. (2006). *The competitive Internet Service Provider.* John Wiley & Sons, Ltd.

[Hill, 1975] Hill, B. M. (1975). A Simple General Approach to Inference About the Tail of a Distribution. *The Annals of Statistics*, 3(5):1163 – 1174.

[Hoeffding, 1963] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301).

[Højsgaard et al., 2012] Højsgaard, S., Edwards, D., and Lauritzen, S. (2012). *Graphical Models with R.* Springer-Verlag New York.

[Johnson and Wichern, 2007] Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis.* Pearson Prentice Hall, New Jersey, 6th edition.

[Juva et al., 2007] Juva, I., Susitaival, R., Peuhkuri, M., and Aalto, S. (2007). Effects of spatial aggregation on the characteristics of origin-destination pair traffic in funet. In Koucheryavy, Y., Harju, J., and Sayenko, A., editors, *Next Generation Teletraffic and Wired/Wireless Advanced Networking*, pages 1–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Kilpi and Lassila, 2006] Kilpi, J. and Lassila, P. (2006). Micro- and Macroscopic Analysis of RTT Variability in GPRS and UMTS Networks. In *Networking 2006*, number 3976 in LNCS, pages 1176–1181. IFIP, Springer.

[Kluppelberg, 1988] Kluppelberg, C. (1988). Subexponential distributions and integrated tails. *Journal of Applied Probablity*, 25(1):132–141.

[Kluppelberg, 1989] Kluppelberg, C. (1989). Subexponential distributions and characterizations of related classes. *Probab. Th. Rel. Fields*, (82):259–269.

[Kotz et al., 2000] Kotz, S., Balakrishnan, N., and Johnson, N. L. (2000). *Continuous multivariate distributions*, volume 1 of *Wiley Series in Probability and Statistics.* 2nd edition.

[Kullback, 1968] Kullback, S. (1968). *Information theory and statistics.* Dover.

[Leland et al., 1994] Leland, W., Taqqu, M., Willinger, W., and Wilson, D. (1994). On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15.

[Leslie, 1989] Leslie, J. R. (1989). On the non-closure under convolution of the subexponential family. *Journal of Applied Probability*, 26(1):58–66.

[Lin et al., 2012] Lin, Y.-D., Hwang, R.-H., and Baker, F. (2012). *Computer Networks: An Open Source Approach.* McGraw-Hill.

[Mannersalo, 2002] Mannersalo, P. (2002). Some notes on prediction of teletraffic. In *Proceedings of 15th ITC Specialist Seminar*, pages 220–229, Würzburg, Germany.

[Medhi and Ramasamy, 2007] Medhi, D. and Ramasamy, K. (2007). *Network Routing, Algorithms, Protocols and Architectures.* Morgan Kaufman.

[Nevzorov, 2001] Nevzorov, V. B. (2001). *Records: Mathematical Theory*, volume 194 of *Translations of Mathematical Monographies.* American Mathematical Society.

[Németh and Zempléni, 2020] Németh, L. and Zempléni, A. (2020). Regression estimator for the tail index. *J Stat Theory Pract*, 14(48).

[Park and Willinger, 2000] Park, K. and Willinger, W. (2000). *Self-Similar Network Traffic: An Overview*, chapter 1, pages 1–38. John Wiley & Sons, Ltd.

[Parzen, 1962] Parzen, E. (1962). *Stochastic Processes.* Dover, San Francisco.

[Pickands, 1975] Pickands, J. I. (1975). Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics*, 3(1):119 – 131.

[Pitman, 1980] Pitman, E. (1980). Subexponential distribution functions. *Journal of the Australian Mathematical Society*, 29:337–347.

[Priestley, 1982] Priestley, M. B. (1982). *Spectral analysis and Time Series.* Academic Press, New York.

[Raatikainen, 1987] Raatikainen, K. (1987). Simultaneous estimation of several percentiles. *Simulation*, pages 159–164.

[Resnick, 1997] Resnick, S. (1997). Heavy tail modeling and teletraffic data. with discussion and a rejoinder by the author. *Annals of Statistics*, 25:1805–1869.

[Roberts et al., 1996] Roberts, J., Mocci, U., and Virtamo, J., editors (1996). *Broadband Network Teletraffic.* Number 1155 in Lecture Notes in Computer Science. Springer-Verlag Berlin Heidelberg.

[Samorodnitsky, 2002] Samorodnitsky, G. (2002). Long range dependence, heavy tails and rare events. Cornell University Operations Research and Industrial Engineering.

[Serfling, 1980] Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley series in probability and statistics. John Wiley & Sons.

[Smirnov, 1949] Smirnov, N. V. (1949). Limit distributions for the terms of a variational series. *Trudy Mat. Inst. Steklov*, 25:3–60.

[Stevens, 1994] Stevens, W. (1994). *TCP/IP Illustrated, Volume 1: The protocols*. Addison-Wesley.

[Valean, 2019] Valean, C. (2019). *(Almost) impossible integrals, sums, and series*. Problem Books in Mathematics. Springer.

Appendix A

Computation of mutual information

In this Appedix we do the computation of the non-trivial but very intuitive result of mutual information $I_n\left(U_{(i)};U_{(n)}\right)$ that we used in the Statistical framework of Chapter 4. We are not aware of similar published results but, since [Ebrahimi et al., 2004] exist, it is quite possible that someone has already done the same computation.

It is convenient to rewrite the densities (4.2) and (4.3) of the order statistics in the uniform distribution case and use the notations $g_{(i)}$ and $g_{(i)(j)}$ for them:

$$\begin{aligned} g_{(i)}(u) &= \frac{n!}{(i-1)!(n-i)!} u^{i-1}(1-u)^{n-i} \\ &= \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i+1)} u^{i-1}(1-u)^{n-i} \\ &= \frac{1}{B(i,n-i+1)} u^{i-1}(1-u)^{n-i}, \qquad 0 \le u \le 1, \end{aligned} \tag{A.1}$$

which is the density of the $Beta(i, n-i+1)$ distribution, and

$$\begin{aligned} g_{(i)(j)}(u,v) &= \frac{n!}{(i-1)!(j-i-1)!(n-j)!} u^{i-1}(v-u)^{j-i-1}(1-v)^{n-j} \\ &= \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(j-i)\Gamma(n-j+1)} u^{i-1}(v-u)^{j-i-1}(1-v)^{n-j}. \end{aligned} \tag{A.2}$$

The joint density (A.2) is defined for $1 \le i < j \le n$ and $0 \le u < v \le 1$. Partial integration allows computing directly ([Valean, 2019]),

$$\int_0^1 u^{n-1} \log(1-u) du = -\frac{1}{n} H_n, \tag{A.3}$$

but it is also a special case of formula (4.12) when $\alpha = n$ and $\beta = 1$:

$$B(n,1)[\psi(1) - \psi(n+1)] = \frac{\Gamma(n)\Gamma(1)}{\Gamma(n+1)}[-\gamma_E - H_n + \gamma_E] = -\frac{1}{n} H_n.$$

The following results make the final computation a little bit shorter: $\Gamma(1) = 1$, $g_{(n)}(v) = nv^{n-1}$, and

$$g_{(i)(n)}(u,v) = \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i)} u^{i-1}(v-u)^{n-i-1}. \tag{A.4}$$

The following integration is computed by changing the order of integration:

$$\begin{aligned}\int_0^1\int_0^v u^{i-1}(v-u)^{n-i-1}\log(1-u)du\,dv &= \int_0^1\int_u^1 u^{i-1}(v-u)^{n-i-1}\log(1-u)dv\,du \\ &= \int_0^1 u^{i-1}\log(1-u)\Big[\int_u^1 (v-u)^{n-i-1}dv\Big]du \\ &= \frac{1}{n-i}\int_0^1 u^{i-1}(1-u)^{n-i}\log(1-u)du \\ &= \frac{\Gamma(i)\Gamma(n-i)}{\Gamma(n+1)}\left[\psi(n-i+1)-\psi(n+1)\right] \qquad (\text{A.5})\end{aligned}$$

A.1 Computation of $I_n\left(U_{(i)}; U_{(n)}\right)$

With (A.5), compute as follows:

$$I_n(U_{(i)}, U_{(n)}) = \int_0^1\int_0^v g_{(i)(n)}(u,v)\log\left(\frac{g_{(i)(n)}(u,v)}{g_{(i)}(u)g_{(n)}(v)}\right)du\,dv$$

$$= \int_0^1\int_0^v \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i)}u^{i-1}(v-u)^{n-i-1}\log\left(\frac{\Gamma(n-i+1)}{n\Gamma(n-i)}\frac{(v-u)^{n-i-1}}{(1-u)^{n-i}v^{n-1}}\right)du\,dv$$

$$= \int_0^1\int_0^v \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i)}u^{i-1}(v-u)^{n-i-1}\left[\log\left(\frac{n-i}{n}\right) + \log\frac{(v-u)^{n-i-1}}{(1-u)^{n-i}v^{n-1}}\right]du\,dv$$

$$= \log\left(\frac{n-i}{n}\right)\underbrace{\frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i)}\int_0^1\int_0^v u^{i-1}(v-u)^{n-i-1}du\,dv}_{=1,\text{ since (A.4) is a density.}}$$

$$+ \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i)}\int_0^1\int_0^v u^{i-1}(v-u)^{n-i-1}\log\frac{(v-u)^{n-i-1}}{(1-u)^{n-i}v^{n-1}}du\,dv$$

$$
\begin{aligned}
&= \log\left(\frac{n-i}{n}\right) + \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i)} \underbrace{\int_0^1 \underbrace{\int_0^v u^{i-1}(v-u)^{n-i-1}\log(v-u)^{n-i-1} du}_{=(n-i-1)B(i,n-i)v^{n-1}[\log v-\psi(n)+\psi(n-i)]}\, dv}_{=(n-i-1)B(i,n-i)[n\psi(n-i)-n\psi(n)-1]/n} \\
&\quad - \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i)} \int_0^1\int_0^v u^{i-1}(v-u)^{n-i-1}\log\left[(1-u)^{n-i}v^{n-1}\right] du\, dv \\
&= \log\left(\frac{n-i}{n}\right) + \left(\frac{n-i-1}{n}\right)[n\psi(n-i) - n\psi(n) - 1] \\
&\quad - \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i)} \int_0^1\int_0^v u^{i-1}(v-u)^{n-i-1}\left[\log(1-u)^{n-i} + \log v^{n-1}\right] du\, dv \\
&= \log\left(\frac{n-i}{n}\right) + \left(\frac{n-i-1}{n}\right)[n\psi(n-i) - n\psi(n) - 1] \\
&\quad - (n-i)\frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i)} \underbrace{\int_0^1\int_0^v u^{i-1}(v-u)^{n-i-1}\log(1-u) du\, dv}_{\text{(A.5) applies here!}} \\
&\quad - (n-1)\frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i)} \underbrace{\int_0^1 (\log v) \underbrace{\int_0^v u^{i-1}(v-u)^{n-i-1} du}_{=B(i,n-i)v^{n-1}}\, dv}_{=-B(i,n-i)/n^2} \\
&= \log\left(\frac{n-i}{n}\right) + (n-i-1)[\psi(n-i) - \psi(n)] - \frac{n-i-1}{n} \\
&\quad - (n-i)[(\psi(n-i+1) - \psi(n+1)] + \frac{n-1}{n} \\
&= \log\left(\frac{n-i}{n}\right) - \psi(n-i) + \psi(n) \qquad \text{(A.6)} \\
&= \log\left(\frac{n-i}{n}\right) - H_{n-i-1} + H_{n-1}. \qquad \text{(A.7)}
\end{aligned}
$$

The result is in terms of the ψ-function (A.6) and in terms of Harmonic numbers (A.7). The case $i = n-1$ provides a partial check since that case is included in the consecutive pair case. There is no symmetry between i and $n-i$. Due to (4.19) it appears the best to write the result in the form

$$
I_n\left(U_{(i)}, U_{(n)}\right) = H_{n-1} - \log n - \left[H_{n-i-1} - \log(n-i)\right]. \qquad \text{(A.8)}
$$

The case $i = 1$ gives

$$I_n\left(U_{(1)}; U_{(n)}\right) = \log\left(1 - \frac{1}{n}\right) + \frac{1}{n-1}, \tag{A.9}$$

which should be of general interest since a typical application of order statistics is to estimate the distribution of the range $R = X_{(n)} - X_{(1)}$ ([Casella and Berger, 2002]).