Statistical models for agroclimate risk analysis

by

Mohamadreza Hosseini

B.Sc., Amirkabir University, 2003
M.Sc., McGill University, 2005

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

November, 2009

Abstract

In order to model the binary process of precipitation and the dichotomized temperature process, we use the conditional probability of the present given the past. We find necessary and sufficient conditions for a collection of functions to correspond to the conditional probabilities of a discrete–time categorical stochastic process $X_1, X_2, \cdots$. Moreover we find parametric representations for such processes and in particular rth–order Markov chains.

To dichotomize the temperature process, quantiles are often used in the literature. We propose using a two–state definition of the quantiles by considering the "left quantile" and "right quantile" functions instead of the traditional definition. This has various advantages such as a symmetry relation between the quantiles of random variables X and $-X$. We show that the left (right) sample quantile tends to the left (right) distribution quantile at $p \in [0, 1]$, if and only if the left and right distribution quantiles are identical at p and diverge almost surely otherwise. In order to measure the loss of estimating (or approximating) a quantile, we introduce a loss function that is invariant under strictly monotonic transformations and call it the "probability loss function." Using this loss function, we introduce measures of distance among random variables that are invariant under continuous strictly monotonic transformations. We use this distance measures to show optimal overall fits to a random variable are not necessarily optimal in the tails. This loss function is also used to find equivariant estimators of the parameters of distribution functions.

We develop an algorithm to approximate quantiles of large datasets which works by partitioning the data or use existing partitions (possibly of non-equal size). We show the deterministic precision of this algorithm and how it can be adjusted to get customized precisions. Then we develop a framework to optimally summarize very large datasets using quantiles and combining such summaries in order to infer about the original dataset.

Finally we show how these higher order Markov models can be used to construct confidence intervals for the probability of frost–free periods.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisors, Prof. Jim Zidek and Prof. Nhu Le for what they have taught me in statistics and a lot more, their encouragements, ideas and financial support through various RA positions during my PhD studies. I feel very grateful and lucky to have them as my supervisors. I should also thank Prof. Matias Salibian-Barrera on my supervisory committee for giving me great ideas and feedbacks. I also like to thank other people in the statistics department at UBC, prof. Paul Gustafson, prof. John Petkau, prof. Constance Van Eden and prof. Ruben Zamar which I owe them a lot of things I know. I also thank Mike Marin (instructor at UBC) for having various interesting discussions about statistics and science, and Viena Tran for helping me regarding many administrative issues. I like to thank my friend Dr. Nathaniel Newlands for insightful comments and good suggestions and Ralph Wright (Alberta Agriculture Food and Rural Development) for making useful comments about the definition of the extremes.

Finally, I like to express my deepest appreciation to all the people who helped me learn and love statistics and mathematics during all my life, from my grandfather who graciously taught me mathematics when I was a child to my mother who has been an inspiration and high school teachers who encouraged me to study mathematics as a university major.

Dedication

To my lovely parents,
my amazing brother: Alireza,
my sweet sister: Fatima,
my best friends: Mostafa Aghajanpour, Mahmoud Sohrabi, Masoud Feizbakhsh, Behruz Khajali, Ali Mehrabian, Prof. Masoud Asgharian, Prof. Niky Kamran, Mirella Simoneova, Kiyouko Futaeda, May Yun, Yuki Ezaki, Soheil Keshmiri, Naoko Yoshimi and Mike Marin.

Chapter 1

Thesis introduction

This thesis develops mathematical and statistical framework to model stochastic processes over time. In particular it develops models for precipitation and extreme (high or low) temperature events occurrences. This is important for Canada's agriculture since agricultural production is dependent on weather and water availability.

We study the quantiles of data and distributions in detail and develop a framework for approximating quantiles in large datasets and inference. We also study categorical Markov chains of higher order and apply them to precipitation and temperature processes. However, the methodologies and theories developed here are general and can be used in many other applications where such processes are encountered (such as physics, chemistry, climatology, economics and so on).

Sample quantiles and quantile function are fundamental concepts in statistics. In the study of extreme events they are often used to pick appropriate thresholds. We use the quantiles specifically to pick thresholds for the temperature process. This motivates us to study the concept of quantiles and extend their classic definition to provide a more intuitively appealing alternative. This alternative also enables us to get interesting asymptotic results about their sample counterparts and a framework to approximate quantiles and make inference. In fact weather datasets (observed weather or output of climate models) are very large in size. This makes computing the quantiles of such large datasets computationally intensive. Along with this alternative definition, we present an algorithm for computing/approximating quantiles in large datasets.

The data used in this thesis come from the climate data CD published by Environment Canada [10], which includes the daily observed precipitation and temperature data for several station from 1895 to 2007 (the years varying with the station). The data are saved in several binary files. We have written a Python module to extract the data in desired formats. The guide to using this module is in Appendix B. For most of the analysis however, we have used the "homogenized" dataset for Alberta. This dataset is adjusted for change of instruments and location of the stations. More information

about the datasets is given in Appendix B and Chapter 2.

Chapter 2 presents results from the exploratory analysis of the dataset. We look at the variables' daily time series, monthly means time series, annual means time series and the distribution of the daily/annual means values. We also look at the relation between the variables as well as some long–term trends by simple techniques such as linear regression. For example it seems that the mean summer daily minimum temperature has increased over time at some locations in Alberta. Then we study the seasonal patterns of these variables over the course of the year. As expected there is a strong seasonal component in these processes. For example, we observe that the daily temperature is more variable in the colder seasons than the warmer ones. The daily values for the minimum and maximum temperature seem to be described fairly well by a Gaussian process. However, some deviations from the Gaussian assumption is seen in the tails. This is particularly important in modeling extreme events and will help us in later chapters to choose our approach to modeling the occurrence of such extremes. As a part of the exploratory analysis, we look the precipitation occurrence. A question that has been addressed by several authors (e.g. Tong in [45] and Gabriel et al. in [18]) is the Markov order of such a chain. The exploratory analysis using the transition probabilities plots leads to the conjecture that a 1st–order Markov chain should be appropriate. This is studied in detail in later chapters. We also look at the spatial–temporal correlation function of these processes. Several interesting features are observed. For example for the maximum and minimum temperature, the correlation seems to be stationary over time. Also the geodesic distance seems to describe the spatial correlation for temperature well. For precipitation on other hand not much spatial correlation is observed. This could be due to the fact that we have only 47 precipitation stations available over Alberta and it is more variable over space compared to temperature.

Let us denote a general weather process by X_t, where t denotes time. The main approach we take to model the process is discrete–time categorical rth–order Markov chains (r a natural number), where we have the following assumption for the conditional probabilities:

$$P(X_t|X_{t-1},\cdots) = P(X_t|X_{t-1},\cdots,X_{t-r}).$$

"Categorical chain" here means that X_t takes only a finite number of possible states. For example it can be a two state space of the (occurrence)/(non-occurrence) of precipitation. Dichotomizing the temperature process, we can consider processes such as (freezing)/(not freezing). Processes with

more than two states can also be considered. For example a process with three states: (not warm)/(warm)/(hot).

Chapter 3 studies the rth–order categorical Markov chains in general. We present a new representation theorem for such chains that expresses the above conditional probability as a linear combination of the monomials of past process values $X_{t-1}, \cdots, X_{t-r}$. We will show the existence and uniqueness of such a representation. In the stationary case since the conditional probability is the same for all time points, some more work on the consistency shows that this representation characterizes all stationary categorical rth–order Markov chains. For the binary case the result is a corollary of a theorem stated in [6]. However, the expression of the theorem in [6] is flawed as also pointed out by Cressie et al. [14]. We present a rigorous statement along with a constructive proof for the theorem. For discrete–time categorical chains with more than two states this theorem does not seem especially useful. We prove a new theorem for this case that gives us representation for all discrete–time categorical chains (rather than only binary). In order to estimate the parameters of such a model in the binary case and infer about them, we use the "Time series following general linear models" as described in [27]. The inferences are similar to generalized linear models. However because of dependencies over time some extensions of the usual theory are needed. Maximizing the "partial likelihood" will give us "consistent" estimators as shown in [48]. We apply the partial likelihood theory to our proposed rth–order Markov models. Simulations show that partial likelihood and the representation together give us satisfactory results for the binary case. We also check the performance of the Bayesian information criterion (BIC), developed in [42] and others, to pick optimal models by simulation studies and we get satisfactory results. This allows us not only to pick the order of the Markov chain but also to compare several Markov chains of the same order. Another advantage of this model to existing ones is the capacity to accommodate other continuous variables. For example, we can add some seasonal processes to get a non–stationary chain. [In previous studies regarding the order of the chain e.g. [45] and [18], it was assumed that the precipitation chain is stationary.] We can also add covariate processes such as temperature of the previous day to the model. Then we apply these techniques to the binary precipitation process in Alberta and pick appropriate models. A 1st–order non-stationary (with one seasonal term) seem to be the most appropriate based on the BIC method for model selection.

To apply these techniques to the temperature processes, we need a way of dichotomizing the temperature process. Usually certain quantiles are cho-

sen in order to do so. Computing the quantiles for large datasets can be computationally challenging. Very large datasets are often encountered in climatology, either from a multiplicity of observations over time and space or outputs from deterministic models (sometimes in petabytes= 1 million gigabytes). Loading a large data vector and sorting it, is impossible sometimes due to memory limitations or computing power. We show that a proposed algorithm to approximating the median, "the median of the median" performs poorly. Instead, we propose a new algorithm that can give us good approximations to the exact quantiles, which is an extension of the algorithm proposed in [3]. In fact, we derive the precision of the algorithm. The algorithm partitions the data, "coarsens" the partitions at every iteration, put the coarsened vectors together and sort it instead of the original vector.

Working on the quantiles, in order to find some theory to justify the usefulness and accuracy of the algorithm motivated us to think about the definition of the quantile function and quantiles for data vectors. The quantile function of a random variable X with distribution function F is traditionally defined as

$$q(p) = \inf\{u|F(u) \geq p\}.$$

Applying this to the fair coin example with $0, 1$ as outputs, we get $q(1/2) = 0$. This is counterintuitive to the fact that the distribution has equal mass on 0 and 1. Also a standard definition for the quantiles does not exist for a data vector. [For example Hyndman et al. [25] point out that there are many definitions of quantiles in different packages.] For example suppose a data vector has an even number of points, then there is no point exactly in the middle, in which case, the average of the two middle values is often proposed as the median. We argue that this is not a good definition. In fact, we present an alternative way of defining quantiles that is motivated by an intuitive experiment and resolve all the above problems. We propose using the two-state definition of right and left quantiles instead of only quantile. The left quantile is defined as above and the right quantile is defined to be

$$rq(p) = \sup\{u|F(u) \leq p\}.$$

We also define left and right quantiles for the data vectors and study the limit properties of the sample quantiles. For example it turns out the sample left and right quantile converge to the distribution quantiles if and only if the left and right quantile are equal. This again shows another interesting aspect of the definition of quantiles and confirms that it is not redundant. This definition is an extension to the concept of upper and lower median

in robustness literature. Also in some books (e.g. [41]) $rq(p)$ is taken to be the definition of quantiles. However, we do not know of any study of their properties or a claim that considering both can lead to many interesting results. We also show that the widely claimed equivariance property of traditional (left) quantile functions under strictly increasing transformations (for example in [21] and [29]) is false. However, we show that the left (right) quantile is equivariant under left (right) continuous increasing transformations. We also provide a neat result for continuous decreasing transformations. We also show that the probability that the random variable is between these the right and left quantile is zero and the left and right quantile are identical except for at most a countable subset of [0,1].

Since our objective is to approximate to the exact quantiles by our algorithm, we need a way of assessing the accuracy such an approximation (a loss function). We introduce a new loss function that is invariant under strictly monotonic transformations of the data or the random variable. This loss function is very natural and in summary the loss of estimating a quantile z by z' is the probability that the random variable is between these two values. In other words, we use the mass of the random variable itself between the two values to judge the goodness of the approximation. We also show some limit theorems to show the empirical loss function tends to the loss function of the distribution. This loss function might be a useful tool in many other contexts and is an interesting topic for future research.

We show by simulations and real data the algorithm performs well. Then we will apply it to the weather data to pick the 95% quantile for the maximum daily temperature. After picking the quantiles we use the rth–order Markov techniques and partial likelihood to find appropriate models to describe the temperature. Using this loss function and the theory developed for the quantiles, we introduce measures to compute "distance" among distribution functions over the reals (random variables) that are invariant under continuous strictly monotonic transformations. We use this distance measures to show optimal overall fits to a random variable are not necessarily optimal in the tails (and hence not appropriate to study extremes). We also find "optimal" ways of picking a limited number of probabilities $0 \leq p_1 < \cdots < p_k \leq 1$ to summarize a random variable by its corresponding quantiles.

Finally we show how these higher order Markov models can be used to construct confidence intervals for the probability of a frost free week at the beginning of August at Medicine Hat (in Alberta).

The last chapter provides a summary of the work and the conclusions. It also points out some interesting questions that are not answered in this

thesis and a research proposal for the future.

Chapter 2

Exploratory analysis of the Canadian weather data

2.1 Introduction

This chapter performs an exploratory analysis for the Homogenized climate dataset for the province of Alberta in Canada. We have access to daily maximum temperature (MT), daily minimum temperature (mt) and precipitation (PN). The temperature data have been provided to us by Vincent, L.A. and the precipitation data have been provided to us by Eva Mekis both from Environment Canada. This dataset has been homogenized for changes of instrument, changes of the location of the stations and so on. More information about these data can be found in [34] and [47]. These data are a homogenized part of a larger dataset published by Environment Canada (2007), which are in binary format and a Python module in order to extract them is provided in Appendix B.

This chapter uses several graphical and analytical tools to examine the behavior of selected climate variables. Looking at the data, we will see some interesting features that suggest future research.

Section 2 describes the dataset. For example the location plots of the stations and their elevation plots are given. In Section 3, we look at the daily and annual time series of temperatures and precipitation. The normality of the distribution of annual values and the associations between different variables are investigated. We have also investigated the seasonal patterns as well as the long–term patterns for different variables over the course of the year. For example, the mean summer daily minimum temperature shows a significant increasing pattern over the course of the past century in Calgary and some other locations. Section 4 looks at the distribution of the daily values. For example, a normal distribution seems to describe the temperature and a Gamma distribution, the precipitation daily values. Confidence intervals for the mean/standard deviation in the normal case and shape/scale parameters in the Gamma case are given. Section 5 looks

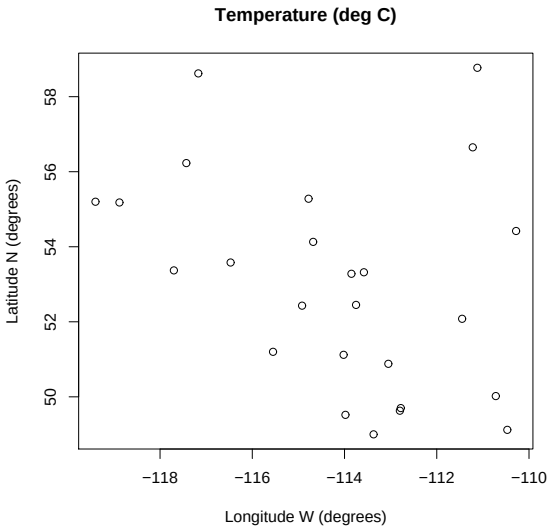

Figure 2.1: Alberta site locations for temperature (deg C) data. There are 25 stations available with temperature data over Alberta.

at the spatial and temporal correlation of different variables.

2.2 Data description

The temperature data comes from 25 stations over Alberta which operated from 1895 to 2006. PN data involve 47 stations from 1895 to 2006. Different stations have different intervals of data available. For example, the PN data for Caldwell is available from 1911 to 1990. Figures 2.1 and 2.2 respectively depict the location of the stations for temperature (both MT and mt) and PN. The number of years available for each station is plotted against the location in Figures 2.3 and 2.4. Another available variable for the location of the stations is the elevation. Figures 2.5 and 2.6 show the elevation in meters. As seen in the plots, some stations have both temperature and precipitation data.

2.3 Temperature and precipitation

To get some initial impression of the data, we look at the time series of MT, mt, and PN at a fixed location. We use the Calgary site since it has a long

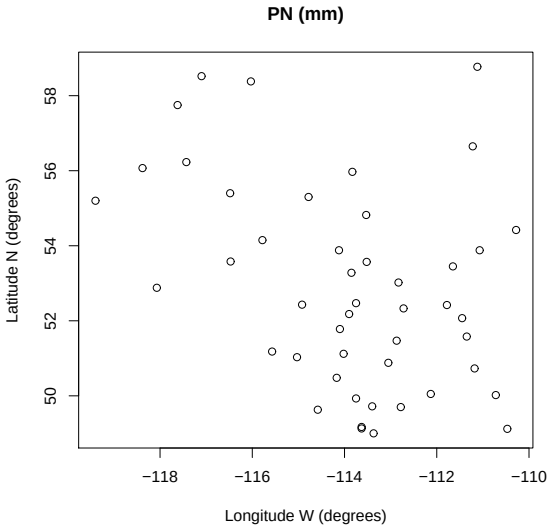

Figure 2.2: Alberta site locations for precipitation (mm) data. There are 47 stations available with precipitations data over Alberta.

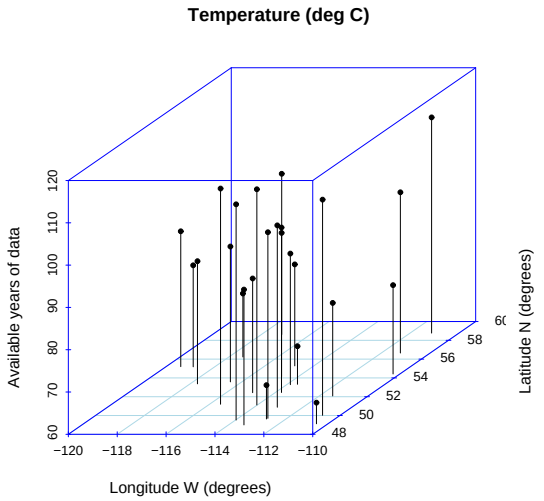

Figure 2.3: The number of years available for sites with temperature (deg C) data.

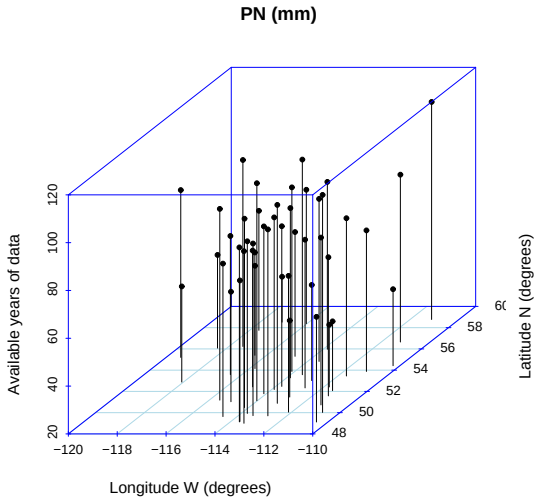

Figure 2.4: The number of years available for sites with precipitation (mm) data available.

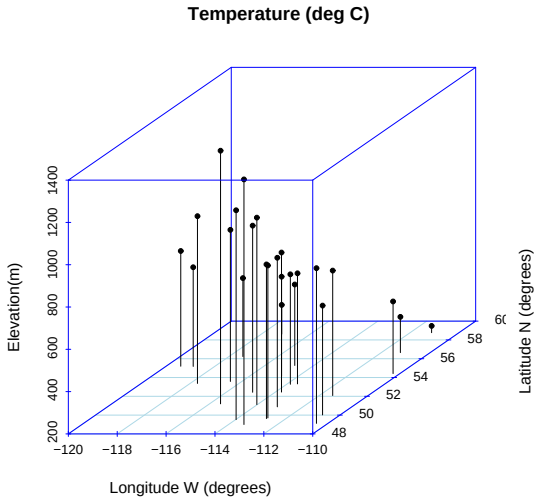

Figure 2.5: The elevation (meters) of sites with temperature data available.

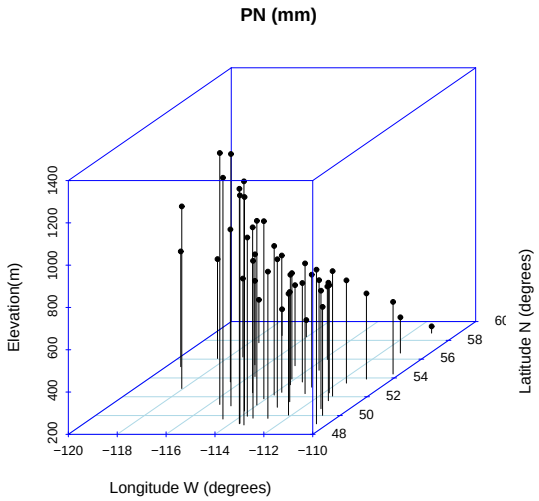

Figure 2.6: The elevation (meters) of the sites with precipitation data available.

period of data available and includes both temperature and precipitation.

Looking at the maximum and minimum temperature, we see the periodic trend over the course of a year as shown in Figures 2.7 and 2.8 which illustrate the MT and mt daily values from 2000 to 2003. A regular seasonal trend is seen in both processes.

Looking at the PN plot in Figure 2.9, we observe a large number of zeros. Moreover, seasonal patterns are hard to see by looking at daily values. To illustrate the seasonal patterns better, we look at the monthly averages for MT, mt and PN over the period 1995 to 2005 in Figures 2.10, 2.11 and 2.12. Now the seasonal patterns for precipitation can be seen better in Figure 2.12.

Next we look at the mean annual values of the three variables for all available years that have less than 10 missing days (Figures 2.13, 2.14 and 2.15). Table 2.1 gives a summary of these annual means.

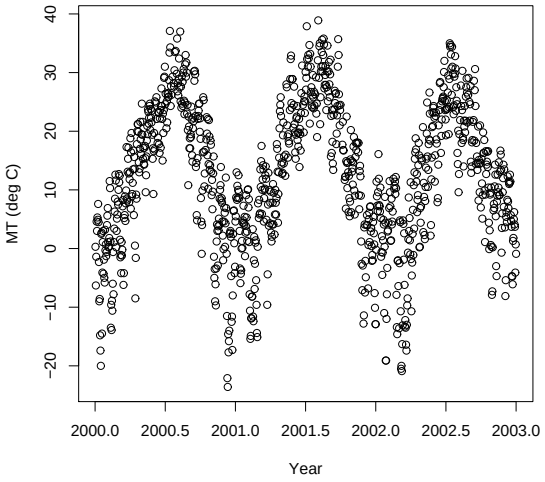

Figure 2.7: The time series of daily maximum temperature (deg C) at the Calgary site from 2000 to 2003.

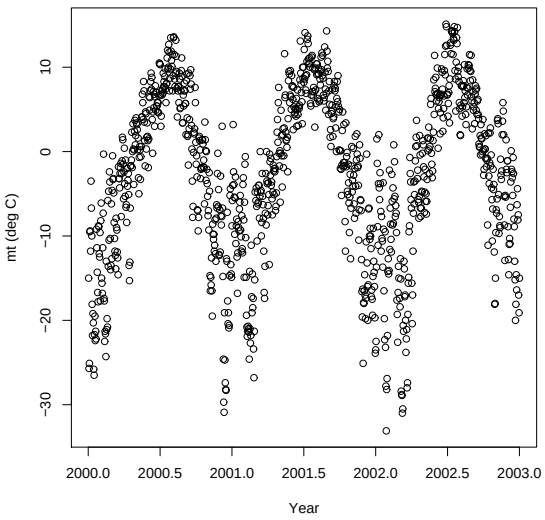

Figure 2.8: The time series of daily minimum temperature (deg C) at the Calgary site from 2000 to 2003.

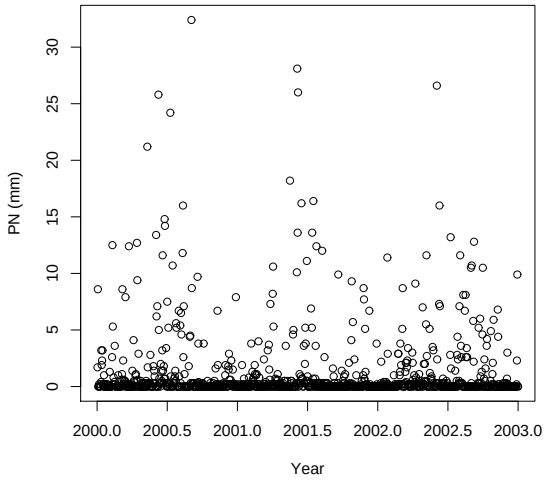

Figure 2.9: The time series of daily precipitation (mm) at the Calgary site from 2000 to 2003.

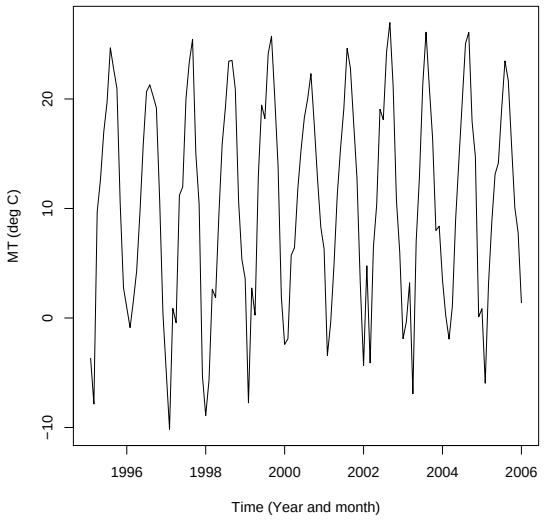

Figure 2.10: The time series of monthly maximum temperature (deg C) at the Calgary site, 1995–2005.

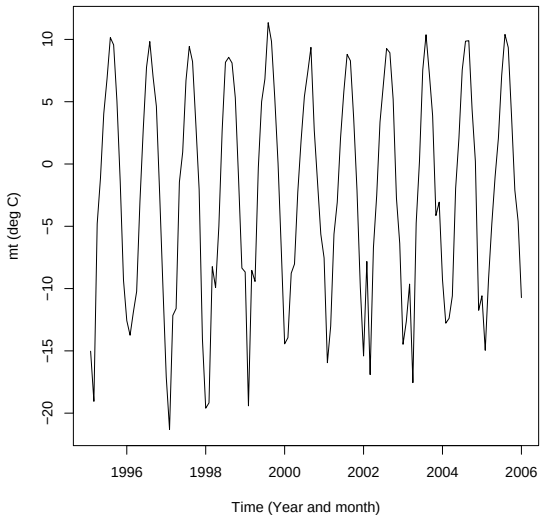

Figure 2.11: The time series of monthly minimum temperature means (deg C) at the Calgary site, 1995–2005.

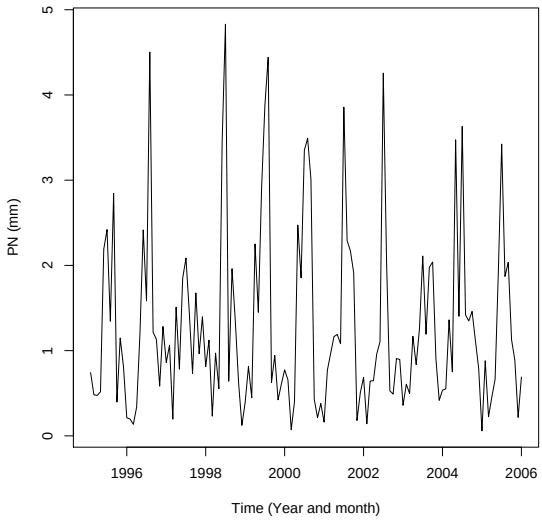

Figure 2.12: The time series of monthly precipitation means (mm) at the Calgary site, 1995–2005.

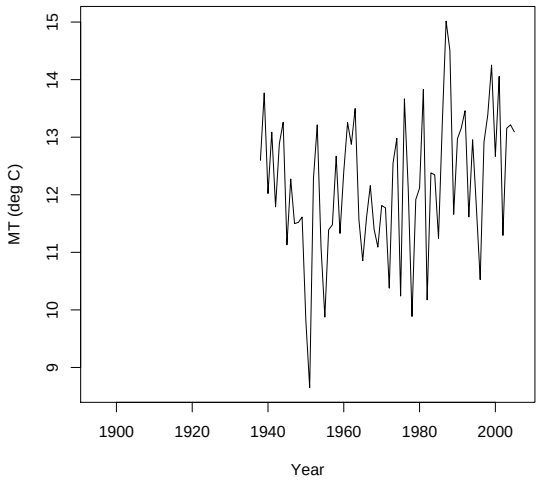

Figure 2.13: The annual mean maximum temperature (C) for Calgary site for all available years.

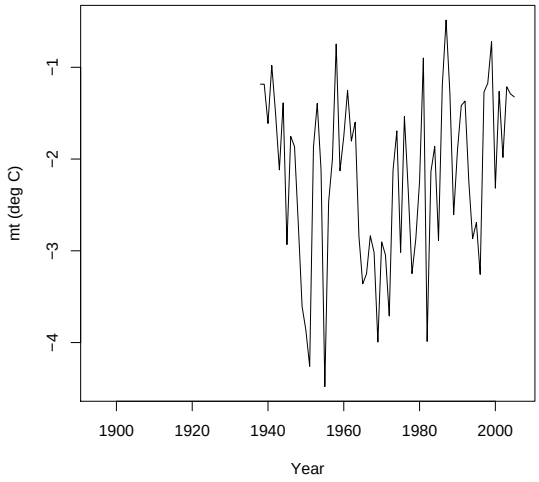

Figure 2.14: The annual mean minimum temperature (C) for Calgary site for all available years.

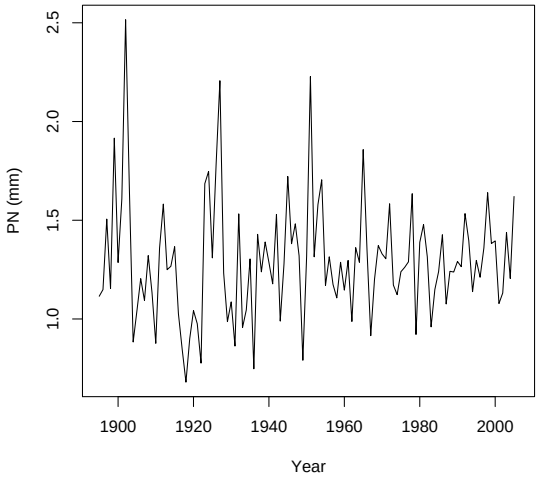

Figure 2.15: The annual mean precipitation (mm) for Calgary site for all available years.

Variable	min	1st quartile	median	mean	3rd quartile	max
MT (deg C)	7.59	9.64	10.37	10.36	11.19	13.46
mt (deg C)	-4.83	-3.40	-2.54	-2.66	-1.95	0.07
PN (mm)	0.68	1.12	1.28	1.29	1.39	2.51

Table 2.1: The summary statistics for the mean annual maximum temperature, min temperature and precipitation at the Calgary site.

Assuming stochastic normality and independence of the observations, we can obtain confidence intervals for all three variables and these are given in Table 2.2. The confidence intervals are fairly narrow.

Variable	95% confidence interval
MT (deg C)	(10.14,10.57)
mt (deg C)	(-2.85,-2.47)
PN (mm)	(1.24,1.35)

Table 2.2: Confidence intervals for the mean annual maximum temperature, min temperature and precipitation at the Calgary site.

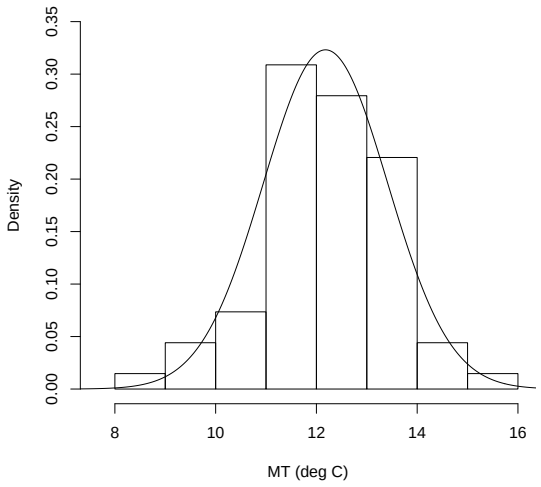

Figure 2.16: The histogram of annual maximum temperature means (deg C) for Calgary with a normal curve fitted to the data.

To investigate the shape of the distribution of annual means, we look at the histogram of each variable with a normal curve fitted in Figures 2.16, 2.18 and 2.20. The corresponding normal qq–plots (quantile–quantile) are also given in Figures 2.17, 2.19 and 2.21 to asses the normality assumption. Both the histogram and the qq–plots for MT validate the normality assumptions. The histogram for mt is slightly left skewed. For PN, some deviation from the normality assumption is seen. This is expected since the daily PN process is very far from normal to start with. Hence, even averaging through the whole year has not quite given us a normal distribution.

We plot all three variables (annual mean MT, mt and PN) in the same graph, Figure 2.22. As shown in that figure, MT and mt show the same trends over time. To get an idea of how the two variables are related, we fit a regression line, taking mt as response and MT as the explanatory variable. As seen in Figure 2.23, the regression fit looks very good. We repeat this analysis this time taking MT as explanatory variable and PN as response. As shown in Figure 2.24, the fit is still reasonable, but the association is not as strong. As shown in Table 3, both fits are significant. One can criticize the use of a simple regression since the independence assumption might not be satisfied. Finding more reliable and sensible relationships among the

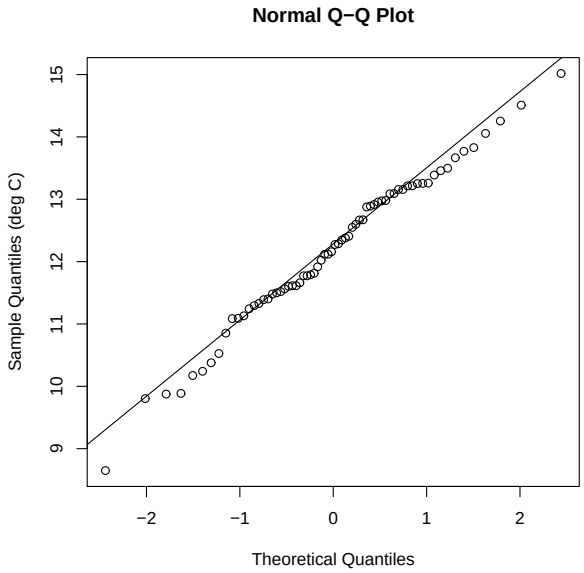

Figure 2.17: The normal qq–plot for annual maximum temperature means (deg C) for Calgary.

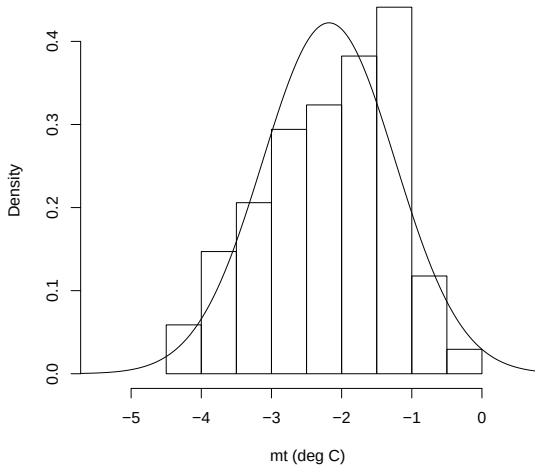

Figure 2.18: The histogram of annual minimum temperature means (deg C) for Calgary with normal curve fitted to the data.

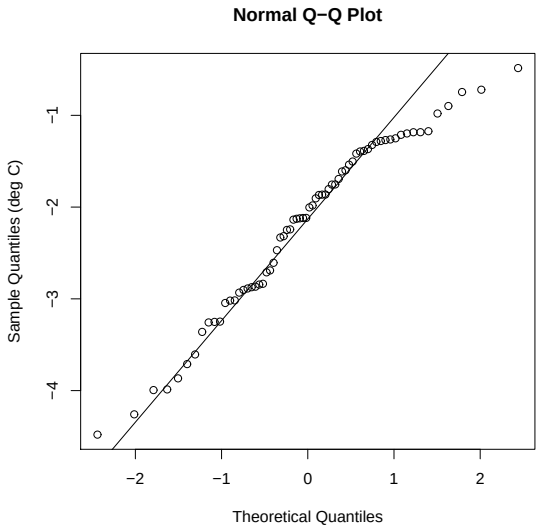

Figure 2.19: The normal qq–plot for annual minimum temperature means (deg C) for Calgary.

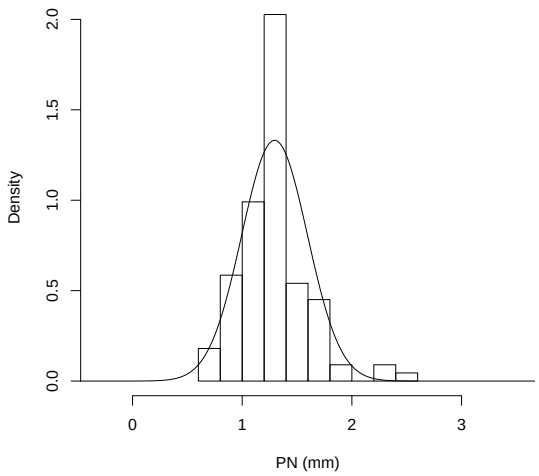

Figure 2.20: The histogram of annual precipitation means (mm) for Calgary with normal curve fitted to the data.

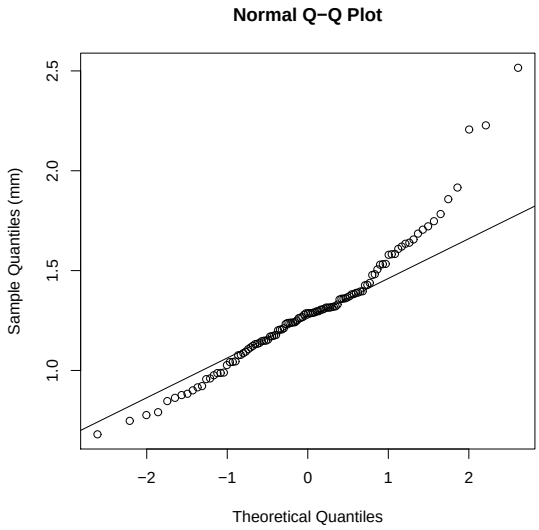

Figure 2.21: The normal qq–plot for annual precipitation means for Calgary.

variables needs a multivariate model taking account of correlation and other aspects of the processes. Also note that these are annual averages which are not as correlated as daily values over time as seen in the annual time series plots.

Variables	Intercept	Slope	p-value for intercept	p-value for slope
mt (deg C)	-10.40	0.746	2×10^{-16}	2×10^{-16}
PN (mm)	2.13	-0.082	1.49×10^{-14}	0.0005

Table 2.3: Lines fitted to annual mean minimum temperature and annual mean precipitation against annual mean maximum temperature.

Next we look at the change in the seasonal means for all three variables. As we noted above there are missing data particularly near the beginning of the time series. This has caused the gap at the beginning of most plots. To get a longer time series of means, we first compute the monthly means allowing 3 missing days and then compute the annual mean using the monthly means. This is reasonable since nearby days have similar values. We do the regression analysis for three locations: Calgary, Banff and Medicine Hat. We fit the regression line to annual means, spring means, summer means, fall means and winter means for each of MT, mt and PN with respect to

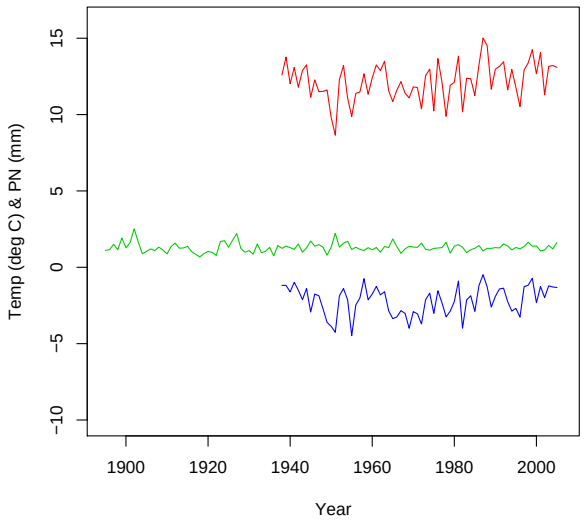

Figure 2.22: The time series plots of maximum temperature (deg C), minimum temperature (deg C) and precipitation (mm) annual means for Calgary. The time series plot in the bottom is minimum temperature, the one in the middle is precipitation and the top curve is maximum temperature.

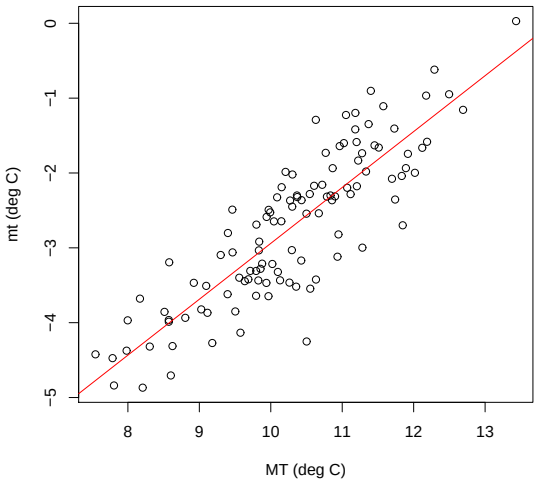

Figure 2.23: The regression line fitted to maximum temperature and minimum temperature annual means for Calgary.

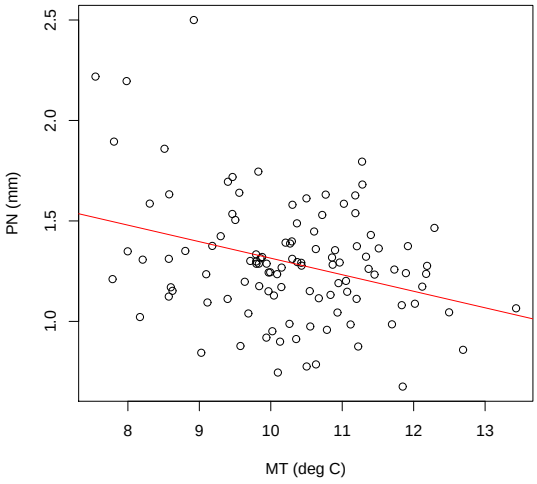

Figure 2.24: The regression line fitted to maximum temperature and precipitation annual means for Calgary.

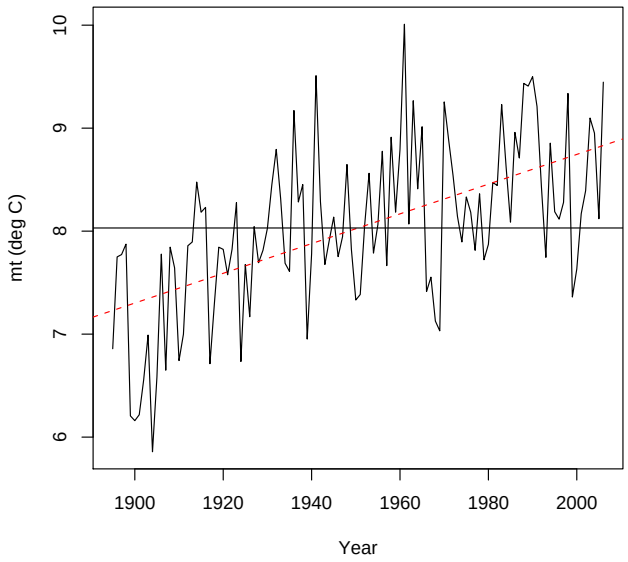

Figure 2.25: The regression line fitted to summer minimum temperature means against time for Calgary.

time. The results are given in Table 4, 5 and 6. We have only included fits that turned out to be significant. Note that PN does not appear in any of the tables. Annual minimum temperature and summer mean temperature show an increase in all three locations. Figure 2.25 depicts one of the time series (mt summer mean for Calgary) with the regression line fitted.

Variable	Season	Intercept	Slope	p-value for intercept	p-value for slope
mt (deg C)	Year	-24.72	0.112	$2 \times 10{-}05$	0.0001
mt (deg C)	Spring	-30.05	0.138	0.0008	0.0024
mt (deg C)	Summer	-20.11	0.0144	6×10^{-7}	3×10^{-11}

Table 2.4: The regression line parameters for the fitted lines for each variable with respect to time for the Calgary site.

Variable	Season	Intercept	Slope	p-value for intercept	p-value for slope
MT (deg C)	Year	-12.99	0.0105	0.019	0.0002
MT (deg C)	Spring	-17.0	0.0048	0.075	0.009
MT (deg C)	Fall	-12.64	0.0106	0.19	0.0326
mt (deg C)	Year	-37.0	0.01666	$2 \times 10{-}10$	2×10^{-8}
mt (deg C)	Spring	-49.8	0.0229	5×10^{-9}	10^{-7}
mt (deg C)	Summer	-36.8	0.0212	2×10^{-15}	2×10^{-16}

Table 2.5: The regression line parameters for the fitted lines for each variable with respect to time for the Banff site.

Variable	Season	Intercept	Slope	p-value for intercept	p-value for slope
MT (deg C)	Year	-24.6	0.0185	0.00102	3×10^{-6}
MT (deg C)	Spring	-34.24	0.0235	0.009	0.0005
mt (deg C)	Year	-39.98	0.0197	$5 \times 10{-}10$	2×10^{-9}
mt (deg C)	Spring	-39.81	0.0196	5×10^{-5}	9×10^{-5}
mt (deg C)	Summer	-10.93	0.0112	0.0199	7×10^{-6}
mt (deg C)	Fall	-24.66	0.0122	0.0110	0.0137

Table 2.6: The regression line parameters for the fitted lines for each variable with respect to time for the Medicine Hat site.

2.4 Daily values, distributions

This section studies the daily values for all three variables. To that end, we pick four days of the year, Jan 1st, April 1st, July 1st and October 1st. Let us look at the time series, histograms and normal qq–plots for each variable over the years. Figures 2.26 to 2.31 give the results. In fact the plots show that a normal distribution fits the data for daily MT and mt for the the selected days fairly well. However, some deviations from the normal distribution is seen, particularly in the tails. We also tried the first day of each month and observed similar results.

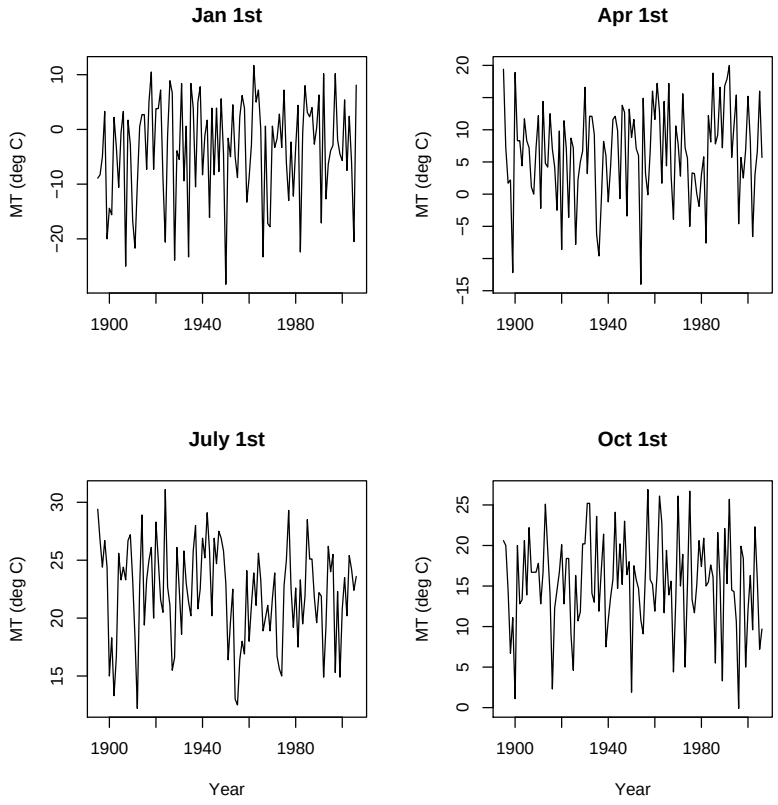

Figure 2.26: The time series of daily maximum temperature at the Calgary site for four given dates: January 1st, April 1st, July 1st and October 1st.

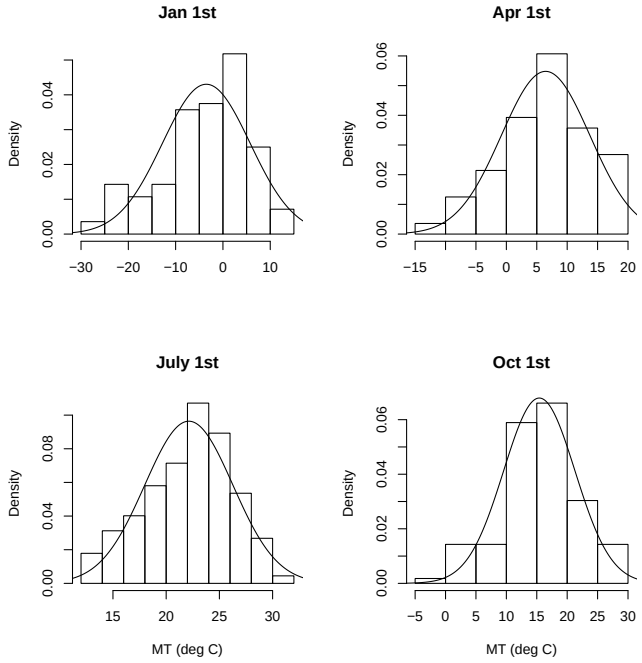

Figure 2.27: The histogram of daily maximum temperature at the Calgary site for four given dates: January 1st, April 1st, July 1st and October 1st.

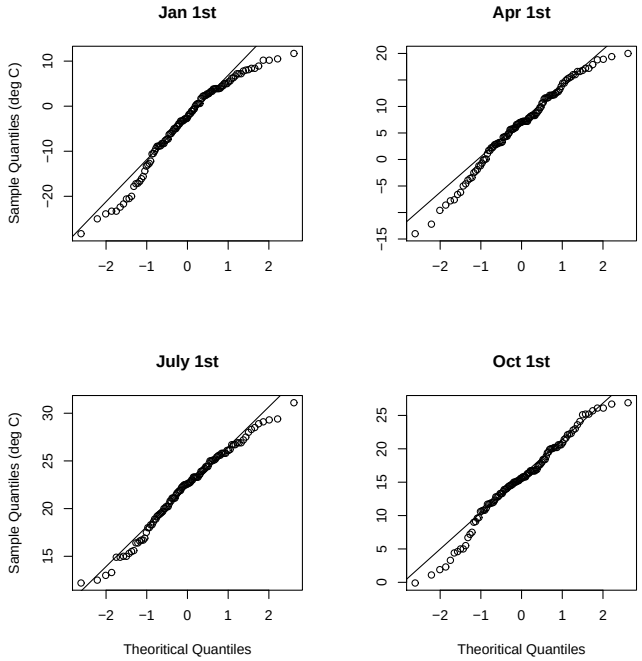

Figure 2.28: The normal qq–plots of of daily maximum temperature at the Calgary site for four given dates: January 1st, April 1st, July 1st and October 1st.

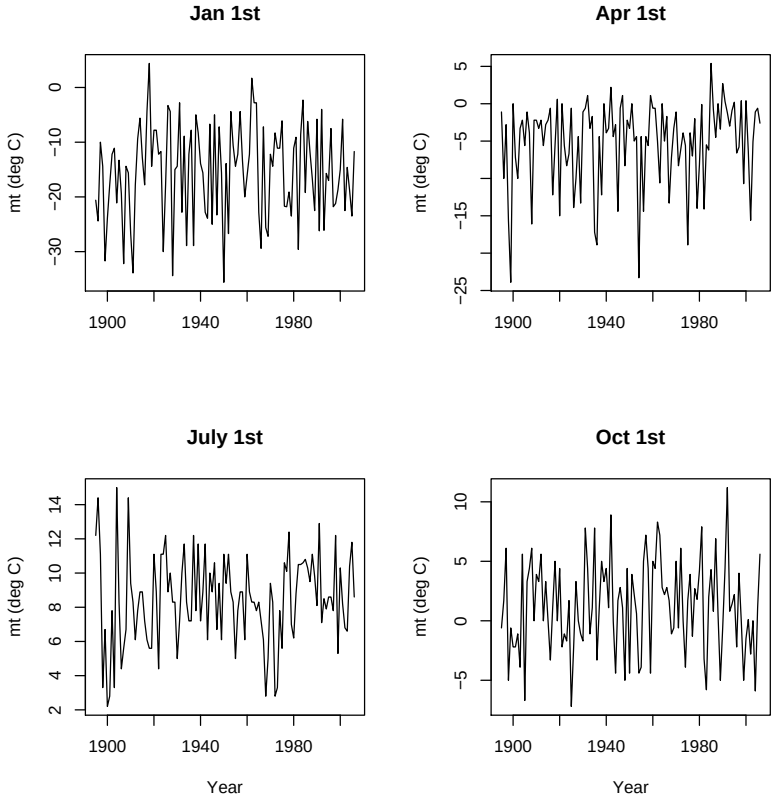

Figure 2.29: The time series of daily minimum temperature for Calgary for four given dates: January 1st, April 1st, July 1st and October 1st.

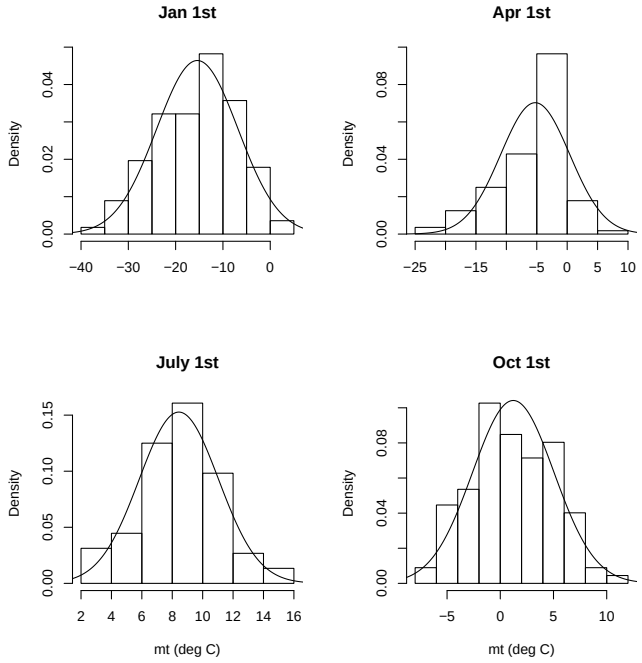

Figure 2.30: The histogram of daily minimum temperature at the Calgary site for four given dates: January 1st, April 1st, July 1st and October 1st.

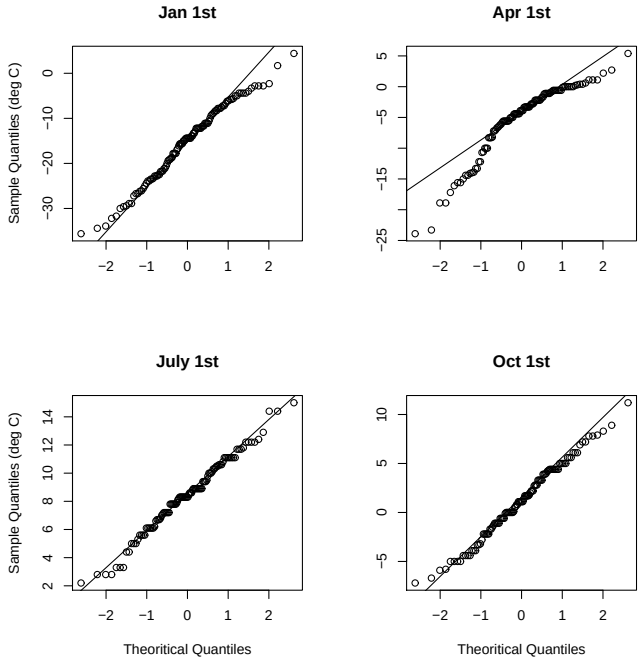

Figure 2.31: The normal qq-plots of daily minimum temperature at the Calgary site for four given dates: January 1st, April 1st, July 1st and October 1st.

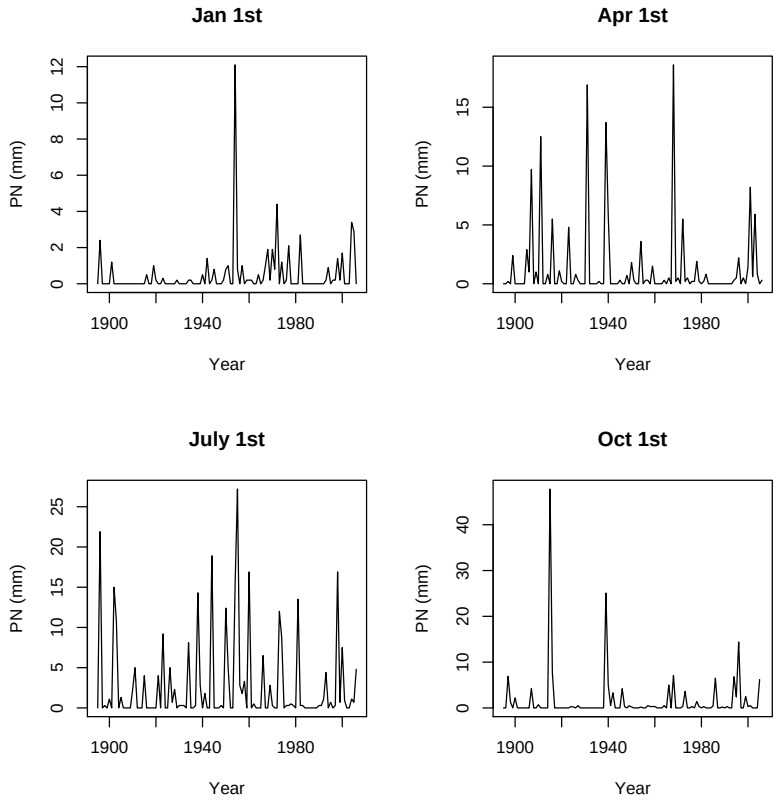

Figure 2.32: The time series of daily precipitation at the Calgary site for four given dates: January 1st, April 1st, July 1st and October 1st.

We plot the histogram for PN as well (Figure 2.33). The distribution is far from normal because of high frequency of no PN (dry) days.

Next, we use the available years to compute the confidence intervals for the mean of every given day of the year for MT and mt. For PN, we construct the confidence intervals for probability of PN. [A PN day is defined to be a day with $PN > 0.2$ (mm). This is because any precipitation amount less than 0.2 (mm) is barely measurable.] Figures 2.34 to 2.36 give the confidence intervals for the means. The confidence interval for the standard deviations (obtained by bootstrap techniques) are given in Figures 2.37 to 2.39. A regular seasonal pattern is seen in the means and standard deviations. For example the maximum for MT and mt occurs around the 200th (in July) day of the year and the minimum occurs at the beginning

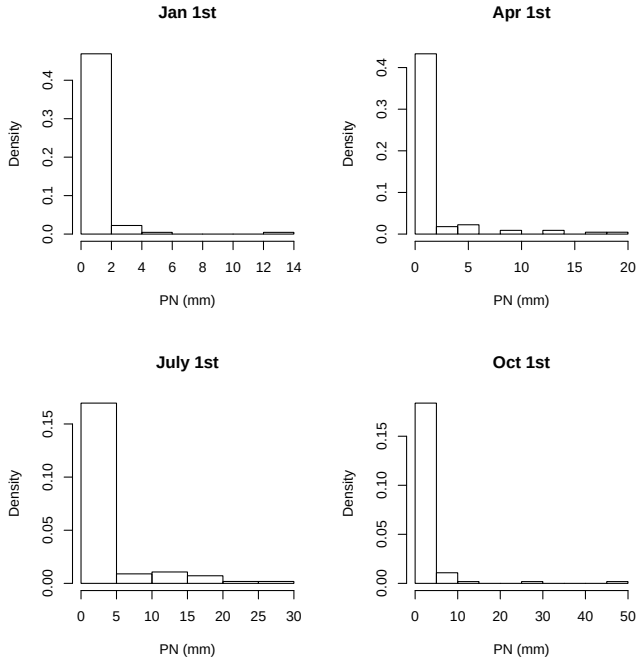

Figure 2.33: The histogram of daily precipitation at the Calgary site for four given dates: January 1st, April 1st, July 1st and October 1st.

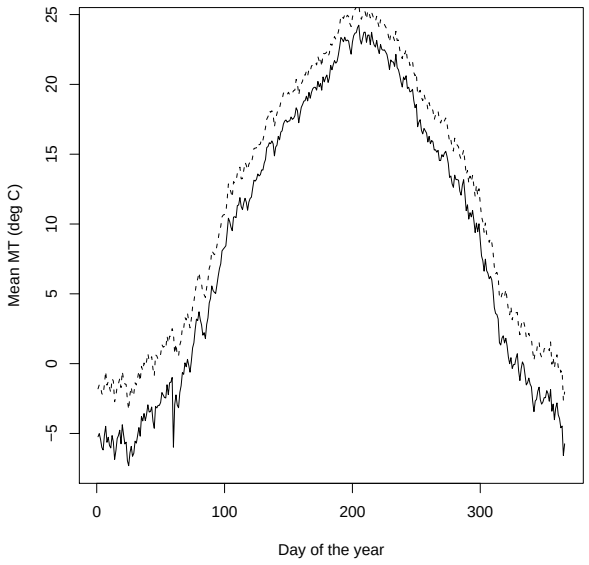

Figure 2.34: The confidence intervals for the daily mean maximum temperature (deg C) at the Calgary site. Dashed line shows the upper bound and the solid line the lower bound of the confidence intervals.

and the end of the year. Comparing the plots of the means and the standard deviations, we observe that warmer days have smaller standard deviations than colder days. For example the minimum standard deviation for the Maximum and minimum temperature happens around the 200th day of the year which correspond to the warmest period of the year. The plots seem to indicate that a simple periodic function suffices to model the seasonal patterns. Contrary to MT and mt, for the 0-1 PN process, the standard deviation is the highest in June, when the probability of precipitation is close to $\frac{1}{2}$.

As shown above, the distribution of daily PN values is far from normal. This time, after removing the zeros, we fit a Gamma distribution to PN (Figure 2.42). The Gamma qq–plots are given in Figure 2.43 and reveal a fairly good fit.

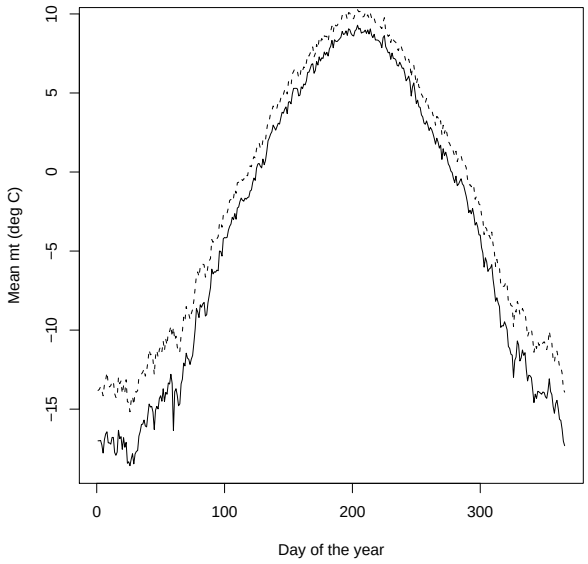

Figure 2.35: The confidence intervals for the daily mean minimum temperature (deg C) at the Calgary site. Dashed line shows the upper bound and the solid the lower bound of the confidence intervals.

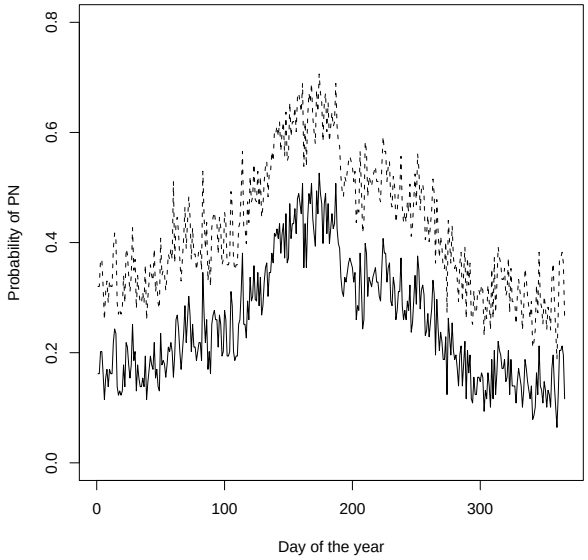

Figure 2.36: The confidence intervals for the probability of precipitation (mm) at the Calgary site for the days of the year. Dashed line shows the upper bound and the solid the lower bound of the confidence intervals.

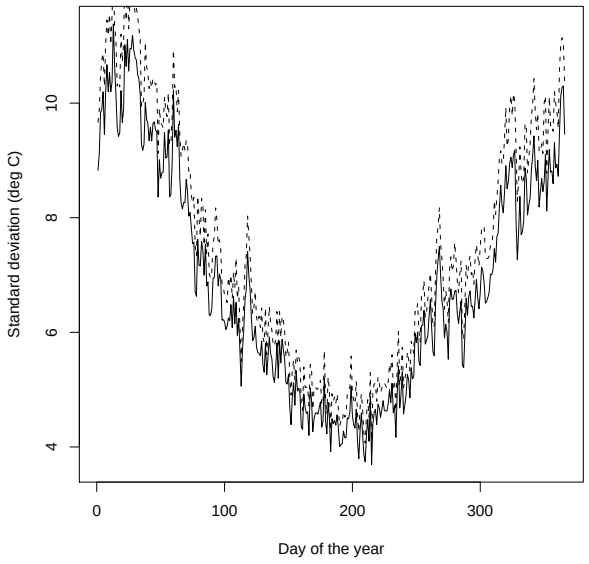

Figure 2.37: The confidence intervals for the standard deviation of each day of the year for maximum temperature (deg C) at the Calgary site. Dashed line shows the upper bound and the solid the lower bound of the confidence intervals.

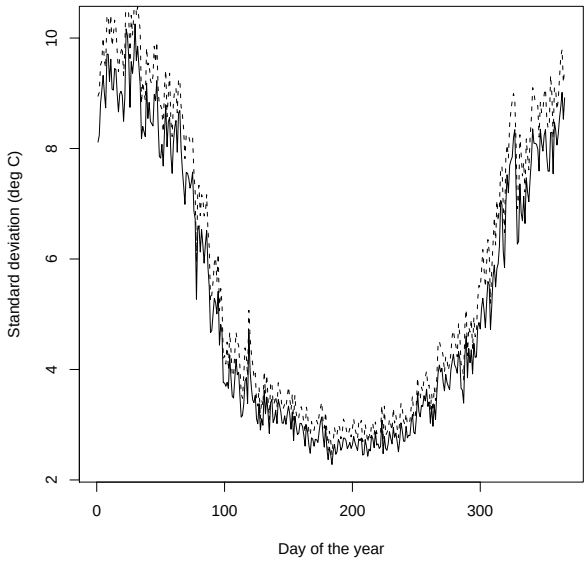

Figure 2.38: The confidence intervals for the standard deviation of each day of the year for minimum temperature (deg C) at the Calgary site. Dashed line shows the upper bound and the solid the lower bound of the confidence intervals.

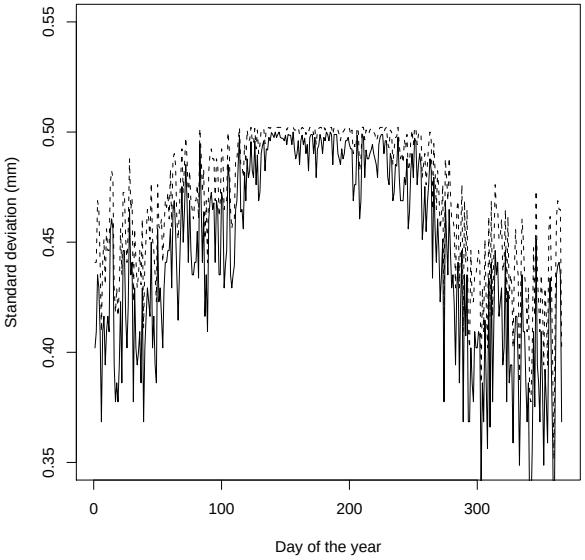

Figure 2.39: The confidence intervals for standard deviation (sd) of each day of the year for the probability of precipitation (mm) (0-1 precipitation process) at the Calgary site. Dashed line shows the upper bound and the solid the lower bound of the confidence intervals. Plot shows $sd \leq 1/2$. This is because $sd = \sqrt{p(1-p)}$ which has a maximum value of $\frac{1}{2}$.

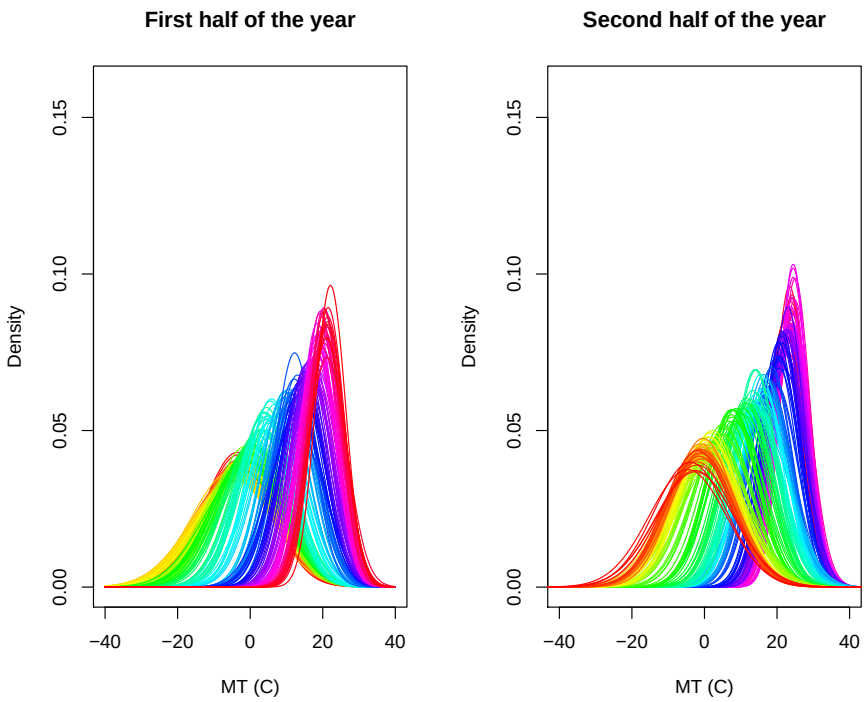

Figure 2.40: The distribution of each day of the year for MT (C) from Jan 1st to Dec 1st. The year has been divided to two halves. In each half rainbow colors are used to show the change of the distribution.

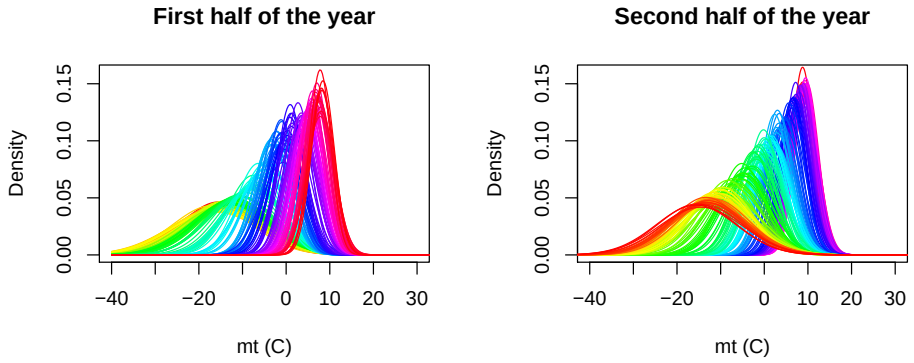

Figure 2.41: The distribution of each day of the year for mt (C) from Jan 1st to Dec 1st. The year has been divided to two halves. In each half rainbow colors are used to show the change of the distribution.

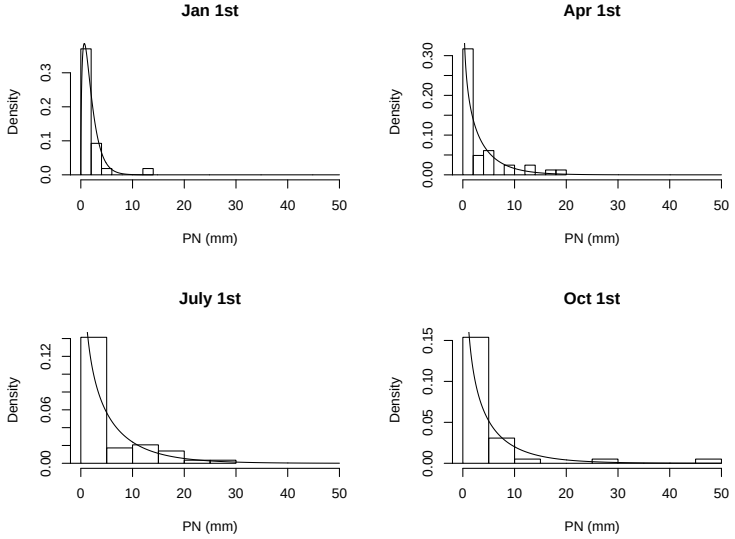

Figure 2.42: The histogram of daily precipitation greater than 0.2 mm at the Calgary site with Gamma density curve fitted using Maximum likelihood.

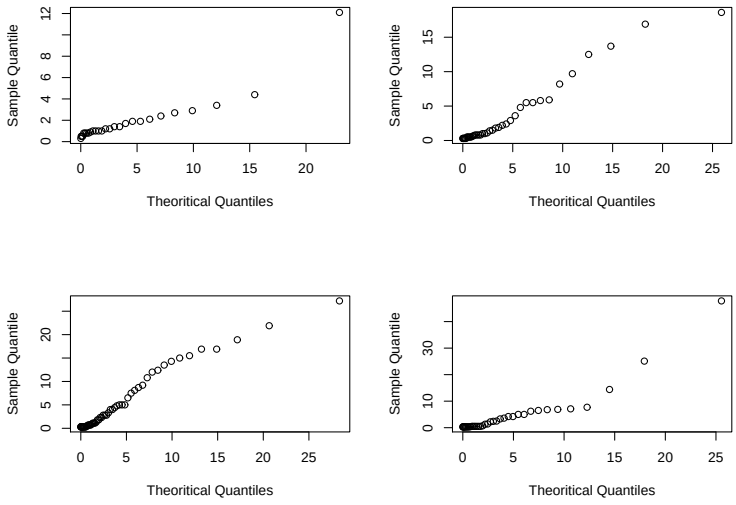

Figure 2.43: The qq–plots of daily precipitation greater than 0.2 mm at the Calgary site with Gamma curve fitted using Maximum likelihood.

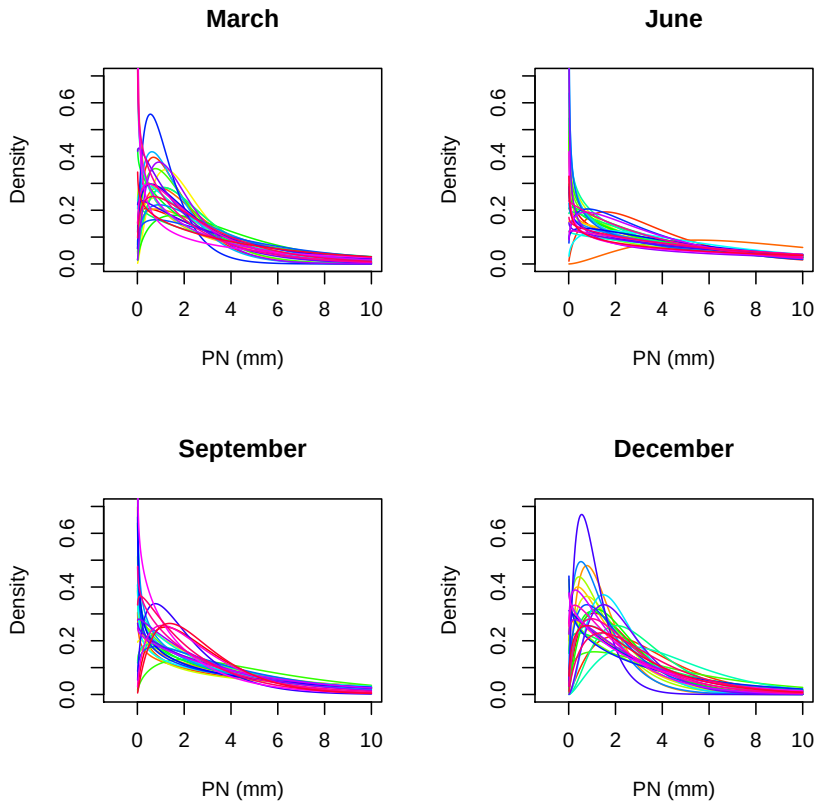

Figure 2.44: The Gamma fit of each day of 4 months for precipitation (mm). In each month rainbow colors are used to show the change of the distribution.

Figures 2.40, 2.41 and 2.44 reveal the result of our investigation of the change in the distribution over a period of time. For MT and mt, we have done that for the course of the year. The figures show how the distribution deforms continuously over the year. We can also notice changes in mean and standard deviation over the year. For PN, we have done the same only for 4 different months because of high irregularity of the process.

Next, we look at the parameters of the Gamma distribution fitted to PN over the course of a year. If we use maximum likelihood estimates (MLE), which we have used above to form the Gamma curve, the confidence intervals, obtained by bootstrap method will be very wide (tend rapidly to infinity). Hence, we use the "method of moments estimates" (MOM), to

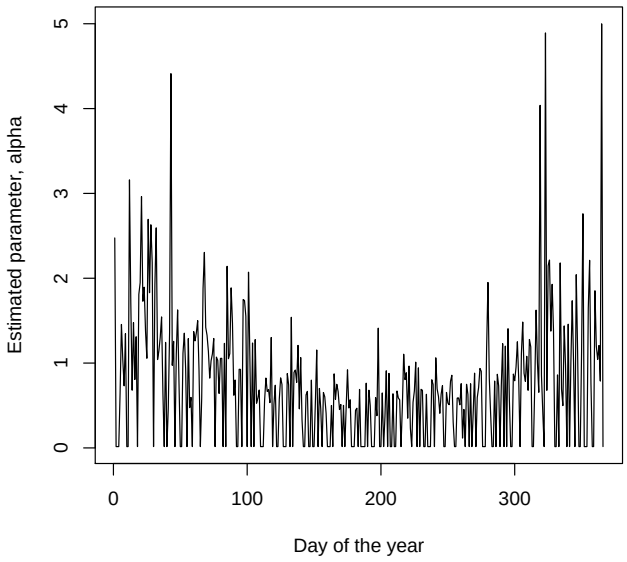

Figure 2.45: The maximum likelihood estimate for α, the shape parameter of the Gamma distribution fitted to the precipitation amounts.

obtain confidence intervals. The MOM confidence intervals are given in Figure 2.46. When using MLE estimates, since there is no closed form for them, we need to use Newton method to find the Maximum values. However, MOM gives us closed form solution. This advantage might explain the better behavior of MOM estimates in forming the confidence intervals. However, even the MOM confidence intervals do not look satisfactory and are rather wide and irregular specially at the beginning and end of the year.

We can also consider the 0-1 process of PN (1 for wet and 0 for dry) and compute the transition probabilities for PN (Figure 2.47). The figure shows the probability of PN is changing continuously over the year and can be modeled by a simple periodic function.

Considering the 0-1 process of PN as a chain leads to the interesting question as the order of the Markov chain. Let us denote by 1 a PN occurrence and 0 otherwise. Suppose $x_t = 1$ denote PN on day t and $x_t = 0$ denote no PN and let $p_{x_{t-r}\cdots x_t}(t)$, denote the probability of observing x_t on day t of the year conditional on the chain $(x_{t-r} \cdots x_{t-1})$. In Figure 2.47, we have plotted the estimated $\hat{p}_{11}(t)$ and $\hat{p}_{01}(t)$ for different days of the year.

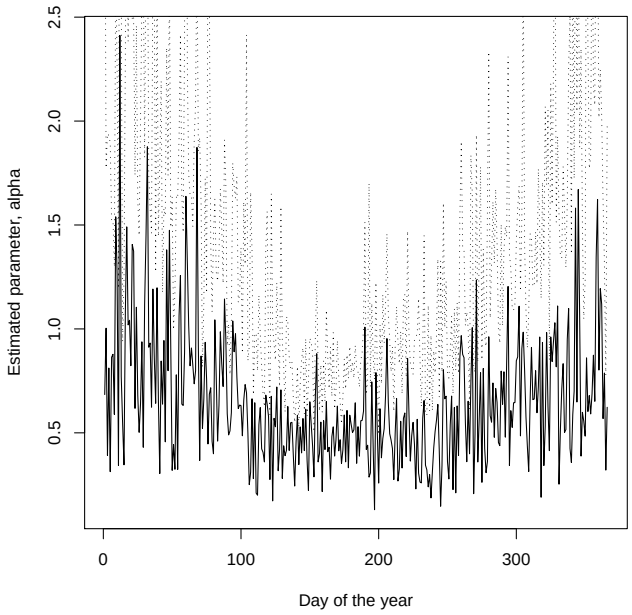

Figure 2.46: The confidence interval for MOM estimate of the shape parameter, α, of the Gamma distribution fitted to daily precipitation amounts. The dotted line is the upper bound and the solid line the lower bound. As seen in the figure the upper bounds at the beginning and end of the year have become very large. We have not shown them because otherwise then the pattern in the rest of the year could not be seen.

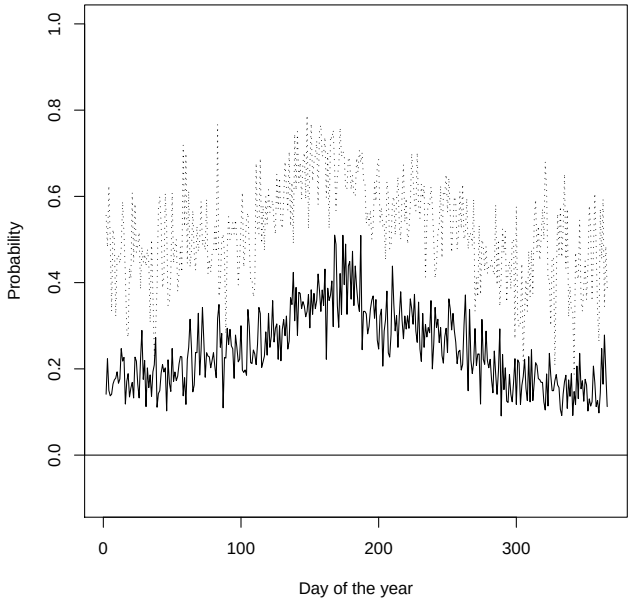

Figure 2.47: The 1st-order transition probabilities. The dotted line is the the probability of precipitation if it happened the day before ($\hat{p_{11}}$) and the dashed is the probability of precipitation if it did not happen the day before ($\hat{p_{01}}$).

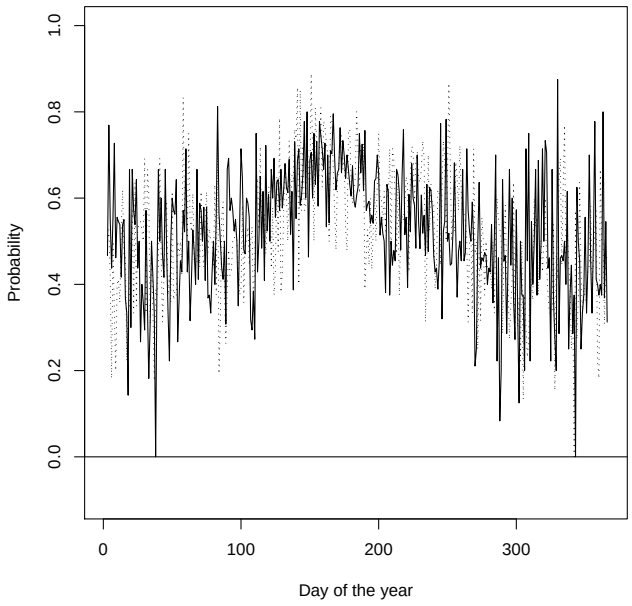

Figure 2.48: The 2nd–order transition probabilities for the precipitation at the Calgary site: $\hat{p}_{111}$ (solid) against $\hat{p}_{011}$ (dotted).

The clear gap between this two estimated probabilities indicate that a 1st–order Markov chain should be preferred to a 0th–order. Figures 2.48 and 2.49 plot $\hat{p}_{111}$ against $\hat{p}_{011}$ and $\hat{p}_{001}$ against $\hat{p}_{101}$. The estimated probabilities seem to be close and overlap heavily over the course of the year. Hence a 1st–order Markov chain seems to suffice for describing the binary process of PN.

2.5 Correlation

The correlation in a spatial–temporal process can depend on time and space. This section studies the temporal and spatial patterns of the correlation function separately.

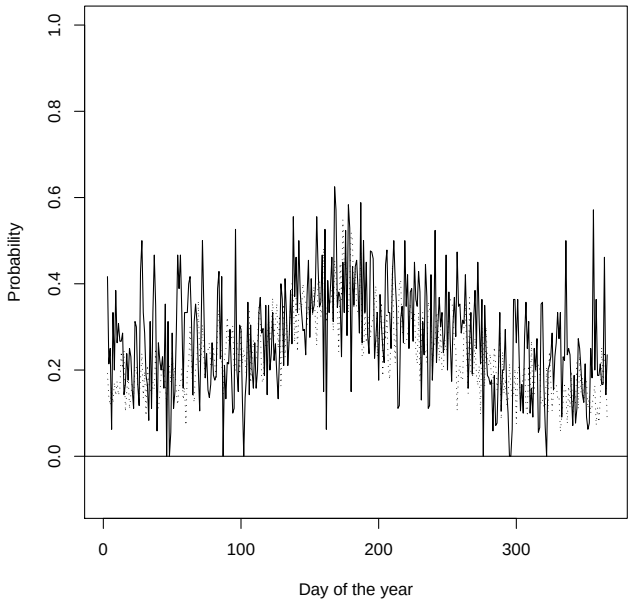

Figure 2.49: The 2nd–order transition probabilities for the precipitation at the Calgary site: $\hat{p}_{001}$ (solid) against $\hat{p}_{101}$ (dotted).

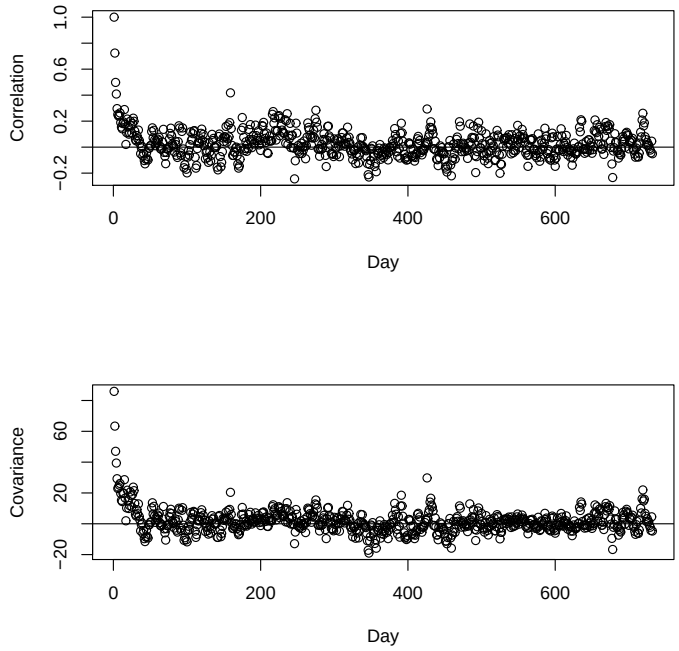

Figure 2.50: The correlation and covariance plot for maximum temperature at the Calgary site for Jan 1st and 732 consequent days.

2.5.1 Temporal correlation

Here we look at the correlation/covariance of the variables as a function of time. The location is taken to be the Calgary site. First we look at the correlation/covariance of a given day and its consequent days. We pick Jan 1st and compute the correlation/covariance with the following days: Jan 2nd, Jan 3rd and etc. Figure 2.50 shows that the correlation and covariance have the same trends for maximum temperature. Figures 2.51 to 2.53 show a decreasing trend for correlation over time for MT, mt and PN. The decrease is far from linear and it looks to be exponentially decreasing. The plots also indicate that only a few consequent days are possibly correlated and in particular two days that are one year apart can be considered independent. This assumption might be useful in building a spatial–temporal model.

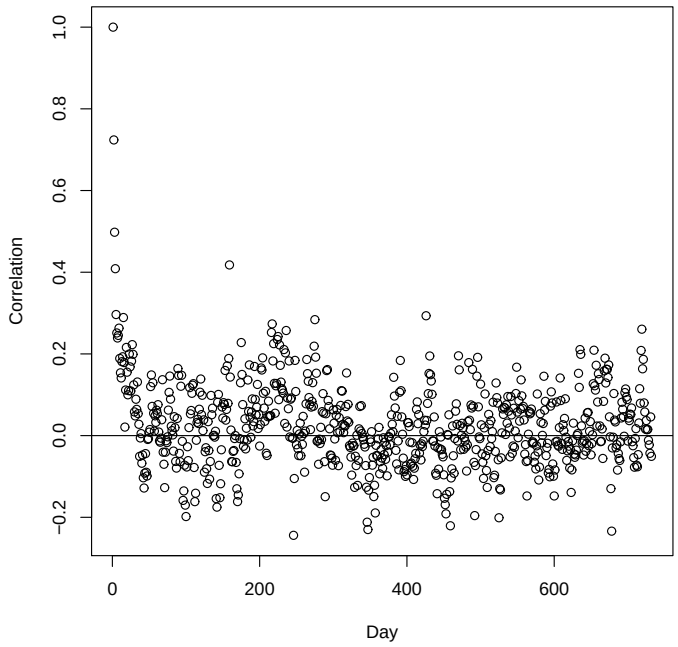

Figure 2.51: The correlation plot for maximum temperature (deg C) at the Calgary site for Jan 1st and 732 consequent days.

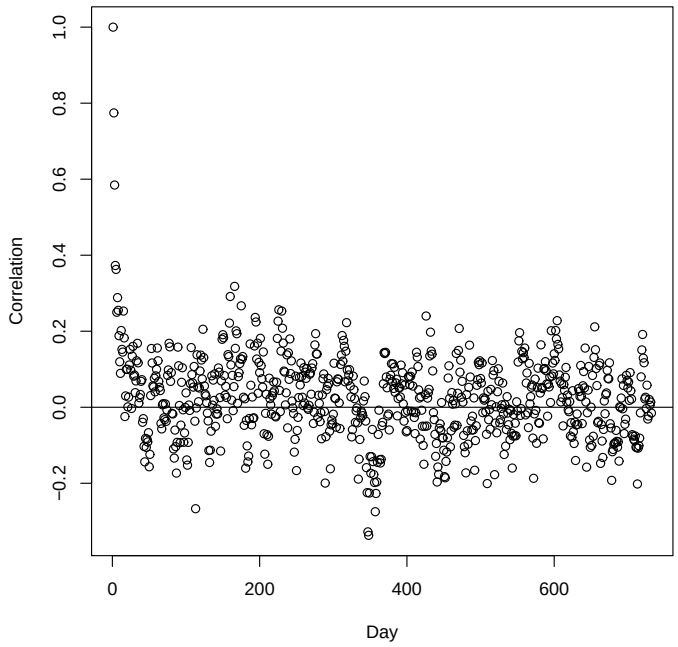

Figure 2.52: The correlation plot for minimum temperature (deg C) at the Calgary site for Jan 1st and 732 consequent days.

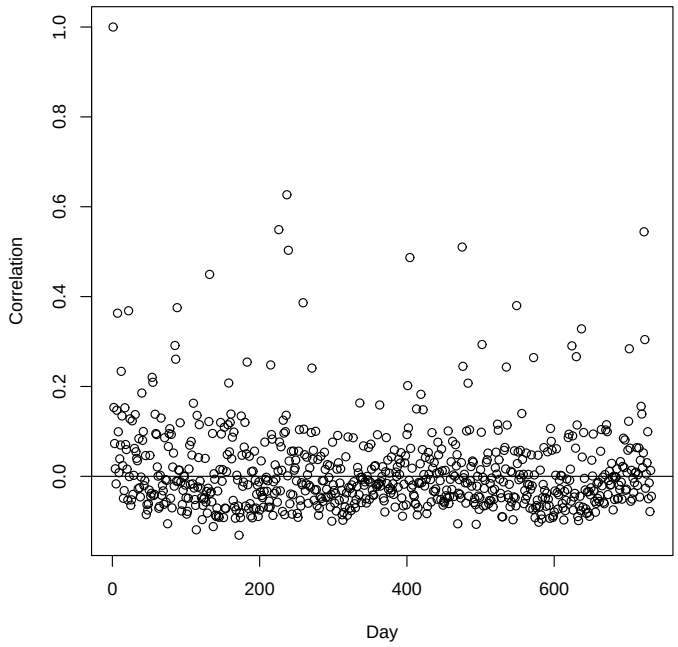

Figure 2.53: The correlation plot for precipitation (mm) at the Calgary site for Jan 1st and 732 consequent days.

Next we look at the correlation of responses on other days of the year with their 30 consecutive days. Our goal is to see if the correlation function has the same behavior over the course of a year. We pick, Feb 1st, April 1st, July 1st, Oct 1st. Figures 2.54 and 2.56 show similar patterns.

Finally, we look at the correlation of two fixed locations over the course of the year (by changing the day). The results are given in Figures 2.57 and 2.58. Strong correlation and clear seasonal patterns are seen for MT and mt. This seems to indicate in particular that the temperature process is not stationary. The correlation in the middle of the year around day 200 which corresponds to the summer season seems to be smaller than the correlation at the beginning and end of the year which correspond to the cold season.

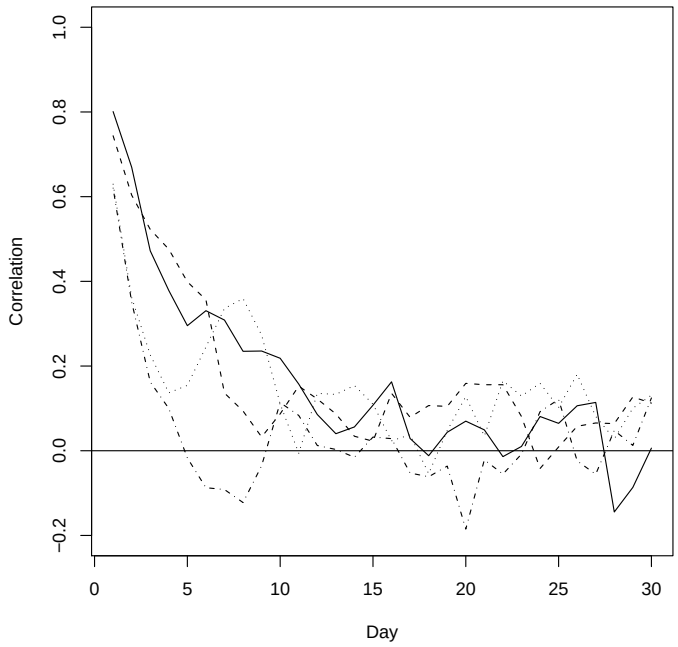

Figure 2.54: The correlation plot for maximum temperature (deg C) at the Calgary site for Feb 1st (solid), April 1st (dashed), July 1st (dotted) and Oct 1st (dot dash) and 30 consequent days.

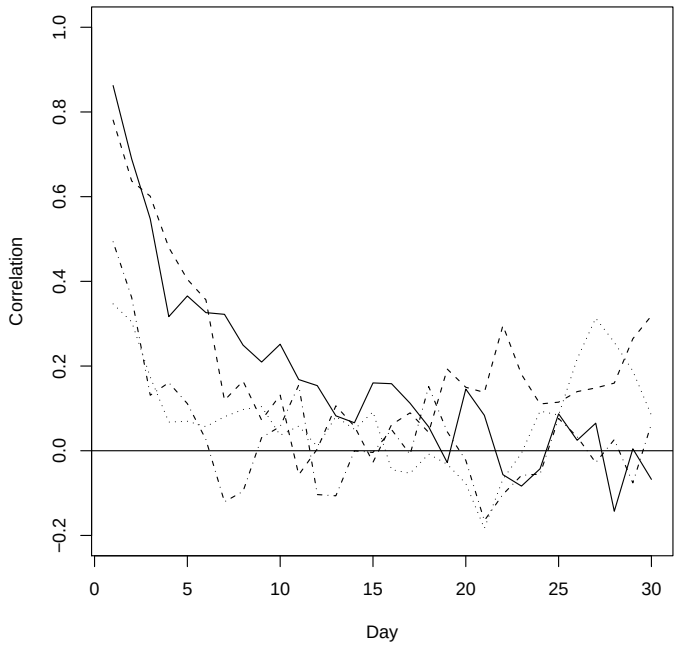

Figure 2.55: The correlation plot for minimum temperature (deg C) at the Calgary site for Feb 1st (solid), April 1st (dashed), July 1st (dotted) and Oct 1st (dot dash) and 30 consequent days.

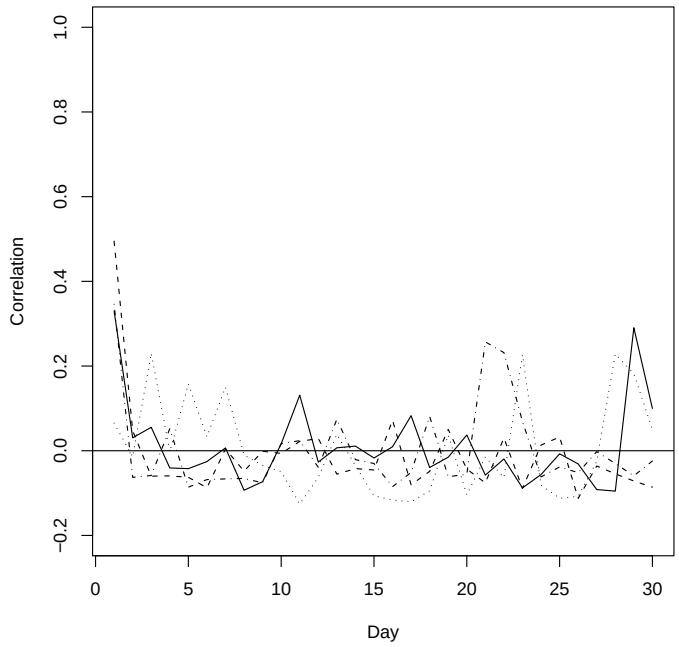

Figure 2.56: The correlation plot for precipitation (mm) at the Calgary site for Feb 1st (solid), April 1st (dashed), July 1st (dotted) and Oct 1st (dot dashed) and 30 consequent days.

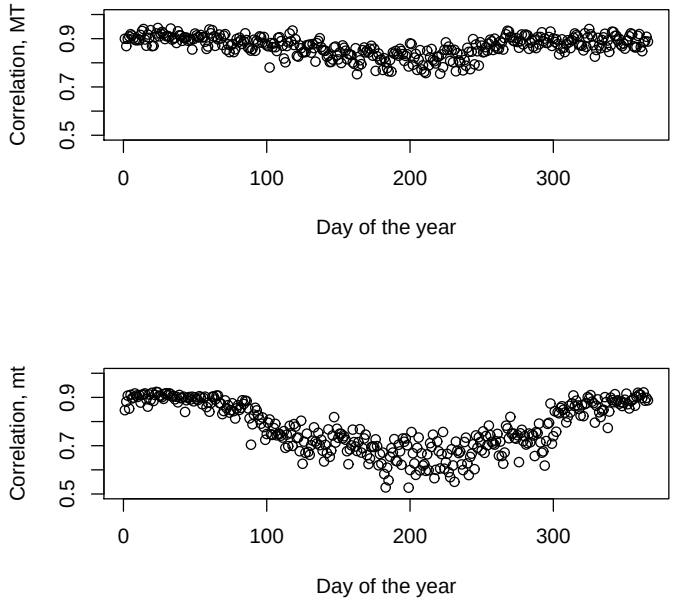

Figure 2.57: The correlation plot for maximum temperature and minimum temperature (deg C) between Calgary and Medicine Hat.

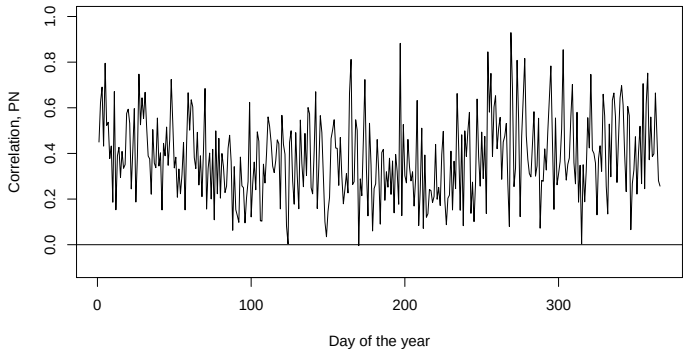

Figure 2.58: The correlation plot for precipitation (mm) between Calgary and Medicine Hat.

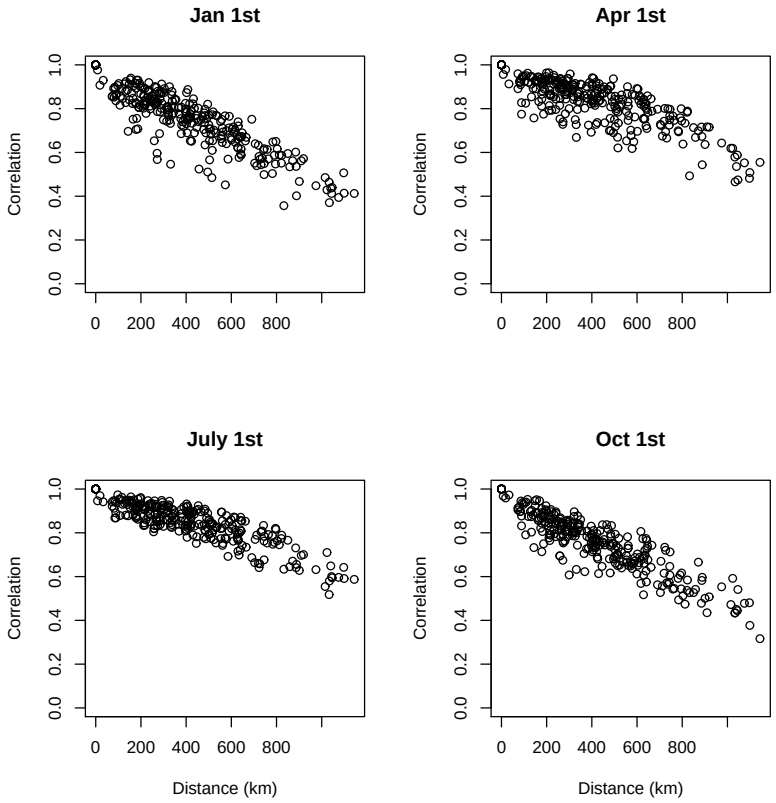

Figure 2.59: The correlation plot for maximum temperature (deg C) with respect to distance (km).

2.5.2 Spatial correlation

This subsection looks at the spatial correlation by fixing the time to a few dates: January 1st, April 1st, July 1st and Oct 1st distributed over year's climate regime. We plot the correlation with respect to the geodesic distance (km) on the surface of the earth. Figures 2.59 to 2.62 show the results for MT, mt, PN and 0-1 PN respectively. For MT and mt, we observe a clear decreasing trend with respect to distance. The trend for PN does not seem to be regular.

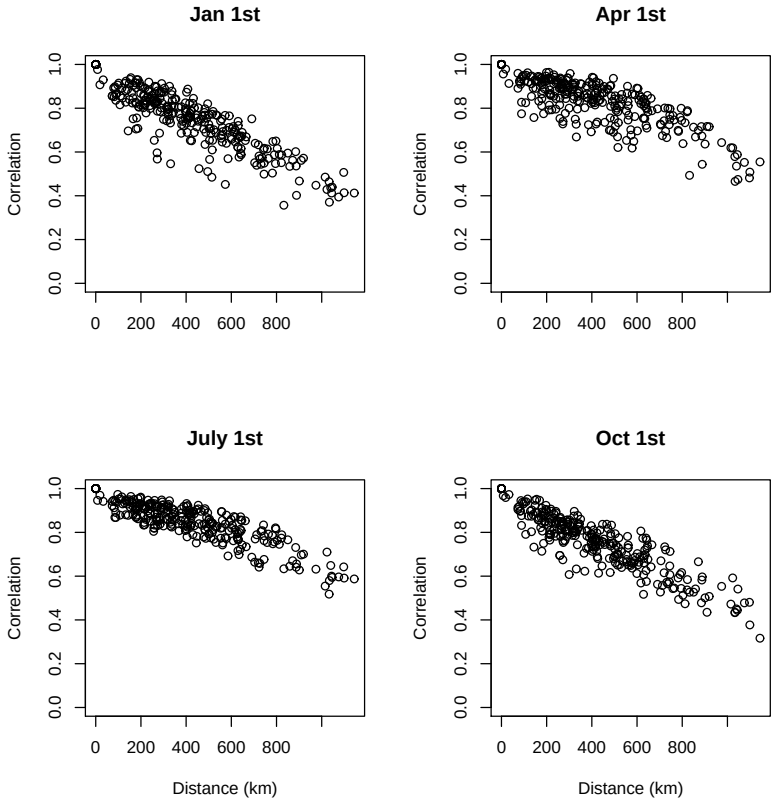

Figure 2.60: The correlation plot for minimum temperature (deg C) with respect to distance(km).

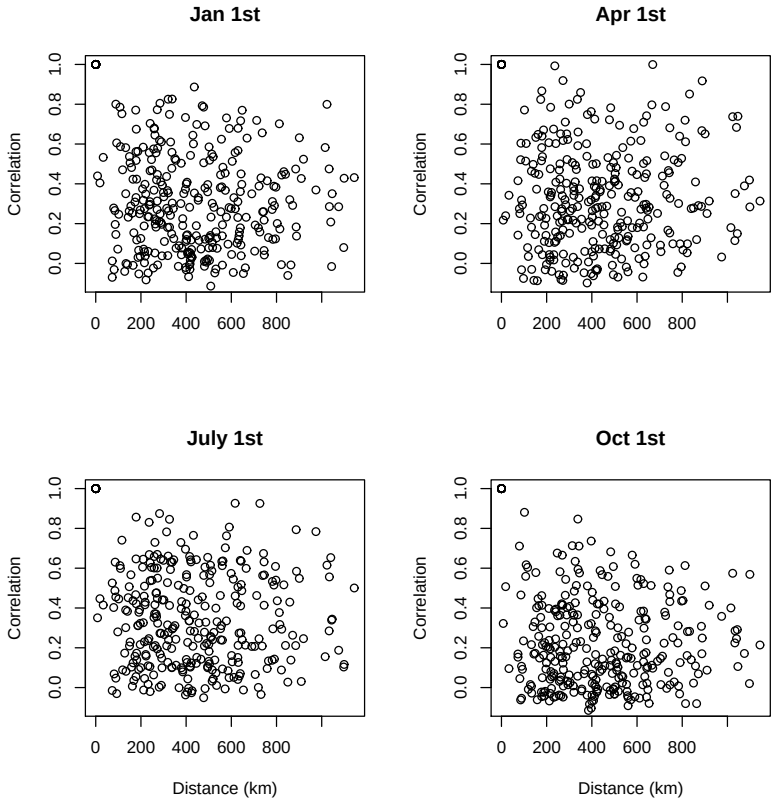

Figure 2.61: The correlation plot for precipitation (mm) with respect to distance (km).

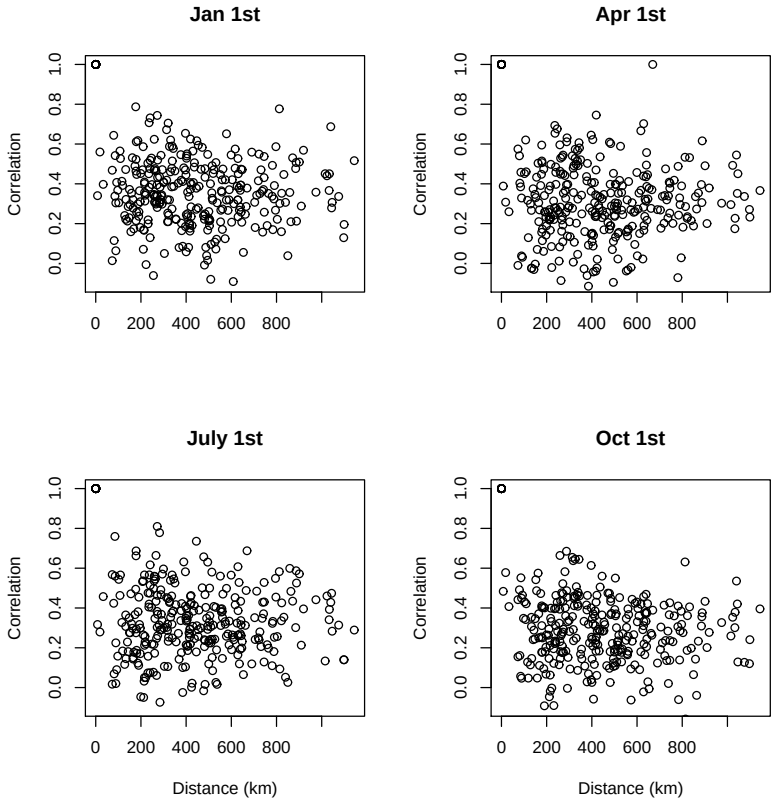

Figure 2.62: The correlation plot for precipitation (mm) 0-1 process with respect to distance (km).

2.6 Summary and conclusions

This section summarizes our findings of the exploratory analysis.

- There is a strong seasonal trend in the temperature and precipitation processes. See Figures 2.7, 2.8, 2.11 and 2.36.
- The summer average minimum temperature has increased over several locations over the past century. See Figure 2.25.
- mt and MT are highly correlated. See Figure 2.23.
- The distributions of daily maximum temperature and minimum temperature are rather close to the Gaussian distribution in the center with some deviations seen in the tails. See Figures 2.27 and 2.29.
- The temperature process in Alberta is less variable in the warm seasons and the converse holds for the precipitation process. See Figures 2.37, 2.38 and 2.39.
- The distribution of the daily temperature varies continuously over the course of the year. This could not be shown for precipitation. (This might be because we need more data.)
- The correlation between two sites depends on the time of the year. They are more correlated in cold seasons. This might be because there are more (strong) global weather regimes in the cold seasons influencing the whole region.
- The correlation over time for MT, mt and PN seems stationary and is decreasing with a nonlinear trend (exponentially) with respect to the time difference.
- The spatial correlations for MT and mt are strong and decreasing almost linearly with respect to the geodesic distance.
- The spatial correlation for PN is not strong. It might be because the sites are too faraway to capture the spatial correlation for PN.

The future chapters investigate some of these items. In particular after developing some theory regarding Markov chains, we investigate the order of the binary precipitation process. Then we will turn to modeling the occurrence of extreme temperature. Instead of using a Gaussian process to

model the temperature and use that to infer about the occurrence of the extremes, we use a categorical chain. This is because of the deviations from normality in the tails as pointed out above.

Chapter 3

rth-order Markov chains

3.1 Introduction

This chapter studies rth–order categorical Markov chains and more generally, categorical discrete–time stochastic processes. By "categorical", we mean chains that have a finite number of possible states at each time point. Such chains have important applications in many areas, one of which is modeling weather processes such as precipitation over time. In fact, we use these chains to model the binary process of precipitation as well as dichotomized temperature processes. In rth–order Markov chains, the conditional probability of the present given the past is modeled. Such a conditional probability is a function of the past r states, where each one of them only takes finite possible values.

It is useful and intuitively appealing to specify or model a discrete process over time by the conditional probabilities rather than the joint distribution. However, one must check the consistency of such a specification i.e. to prove that it corresponds to a full joint distribution. In the case of discrete–time categorical processes, we prove a theorem that shows the conditional probabilities can be used to specify the process. Also we prove a representation theorem which states that every such conditional probability after an appropriate transformation can be written as a linear summation of monomials of the past processes. In fact, we represent all categorical discrete–time stochastic processes over time, in particular rth–order Markov chains and more particularly stationary rth–order Markov chains. For the binary case the result is a consequence of an expansion theorem due to Besag [6]. To generalize the result to arbitrary categorical Markov chains, we prove a new expansion theorem which generalizes the result to the case of arbitrary categorical rth–order Markov chains (rather than binary only).

The result simplifies the task of modeling categorical stochastic processes. Since we have written the conditional probability as a linear combination, we can simply add other covariates as linear terms to the model to build non-stationary chains. For example, we can add seasonal terms or geographical coordinates (longitude and latitude). The theory of "partial

likelihood" allows us to estimate the parameters of such chain models for the binary case. By restricting the degree of those polynomials or by requiring that some of their coefficients be the same, we can find simpler models. Simulation studies show that the "BIC" criterion (Bayesian information criterion) combined with the partial likelihood works well in that they recover the correct simulation model. Since we are only dealing with the categorical case all the density functions in this chapter are densities with the respect to the counting measure on the real line.

Specifying a categorical chain over time (with positive joint densities) using conditional probabilities of the present given the past is quite common in statistics and probability. However, we did not find a rigorous result for sufficient and necessary condition for a collection of function to correspond to the conditionals of a unique stochastic process. The proof is given in Theorem 3.5.6. This is an easy consequence of Lemma 3.3.2 that states that the "ascending" joint densities can uniquely determine such a stochastic process.

Another commonly used technique in statistics is transforming a discrete probability density from $(0, 1)$ to the real numbers using a transformations such as "log" for example in logistic regression. This is done to remove the restriction of these quantities and ease modeling of such probabilities. Theorem 3.4.1 provides a characterization of all such density functions given any bijective transformation between positive numbers and reals. Hence any positive discrete density function (mass function) correspond to a a unique function on reals and any arbitrary function on reals correspond to a positive function (after fixing the transformation and one element with positive probability). We do not know of a result in this generality elsewhere. Obviously now modeling such arbitrary function on reals which can only take finite values is easier.

In order to find a parametric form for an arbitrary function over the reals that only takes finite values, for the binary case, we use a corollary of a result stated by Besag [6] who used such functions in modeling Markov random fields. However, Besag did not provide a rigorous proof and the statement of the theorem is flawed as also pointed out by Cressie et al. in [14]. They also state a correct version of the theorem without offering a proof. We provide a rigorous statement and proof in Theorem 3.5.1. The corollary can only be obtained if the flaw in the statement is fixed. In order to extend to stochastic processes that can have more than two states at some times, we prove a new representation theorem in Theorem 3.5.6. Some novel simplified models with less parameters for such processes are given in Subsection 3.5.3 and many of them have been investigated in later chapters to model precipitation and

extreme temperature events occurrences.

3.2 Markov chains

Let $\{X_t\}_{t\in T}$ be a stochastic process on the index set T, where $T = \mathbb{Z}$, $T = \mathbb{N}$ (the integers or natural numbers respectively) or $T = \{0, 1, \cdots, n\}$. It is customary to call $\{X_t\}_{t\in T}$ a chain, since T is countable and has a natural ordering. $\{X_t\}_{t\in T}$ is called an rth–order Markov chain if:

$$P(X_t|X_{t-1}, \cdots) = P(X_t|X_{t-1}, \cdots, X_{t-r}),\ \forall t \text{ such that } t, t-r \in T.$$

We call the Markov chain homogenous if

$$P(X_t = x_t|X_{t-1} = x_{t-1}, \cdots, X_{t-r} = x_{t-r}) =$$
$$P(X_{t'} = x_t|X_{t'-1} = x_{t-1}, \cdots, X_{t'-r} = x_{t-r}),$$

$\forall t, t' \in T$ such that $t - r$ and $t' - r$ are also in T. Note that Markovness can be defined as a local property. We call $\{X_t\}_{t\in T}$ locally rth–order Markov at t if

$$P(X_t|X_{t-1}, \cdots) = P(X_t|X_{t-1}, \cdots, X_{t-r}).$$

Hence, we can have chains with a different Markov order at different times.

Let X_t be the binary random variable for precipitation on day t, with 1 denoting the occurrence of precipitation and 0 non-occurrence. In particular, consider the precipitation (PN) for Calgary site from 1895 to 2006. This process can be considered in two possible ways:

1. Let $X_1, X_2, \cdots, X_{366}$ denote the binary random variable of precipitation for days of a year. Suppose we repeatedly observe this chain year-by-year from 1895 to 2006 and take these observed chains to be independent and identically distributed from one year to the next. With this assumption, techniques developed in [4] can be applied in order to infer the Markov order of the chain. However, this approach presents three issues. Firstly independence of the successive chains seems questionable. In particular, the end of any one year will be autocorrelated with the beginning of the next. Secondly this model unrealistically assumes the 0-1 precipitation stochastic process is identically distributed over all years. Thirdly and more technically, leap years have 366 days while non–leap years have 365. We can resolve this last issue by formally assuming a missing data day in the non–leap years, by dropping

the last day in the leap year or by using other methods. However, none of these approaches seem completely satisfactory.

2. Alternatively, we could consider the observations of Calgary daily precipitation as coming from a single process that spans the entire time interval from 1895 to 2006. In this case, we will show below that we can still build models that bring in the seasonality effects within a year.

3.3 Consistency of the conditional probabilities

To represent a stochastic process, we only need to specify the joint probability distributions for all finite collections of states. The Kolmogorov extension theorem then guarantees the existence and uniqueness of an underlying stochastic process from which these distributions derive, provided they are consistent as described below. (See [9] for example.)

To state the version of that celebrated theorem we require, let T denote some interval (that can be thought of as "time"), and let $n \in \mathbb{N} = \{1, 2, \dots\}$. For each $k \in \mathbb{N}$ and finite sequence of times $t_1, \cdots, t_k$, let $\nu_{t_1 \cdots t_k}$ be a probability measure on $(\mathbb{R}^n)^k$. Suppose that these measures satisfy two consistency conditions:

1. **Permutation invariance.** For all permutations π (a bijective and one–to–one map from a set to itself) of $1, \cdots, k$ and measurable sets $F_i \subset \mathbb{R}^n$,

$$\nu_{t_{\pi(1)} \cdots t_{\pi(k)}} (F_1 \times \cdots \times F_k) = \nu_{t_1 \dots t_k} \left(F_{\pi^{-1}(1)} \times \cdots \times F_{\pi^{-1}(k)} \right).$$

2. **Marginalization consistency.** For all measurable sets $F_i \subseteq \mathbb{R}^n$, $m \in \mathbb{N}$:

$$\nu_{t_1 \dots t_k} (F_1 \times \cdots \times F_k) = \nu_{t_1 \dots t_k t_{k+1}, \dots, t_{k+m}} (F_1 \times \cdots \times F_k \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n).$$

Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a stochastic process

$$X : T \times \Omega \to \mathbb{R}^n,$$

such that:

$$\nu_{t_1 \dots t_k} (F_1 \times \cdots \times F_k) = \mathbb{P} (X_{t_1} \in F_1, \dots, X_{t_k} \in F_k),$$

for all $t_i \in T$, $k \in \mathbb{N}$ and measurable sets $F_i \subseteq \mathbb{R}^n$, i.e. X has the $\nu_{t_1 \dots t_k}$ as its finite–dimensional distributions. (See [37] for more details.)

Remark. Note that Condition 1 is equivalent to

$$\nu_{t_{\pi(1)}\dots t_{\pi(k)}}\left(F_{\pi(1)}\times\cdots\times F_{\pi(k)}\right)=\nu_{t_1\dots t_k}\left(F_1\times\cdots\times F_k\right).$$

This is seen by replacing $F_1\times\cdots\times F_k$ by $F_{\pi(1)}\times\cdots\times F_{\pi(k)}$ in the first equality.

Remark. We are only concerned about the case $n=1$. This is because we consider stochastic processes, a collection of random variables from the same sample space to $\mathbb{R}^1=\mathbb{R}$.

When working on (higher order) Markov chains over the index set $\mathbb{N}$, it is natural to consider the conditional distributions of the present, time t, given the past instead of the finite joint distributions, in other words

$$P_t(x_0,\cdots,x_t)=P(X_t=x_t|X_{t-1}=x_{t-1},\cdots,X_0=x_0),$$

for $\{X_t\}_{t\in\mathbb{N}\cup 0}$ plus the starting distribution

$$P_0(x_0)=P(X_0=x_0).$$

However that raises a fundamental question – does there exist a stochastic process whose conditional distributions match the specified ones and if so, is it unique? We answer this question affirmatively in this section for the case of discrete–time categorical processes, in particular higher order categorical Markov chains. We also restrict ourselves to chains for which all the joint probabilities are positive. Let $M_0,M_1,\cdots\subset\mathbb{R}$ be the state spaces for time $0,1,\cdots$, where each one of them is of finite cardinality. A probability measure on the finite space M_0 can be represented through its density function, a positive function $P_0:M_0\to\mathbb{R}$ satisfying the condition

$$\sum_{m\in M_0}P_0(m)=1.$$

The following theorem ensures the consistency of our probability model.

Theorem 3.3.1 *Suppose $M_0,M_1,\cdots\subset\mathbb{R}$, $|M_t|=c_t<\infty$, $t=0,1,\cdots$. Let $P_0:M_0\to\mathbb{R}$ be the density of a probability measure on M_0 and more generally for $n=1,\dots,$ $P_n(x_0,x_1,\cdots,x_{n-1},.)$ be a positive probability density on M_n, $\forall(x_0,\cdots,x_{n-1})\in M_0\times\cdots\times M_{n-1}$. Then there exists a unique stochastic process (up to distributional equivalence) on a probability space (Ω,Σ,P) such that*

$$P(X_n=x_n|X_{n-1}=x_{n-1},\cdots,X_0=x_0)=P_n(x_0,x_1,\cdots,x_{n-1},x_n).$$

To prove this theorem, we first consider a related problem whose solution is used in the proof. More precisely, we consider stochastic processes $\{X_n\}_{n\in\mathbb{N}\cup\{0\}}$, where the state space for X_n is M_n, $i = 0, 1, 2, \cdots$ and finite. Suppose $p_n : M_0 \times M_1 \times \cdots \times M_n \to \mathbb{R}$ is the joint probability distribution (density) of a random vector $\{X_0, \ldots, X_n\}$, i.e.

$$p_n(x_0, \cdots, x_n) = P(X_0 = x_0, \cdots, X_n = x_n).$$

We call a such sequence of functions, $\{p_n\}_{n\in\mathbb{N}}$, the "ascending joint distributions" of the stochastic process $\{X_n\}_{n\in\mathbb{N}\cup\{0\}}$. It is clear that given a family of functions $\{p_n\}_{n\in\mathbb{N}}$, other joint distributions such as

$$P(X_{t_1} = x_{t_1}, \cdots, X_{t_k} = x_{t_k}),$$

are obtainable by summing over appropriate components. Now consider the inverse problem. Given the $\{p_n\}_{n\in\mathbb{N}}$ and some type of consistency between them, is there a (unique) stochastic process that matches these joint distributions? The following lemma gives an affirmative answer.

Lemma 3.3.2 *Suppose $M_t \subset \mathbb{R}$, $t = 0, 1, \cdots$ are finite, $p_0 : M_0 \to \mathbb{R}$ represents a probability density function (i.e. $\sum_{x_0\in M_0} p(x_0) = 1$) and functions $p_n : M_1 \times \cdots \times M_n \to \mathbb{R}^+ \cup \{0\}$ satisfy the following (consistency) condition:*

$$\sum_{x_n\in M_n} p_n(x_0, \cdots, x_n) = p_{n-1}(x_0, \cdots, x_{n-1}).$$

Then there exist a unique stochastic process (up to distributional equivalence) $\{X_t\}_{t\in\mathbb{N}\cup\{0\}}$ such that

$$P(X_0 = x_0, \cdots, X_n = x_n) = p_n(x_0, \cdots, x_n)$$

Proof
Existence: By the Kolmogorov extension theorem quoted above, we only need to show there exists a consistent family of measures (density functions)

$$\{q_{t_1,\cdots,t_k} | k \in \mathbb{N}, (t_1, \cdots, t_k) \in \mathbb{N}^k\},$$

such that $q_{1,\cdots,t} = p_t$. We define such a family of functions, prove they are measures and consistent.

For any sequence, $t_1, \cdots, t_k$, let $t = \max\{t_1, \cdots, t_k\}$ and define

$$q_{t_1,\cdots,t_k}(x_{t_1}, \cdots, x_{t_k}) = \sum_{x_u\in M_u, u\in\{1,\cdots,t\}-\{t_1,\cdots,t_k\}} p_t(x_1, \cdots, x_t).$$

We need to prove three things:

a) Each $q_{t_1,\cdots,t_k}$ is a density function. It suffices to show that q_t is a measure because the $q_{t_1,\cdots,t_k}$ are sums of such measures and so are measures themselves. But p_t is nonnegative by assumption. It only remains to show that p_t sums up to one. For $t = 1$ it is in the assumptions of the theorem. For $t > 1$, it can be done by induction because of the following identity

$$\sum_{x_i \in M_i, i=0,1,\cdots,t} p_t(x_0, \cdots, x_t) = \sum_{x_i \in M_i, i=0,1,\cdots,t-1} p_{t-1}(x_0, \cdots, x_{t-1})$$

where the right hand side is obtained by the assumption $\sum_{M_n} p_n = p_{n-1}$.

b) In order to satisfy the first condition of Kolmogorov extension theorem, we need to show

$$q_{t_1,\cdots,t_k}(x_{t_1}, \cdots, x_{t_k}) = q_{t_{\pi(1)},\cdots,t_{\pi(k)}}(x_{t_{\pi(1)}}, \cdots, x_{t_{\pi(k)}}),$$

for π a permutation of $\{1, 2, \cdots, k\}$. But this is obvious since

$$\max\{t_1, \cdots, t_k\} = \max\{t_{\pi(1)}, \cdots, t_{\pi(k)}\}.$$

c) In order to satisfy the second condition of Kolmogorov extension theorem, we need to show

$$\sum_{x_{t_i} \in M_{t_i}} q_{t_1,\cdots,t_i,\cdots,t_k}(x_{t_1}, \cdots, x_{t_i}, \cdots, x_{t_k}) = q_{t_1,\cdots,\hat{t_i},\cdots,t_k}(x_{t_1}, \cdots, \hat{x_{t_i}}, \cdots, x_{t_k}),$$

where the notation $\hat{}$ above a component means that component is omitted.

To prove this, we consider two cases:

Case I: $t = \max\{t_1, \cdots, t_k\} = \max\{t_1, \cdots, \hat{t_i}, \cdots, t_k\}$:

$$\begin{aligned}
&\sum_{x_{t_i} \in M_{t_i}} q_{t_1,\cdots,t_i,\cdots,t_k}(x_{t_1}, \cdots, x_{t_i}, \cdots, x_{t_k}) = \\
&\sum_{x_{t_i} \in M_{t_i}} \sum_{x_u \in M_u, u \in \{1,\cdots,t\} - \{t_1,\cdots,t_i,\cdots,t_k\}} p_t(x_0, \cdots, x_t) = \\
&\sum_{x_u \in M_u, u \in \{1,\cdots,t\} - \{t_1,\cdots,\hat{t_i},\cdots,t_k\}} p_t(x_0, \cdots, x_t) = \\
&p_{t_1,\cdots,\hat{t_i},\cdots,t_k}(x_{t_1}, \cdots, \hat{x_{t_i}}, \cdots, x_{t_k})
\end{aligned}$$

Case II: $\max\{t_1, \cdots, \hat{t_i}, \cdots, t_k\} = t' < t = t_i$:

$$\sum_{x_{t_i} \in M_{t_i}} q_{t_1,\cdots,t_i,\cdots,t_k}(x_{t_1}, \cdots, x_{t_i}, \cdots, x_{t_k}) =$$

$$\sum_{x_{t_i} \in M_{t_i}} \sum_{x_u \in M_u, u \in \{1,\cdots,t\} - \{t_1,\cdots,t_i,\cdots,t_k\}} p_t(x_0, \cdots, x_t) =$$

$$\sum_{x_u \in M_u, u \in \{1,\cdots,t\} - \{t_1,\cdots,\hat{t_i},\cdots,t_k\}} p_t(x_0, \cdots, x_t) =$$

$$\sum_{x_u \in M_u, u \in \{1,\cdots,t'\} - \{t_1,\cdots,\hat{t_i},\cdots,t_k\}} \sum_{x_v \in M_v, v \in \{t'+1,\cdots,t\}} f_t(x_0, \cdots, x_t) =$$

$$\sum_{x_u \in M_u, u \in \{1,\cdots,t'\} - \{t_1,\cdots,\hat{t_i},\cdots,t_k\}} p_{t'}(x_0, \cdots, x_{t'}) =$$

$$q_{t_1,\cdots,\hat{t_i},\cdots,t_k}(x_{t_1}, \cdots, \hat{x_{t_i}}, \cdots, x_{t_k}).$$

Uniqueness: Suppose $\{Y_t\}_{t \in \mathbb{N} \cup \{0\}}$ is another stochastic process satisfying the conditions of the theorem with the $p'_{t_1,\cdots,t_k}$ as the joint measures.

$$p'_{1,\cdots,t} = p_t = p_{1,\cdots,t},$$

by the assumption. Taking the appropriate sums on the two sides, we get $p'_{t_1,\cdots,t_k} = p_{t_1,\cdots,t_k}$. Now the uniqueness is a straight consequence of the Kolmogorov Extension Theorem. ■

Remark. Note that we did not impose the positivity of the functions for this case.

Now we are ready to prove Theorem 3.3.1.

Proof

Existence: In Lemma 3.3.2, let

$p_0 = P_0$,

$p_1 : M_0 \times M_1 \to \mathbb{R},\ p_1(x_0, x_1) = p_0(x_0) P_1(x_0, x_1)$,

$\vdots$

$p_n : M_1 \times M_2 \times \cdots \times M_n \to \mathbb{R},\ p_n(x_0, \cdots, x_n) = p_{n-1}(x_0, \cdots, x_{n-1}) P_n(x_0, \cdots, x_n)$.

To see that the $\{p_i\}$ satisfy the conditions of Lemma 3.3.2, note that

$$\sum_{x_n \in M_n} p_n(x_0, \cdots, x_n) =$$
$$\sum_{x_n \in M_n} p_{n-1}(x_0, \cdots, x_{n-1}) P_n(x_0, \cdots, x_n) =$$
$$p_{n-1}(x_0, \cdots, x_{n-1}) \sum_{x_n \in M_n} P_n(x_0, \cdots, x_n) =$$
$$p_{n-1}(x_0, \cdots, x_{n-1}).$$

Lemma 3.3.2 shows the existence of a stochastic process with joint distributions matching the p_i. Furthermore, the positivity of the $\{P_i\}$ implies that of the $\{p_i\}$. Thus all the conditionals exist for such a process and they match the P_i by the definition of the conditional probabilities.
Uniqueness. Any stochastic process satisfying the above conditions, has a joint distribution that matches those of the $\{p_i\}$ and hence by the above theorem they are unique. ■

3.4 Characterizing density functions and rth–order Markov chains

The previous section saw discrete–time categorical processes represented in terms of conditional probability density functions. However such densities on finite domains satisfy certain restrictions that can make modeling them difficult. That leads to the idea of linking them to unrestricted functions on $\mathbb{R}$ in much the same spirit as a single probability can profitably be *logit* transformed in logistic regression.

To begin, let X be a random variable with probability density p defined on a finite set $M = \{m_1, \cdots, m_n\}$. The section finds the class of all possible such ps with $p(m_i) > 0$, $i = 1, \cdots, n$ and $g : \mathbb{R} \to \mathbb{R}^+$, a fixed bijection. For example $g(x) = \exp(x)$. The following theorem characterizes the relationship between p and g. While particular examples of the following theorem are used commonly in statistical modeling we are not aware of a reference which contains this result or the proof in this generality.

Theorem 3.4.1 *Let $g : \mathbb{R} \to \mathbb{R}^+$ a bijection. For every choice of probability density p on $M = \{m_1, \cdots, m_n\}$, $n \geq 2$, there exists a unique function $f : M - \{m_1\} \to \mathbb{R}$, such that*

$$p(m_1) = \frac{1}{1 + \sum_{y \in M - \{m_1\}} h(y)}, \tag{3.1}$$

$$p(x) = \frac{h(x)}{1 + \sum_{y \in M - \{m_1\}} h(y)},\ x \neq m_1, \tag{3.2}$$

where $h = g \circ f$. Moreover, $h(x) = p(x)/p(m_1)$. Inversely, for an arbitrary function $f : M - \{m_1\} \to \mathbb{R}$, the p defined above is a density function.

Proof
Existence: Suppose $p : M \to (0, 1)$ is given. Let $h(x) = \frac{p(x)}{p(m_1)}$, $x \neq m_1$ and $f : M - \{m_1\} \to \mathbb{R}$, $f(x) = g^{-1} \circ h(x)$. Obviously $h = g \circ f$. Moreover

$$\frac{1}{1 + \sum_{y \in M - \{m_1\}} h(y)} = \frac{1}{1 + \sum_{y \in M - \{m_1\}} p(y)/p(m_1)} =$$

$$\frac{1}{1 + (1 - p(m_1))/p(m_1)} = p(m_1)$$

and

$$\frac{h(x)}{1 + \sum_{y \in M - \{m_1\}} h(y)} = \frac{p(x)/p(m_1)}{1 + (1 - p(m_1))/p(m_1)} = p(x),$$

thereby establishing the validity of equations (3.1) and (3.2).
Uniqueness: Suppose for f_1, f_2, we get the same p. Let $h_1 = g \circ f_1$, $h_2 = g \circ f_2$, by dividing 3.2 by 3.1 for h_1 and h_2, we get $h_1(x) = p(x)/p(m_1) = h_2(x)$ hence $g \circ f_1 = g \circ f_2$. Since g is a bijection $f_1 = f_2$. ■

Corollary 3.4.2 *Fixing a bijection g and $m_1 \in M$, every density function corresponds to an arbitrary vector of length $n - 1$ over $\mathbb{R}$.*

Example Consider the binomial distribution with a trials and probability of success π and the transformation $g(x) = \exp x$. Then $M = \{0, 1, \cdots, a\}$. Let $m_1 = 0$ then for $x \neq 0$

$$f(x) = g^{-1}(h(x)) = \log p(x)/p(0) = \log \binom{n}{x} p^x (1 - p)^{n-x}/(1 - p)^n =$$

$$\log \binom{n}{x} + x \log\{p/(1 - p)\}.$$

Theorem 3.4.3 *Fix a bijection $g : \mathbb{R} \to \mathbb{R}^+$, $m_1^n \in M_n$. Let $M_n, n = 0, 1, \cdots$ be finite subsets of $\mathbb{R}$ with cardinality greater than or equal to 2 and $M_n' = M_n - \{m_1^n\}$, $\forall n$. Then every categorical stochastic process with positive joint distribution on the M_n having initial density $P_0 : M_0 \to \mathbb{R}$ and conditional probabilities P_n at stage n given the past, can be uniquely represented by means of unique functions:*

$$\begin{aligned} g_0 &: M_0' \to \mathbb{R} \\ &\vdots \\ g_n &: M_0 \times \cdots \times M_{n-1} \times M_n' \to \mathbb{R} \\ &\vdots \end{aligned}$$

for $n = 1, \ldots,$ where

$$P_0(m_1^0) = \frac{1}{1 + \sum_{y \in M_0 - \{m_1^0\}} h_0(y)}, \tag{3.3}$$

$$P_0(x) = \frac{h_0(x)}{1 + \sum_{y \in M_0 - \{m_1^0\}} h_0(y)}, \quad x \neq m_1^0 \in M_0, \tag{3.4}$$

and $h_0 = g \circ g_0$. Moreover $h_0(x) = \frac{P(X_0 = x)}{P(X_0 = m_0^1)}$.
The conditional probabilities P_n are given by

$$P_n(x_0, \cdots, x_{n-1}, m_1^n) = \frac{1}{1 + \sum_{y \in M_n - \{m_1^n\}} h_n(y)}, \tag{3.5}$$

$$P_n(x_0, \cdots, x_{n-1}, x) = \frac{h(x)}{1 + \sum_{y \in M_n - \{m_1^n\}} h_n(y)}, \quad x \neq m_1^n \in M_n, \tag{3.6}$$

where, $h_n = g \circ g_n$. Moreover $h_n(x_0, \cdots, x) = \frac{P(X_n = x | X_{n-1} = x_{n-1}, \cdots, X_0 = x_0)}{P(X_n = m_n^1 | X_{n-1} = x_{n-1}, \cdots, X_0 = x_0)}$. Conversely, any collection of arbitrary functions $g_0, g_1, \cdots$ gives rise to a unique stochastic process by the above relations.

Proof

The result is immediate by Theorems 3.3.1 and 3.4.1. ■

Remark. We can view the arbitrary functions $g_0, \cdots, g_n$ on $M_0', M_0 \times M_1', \cdots, M_0 \times \cdots \times M_{n-1} \times M_n'$ as arbitrary functions g_0 on M_0', $g_1(., x_1),\ x_1 \neq m_1^1$ on M_0 and $g_n(., x_n),\ x_n \neq m_1^n$ on $M_0 \times \cdots \times M_{n-1}$. As a check we can compute the number of free parameters of such a stochastic process on $M_0, \cdots, M_n$. We can specify such a process by $c_0 c_1 \cdots c_n - 1$ parameters by specifying the joint distribution on $M_0 \times M_1 \times \cdots \times M_n$. If we specify the stochastic process using the above theorems and the g_i functions, we need $(m_0 - 1) + m_0(m_1 - 1) + m_0 m_1 (m_2 - 1) + \cdots + m_0 m_1 \cdots m_{n-1}(m_n - 1)$ which is the same number after expanding the terms and canceling out.

Remark. In the case of rth–order Markov chains, $g_n(x_0, \cdots, x_n)$ only depends on the last $r + 1$ components for $n > r$.

Remark. In the case of homogenous rth–order Markov chains, $M_i = M_0,\ \forall i$. Fix $m_0 \in M_0$ and suppose $|M_0| = c_0$. We only need to specify g_0 to g_r, which are completely arbitrary functions. We only need to specify g_0 on M_0', g_1 on $M_0 \times M_1'$ to g_r on $M_0 \times \cdots \times M_{r+1}'$. This also shows every homogenous Markov chain of order at most r is characterized by $(c_0 - 1) \sum_{i=0}^{r} c_0^r = c_0^{r+1} - 1$ elements in $\mathbb{R}$. We could have also counted all such Markov chains by noting they are uniquely represented by the joint probability density p_{r+1} on M_0^{r+1} which has $c_0^{r+1} - 1$ free parameters (since it has to sum up to 1).

To describe processes using Markov chains, we need to find appropriate parametric forms. We investigate the generality of these forms in the following section and use the concept of partial likelihood to estimate them. We find appropriate parametric representations of g_n which are functions of $n + 1$ finite variables. In the next section we study the properties of such functions. We call a variable "finite" if it only takes values in a finite subset of $\mathbb{R}$.

3.5 Functions of r variables on a finite domain

This section studies the properties of functions of r variables with finite domain. First, we present a result of Besag [6] who studied such functions in the context of Markov random fields. However the statement of the result in his paper is inaccurate and moreover it gives no rigorous proof of his result. We present a rigorous statement, proof of the result and generalization of Besag's theorem.

3.5.1 First representation theorem

This subsection presents a corrected version of a theorem stated by Besag in [6] and a constructive proof. Then we generalize this theorem and apply it to stationary binary Markov chains to get a parametric representation.

Theorem 3.5.1 *Suppose,* $f : \prod_{i=1,\cdots,r} M_i \to \mathbb{R}$, M_i *being finite with* $|M_i| = c_i$ *and* $0 \in M_i$, $\forall i$, $1 \le i \le r$. *Let* $M_i' = M_i - \{0\}$. *Then there exist a unique family of functions*

$$\{G_{i_1,\cdots,i_k} : M_{i_1}' \times M_{i_2}' \times \cdots \times M_{i_k}' \to \mathbb{R},\ 1 \le k \le r,\ 1 \le i_1 < i_2 < \cdots < i_k \le r\},$$

such that

$$f(x_1,\cdots,x_r) = f(0,\cdots,0) + \sum_{i=1}^{r} x_i G_i(x_i) + \cdots + \sum_{1\le i_1<i_2<\cdots<i_k\le r} (x_{i_1}\cdots x_{i_k}) G_{i_1,\cdots,i_k}(x_{i_1},\cdots,x_{i_k}) + \cdots + (x_1 x_2 \cdots x_r) G_{12\cdots r}(x_1,\cdots,x_r).$$

.

Remark. In [6], Besag claims that $\{G_{i_1,\cdots,i_k} : M_{i_1} \times M_{i_2} \times \cdots \times M_{i_k} \to \mathbb{R}\}$ (without removing one element from each set) are unique.

Proof Denote by I_A the indicator function of a set A and

$$N_k = \{(x_1,\cdots,x_r) : \sum_{i=1}^{r} I_{\{0\}}(x_i) \le k\}.$$

Existence: The proof is by induction. For $i = 1,\cdots,r$, define

$$G_i : M_i' \to \mathbb{R},$$
$$G_i(x_i) = \frac{f(0,\cdots,0,x_i,0,\cdots,0) - f(0,\cdots,0)}{x_i},$$

where x_i is the i^{th} coordinate. Then let $f_1(x_1,\cdots,x_r) = f(0,\cdots,0) + \sum_{i=1}^{r} x_i G_i(x_i)$. Note that $f_1 = f$ on N_1.

Next define $G_{i_1,i_2} : M_{i_1}' \times M_{i_2}' \to \mathbb{R}$ by

$$G_{i_1,i_2}(x_{i_1},x_{i_2}) = \frac{f(0,\cdots,0,x_{i_1},0,\cdots,0,x_{i_2},0,\cdots,0) - f_1(0,\cdots,0,x_{i_1},0,\cdots,0,x_{i_2},0,\cdots,0)}{x_{i_1}x_{i_2}},$$

where, x_{i_1}, x_{i_2} are the i_1^{th} and i_2^{th} coordinates, respectively. Using the $\{G_{i_1,i_2}\}$, we can define f_2 on N_2 by

$$f_2(x_1,\cdots,x_r) = f(0,\cdots,0) + \sum_{i=1}^{r} x_i G_i(x_i) + \sum_{1\le i_1<i_2\le r} x_{i_1}x_{i_2}G_{i_1,i_2}(x_{i_1},x_{i_2}).$$

Or equivalently,

$$f_2(x_1,\cdots,x_r) = f_1(x_1\cdots,x_r) + \sum_{1\le i_1<i_2\le r} x_{i_1}x_{i_2}G_{i_1,i_2}(x_{i_1},x_{i_2}).$$

It is easy to see that $f_2 = f$ on N_2.

In general, suppose we have defined $G_{i_1,\cdots,i_{k-1}}$ and f_{k-1}, let

$$G_{i_1,\cdots,i_k}(x_{i_1},\cdots,x_{i_k}) = \frac{f(0,\cdots,0,x_{i_1},0,\cdots,0,x_{i_k},0,\cdots,0) - f_{k-1}(0,\cdots,0,x_{i_1},0,\cdots,0,x_{i_k},0,\cdots,0)}{x_{i_1}\cdots x_{i_k}},$$

for $(x_{i_1},\cdots,x_{i_k}) \in M'_{i_1}\times\cdots\times M'_{i_k}$.
Also let

$$f_k(x_1,\cdots,x_r) = f_{k-1}(x_1,\cdots,x_r) + \sum_{1\le i_1<i_2<\cdots<i_k\le r} x_{i_1}\cdots x_{i_k}G_{i_1,\cdots,i_k}(x_{i_1},\cdots,x_{i_k})$$

We claim $f = f_k$ on N_k.

To see that, fix $x = (x_1,\cdots,x_r)$. If x has less than k nonzero elements, the second term in the above expansion will be zero and

$$f_k(x_1,\cdots,x_r) = f_{k-1}(x_1,\cdots,x_r) = f(x_1,\cdots,x_r),$$

by the induction hypothesis and we are done.

However if x has exactly k nonzero elements

$$x = (x_1,\cdots,x_r) = (0,\cdots,0,x_{j_1},0,\cdots,0,x_{j_k},0\cdots).$$

Then

$$\sum_{1\le i_1<i_2<\cdots<i_k\le r} x_{i_1}\cdots x_{i_k}G_{i_1,\cdots,i_k}(x_{i_1},\cdots,x_{i_k}) = x_{j_1}\cdots x_{j_k}G_{j_1,\cdots,j_k}(x_{j_1},\cdots,x_{j_k}).$$

Hence

$$\begin{aligned} f_k(x_1,\cdots,x_r) &= f_{k-1}(x_1,\cdots,x_r) + (x_{j_1},\cdots,x_{j_k})G_{j_1,\cdots,j_k}(x_{j_1},\cdots,x_{j_k}) \\ &= f_{k-1}(x_1,\cdots,x_r)+ \\ x_{j_1}\cdots x_{j_k}&\frac{f(\cdots,0,x_{j_1},0,\cdots,0,x_{j_k},0,\cdots)-f_{k-1}(\cdots,0,x_{j_1},0,\cdots,0,x_{j_k},0,\cdots)}{x_{j_1}\cdots x_{j_k}} \\ &= f(x_1,\cdots,x_r). \end{aligned}$$

By induction, $f = f_r$ on $N_r = \prod_{i=1,\cdots,r} M_i$. Hence, the family of functions satisfies the conditions.
Uniqueness: To prove uniqueness, suppose

$$\{G_{i_1,\cdots,i_k} : M'_{i_1}\times M'_{i_2}\times\cdots\times M'_{i_k}\to\mathbb{R},\ 1\le k\le r,\ 1\le i_1<i_2<\cdots<i_k\le r\},$$

and

$$\{H_{i_1,\cdots,i_k} : M'_{i_1}\times M'_{i_2}\times\cdots\times M'_{i_k}\to\mathbb{R},\ 1\le k\le r,\ 1\le i_1<i_2<\cdots<i_k\le r\},$$

are two families of functions satisfying the equation. Also assume f_k^G and f_k^H are the summation functions as defined above corresponding to the two families. We need to show $G_{i_1,\cdots,i_k} = H_{i_1,\cdots,i_k}$ on $M'_{i_1}\times\cdots\times M'_{i_k}$. We use induction on k. It is easy to verify the result for the case $k = 1$. Now suppose $x = (x_{i_1},\cdots,x_{i_k}) \in M'_{i_1}\times M'_{i_2}\times\cdots\times M'_{i_k}$. Then by definition

$$G_{i_1,\cdots,i_k}(x_{i_1},\cdots,x_{i_k}) = \frac{f(0,\cdots,0,x_{i_1},0,\cdots,0,x_{i_k},0,\cdots,0)-f_{k-1}^G(0,\cdots,0,x_{i_1},0,\cdots,0,x_{i_k},0,\cdots,0)}{x_{i_1}\cdots x_{i_k}},$$

and

$$H_{i_1,\cdots,i_k}(x_{i_1},\cdots,x_{i_k}) = \frac{f(0,\cdots,0,x_{i_1},0,\cdots,0,x_{i_k},0,\cdots,0)-f_{k-1}^H(0,\cdots,0,x_{i_1},0,\cdots,0,x_{i_k},0,\cdots,0)}{x_{i_1}\cdots x_{i_k}}.$$

But by induction hypothesis $f^G_{k-1} = f^H_{k-1}$. Hence we are done. ■

We can think of this representation of f as an expansion around $(0, \cdots, 0)$. However, $(0, \cdots, 0)$ has no intrinsic role and we can generalize the above theorem as follows.

Theorem 3.5.2 *Suppose, $f : M = \prod_{i=1,\cdots,r} M_i \to \mathbb{R}$, M_i being finite and $|M_i| = c_i$. For any fixed $(\mu_1, \cdots, \mu_r) \in M$, let $M'_i = M_i - \{\mu_i\}$. Then there exist unique functions*

$$\{H_{i_1,\cdots,i_k} : M'_{i_1} \times M'_{i_2} \times \cdots \times M'_{i_k} \to \mathbb{R},\ 1 \le k \le r,\ 1 \le i_1 < i_2 < \cdots < i_k \le r\},$$

such that

$$f(x_1, \cdots, x_r) = f(\mu_1, \cdots, \mu_r) + \sum_{i=1}^{r} (x_i - \mu_i) H_i(x_i) + \cdots +$$
$$\sum_{1 \le i_1 < i_2 < \cdots < i_k \le r} (x_{i_1} - \mu_{i_1}) \cdots (x_{i_k} - \mu_{i_k}) H_{i_1,\cdots,i_k}(x_{i_1}, \cdots, x_{i_k}) +$$
$$\cdots + (x_1 - \mu_1)(x_2 - \mu_2) \cdots (x_r - \mu_r) H_{12\cdots r}(x_1, \cdots, x_r).$$

Proof Let $N_i = M_i - \mu_i$ (meaning that we subtract μ_i from all elements of M_i) so that N_i and M_i have the same cardinality. Also let $N = \prod_{i=1,\cdots,r} N_i$ and $N'_i = N_i - \{0\}$. Then define a bijective mapping

$$\phi_i : N_i \to M_i,$$

$$\phi_i(x_i) = x_i + \mu_i.$$

This will induce a bijective mapping Φ between N and M that takes $(0, \cdots, 0)$ to $(\mu_1, \cdots, \mu_r)$. Now consider $f \circ \Phi : \prod_{i=1,\cdots,r} N_i \to \mathbb{R}$. By the previous theorem, unique functions

$$\{G_{i_1,\cdots,i_k} : N'_{i_1} \times N'_{i_2} \times \cdots \times N'_{i_k} \to \mathbb{R},\ 1 \le k \le r,\ 1 \le i_1 < i_2 < \cdots < i_k \le r\}$$

exist such that

$$f \circ \Phi(x_1, \cdots, x_r) = f \circ \Phi(0, \cdots, 0) + \sum_{i=1}^{r} x_i G_i(x_i) + \cdots +$$
$$\sum_{1 \le i_1 < i_2 < \cdots < i_k \le r} x_{i_1} \cdots x_{i_k} G_{i_1,\cdots,i_k}(x_{i_1}, \cdots, x_{i_k}) + \cdots + x_1 x_2 \cdots x_r G_{12\cdots r}(x_1, \cdots, x_r).$$

Hence,

$$f(\phi_1(x_1),\cdots,\phi_r(x_r)) = f(\phi_1(0),\cdots,\phi_r(0)) + \sum_{i=1}^{r} x_i G_i(x_i) + \cdots + \sum_{1\le i_1<i_2<\cdots<i_k\le r} x_{i_1}\cdots x_{i_k} G_{i_1,\cdots,i_k}(x_{i_1},\cdots,x_{i_k}) + \cdots + x_1 x_2 \cdots x_r G_{12\cdots r}(x_1,\cdots,x_r).$$

We conclude,

$$f(x_1+\mu_1,\cdots,x_r+\mu_r) = f(\mu_1,\cdots,\mu_r) + \sum_{i=1}^{r} x_i G_i(x_i) + \cdots + \sum_{1\le i_1<i_2<\cdots<i_k\le r} x_{i_1}\cdots x_{i_k} G_{i_1,\cdots,i_k}(x_{i_1},\cdots,x_{i_k}) + \cdots + x_1 x_2 \cdots x_r G_{12\cdots r}(x_1,\cdots,x_r).$$

This gives

$$f(x_1,\cdots,x_r) = f(\mu_1,\cdots,\mu_r) + \sum_{i=1}^{r}(x_i-\mu_1)G_i(x_i-\mu_i) + \cdots + \sum_{1\le i_1<i_2<\cdots<i_k\le r}(x_{i_1}-\mu_{i_1})\cdots(x_{i_k}-\mu_{i_k})G_{i_1,\cdots,i_k}(x_{i_1}-\mu_{i_1},\cdots,x_{i_k}-\mu_{i_k}) + \cdots + (x_1-\mu_1)(x_2-\mu_2)\cdots(x_r-\mu_r)G_{12\cdots r}(x_1-\mu_1,\cdots,x_r-\mu_r).$$

To prove the existence, let

$$H_{i_1,\cdots,i_k}(x_{i_1},\cdots,x_{i_k}) = G_{i_1,\cdots,i_k}(x_{i_1}-\mu_{i_1},\cdots,x_{i_k}-\mu_{i_k}).$$

The uniqueness can be obtained as in the previous theorem. ■
We call this expression the Besag expansion around $(\mu_1,\cdots,\mu_r)$.

Corollary 3.5.3 *In the case of binary $\{0,1\}$ variables, the G functions are simply real numbers, since $M'_{i_1}\times\cdots\times M'_{i_k}$ has exactly one element: $(1,\cdots,1)$. Hence, we have found a linear representation of f in terms of the $x_{i_1}\cdots x_{i_k}$.*

Corollary 3.5.4 *Suppose that $\{X_t\}$ is an rth–order Markov chain, X_t taking values in $M_t=\{0,1\}$ and the conditional probability*

$$P(X_t=1|X_{t-1},\cdots,X_0),$$

is well-defined and in (0,1). Let $g : \mathbb{R} \to \mathbb{R}^+$ be a given bijective transformation. Then

$$g_t(x_{t-1}, \cdots, x_0) = g^{-1}\{\frac{P(X_t = 1|X_{t-1} = x_{t-1}, \cdots, X_0 = x_0)}{P(X_t = 0|X_{t-1} = x_{t-1}, \cdots, X_0 = x_0)}\},$$

is a function of t variables, $(x_{t-1}, \cdots, x_0)$, for $t < r$ and is a function of r variables, $(x_{t-1}, \cdots, x_{t-r})$, for $t > r$. Hence there exist unique parameters α_0^t, $\{\alpha^t_{i_1,\cdots,i_t}\}_{1\le i_1,\cdots,i_t\le t}$ for $t < r$ and $\alpha_0^t, \{\alpha^t_{i_1,\cdots,i_r}\}_{1\le i_1,\cdots,i_r\le r}$ for $t \ge r$ such that
for $t < r$:

$$g^{-1}\{\frac{P(X_t = 1|X_{t-1}, \cdots, X_0)}{P(X_t = 0|X_{t-1}, \cdots, X_0)}\} =$$
$$\alpha_0^t + \sum_{i=1}^{t} X_{t-i}\alpha_i^t + \cdots +$$
$$\sum_{1\le i_1<i_2<\cdots<i_k\le t} \alpha^t_{i_1,\cdots,i_k} X_{t-i_1}\cdots X_{t-i_k} + \cdots +$$
$$\alpha^t_{12\cdots t} X_{t-1}X_{t-2}\cdots X_0.$$

and for $t \ge r$:

$$g^{-1}\{\frac{P(X_t = 1|X_{t-1}, \cdots, X_0)}{P(X_t = 0|X_{t-1}, \cdots, X_0)}\} =$$
$$\alpha_0^t + \sum_{i=1}^{r} X_{t-i}\alpha_i^t + \cdots +$$
$$\sum_{1\le i_1<i_2<\cdots<i_k\le r} \alpha^t_{i_1,\cdots,i_k} X_{t-i_1}\cdots X_{t-i_k} + \cdots +$$
$$\alpha^t_{12\cdots r} X_{t-1}X_{t-2}\cdots X_{t-r}.$$

Moreover, given any collection of parameters, α_0^t, $\{\alpha^t_{i_1,\cdots,i_t}\}_{1\le i_1,\cdots,i_t\le t}$ for $t < r$ and $\alpha_0^t, \{\alpha^t_{i_1,\cdots,i_r}\}_{1\le i_1,\cdots,i_r\le r}$ for $t \ge r$ a unique stochastic process (upto distribution) is specified using the above relations.
In the case of homogenous Markov chains the α_0^t, $\alpha^t_{i_1,\cdots,i_k}$ do not depend on t for $t > r$.

The above corollary shows that the conditional probability of a Markov chain after an appropriate transformation can be uniquely represented as a linear combination of monomial products of previous states.

One might conjecture that the same result holds for all categorical–valued Markov chains (with a finite number of states) using the above theorem. This is not true in general since the $\{G_{i_1,\cdots,i_k}\}$ are functions. In the next section, we prove another representation theorem which paves the way for the categorical case. As it turns out, we need more terms in order to write down the transformed conditional probability as a linear combination of past processes.

3.5.2 Second representation theorem

In this section, we prove a new representation theorem for functions of r finite variables. We start with the trivial finite–valued one–variable function and then extend the result to r–variable functions. The proof for the general case is non–trivial and is done again by induction.

Lemma 3.5.5 *Suppose $f : M \to \mathbb{R}$, $M \subset \mathbb{R}$ being finite of cardinality c. Let $d = c - 1$. Then f has a unique representation of the form*

$$f(x) = \sum_{0 \le i \le d} \alpha_i x^i, \ \forall x \in M.$$

Remark. The lemma states that, if we consider the vector space $V = \{f : M \to \mathbb{R}\}$, then the monomial functions $\{p_i\}_{0 \le i \le d}$, where $p_i : M \to \mathbb{R}$, $p_i(x) = x^i$ form a basis for V.

Proof First note that the dimension of V is c. To show this, suppose $M = \{m_1, \cdots, m_c\}$ and consider the following isomorphism of vector spaces,

$$I : V \to \mathbb{R}^c$$

$$f \mapsto (f(m_1), \cdots, f(m_c)).$$

It only remains to show that $\{p_i\}_{0 \le i \le d}$ is an independent set. To prove this suppose,

$$\sum_{0 \le i \le d} \alpha_i x^i = 0, \quad \forall x \in M.$$

That would mean that the d–th degree polynomial $p(x) = \sum_{0 \le i \le d} \alpha_i x^i$ has at least $c = d + 1$ disjoint roots which is greater than its degree. This contradicts the fundamental theorem of algebra. ■

Theorem 3.5.6 *(Categorical Expansion Theorem) Suppose M_i is a finite subset of $\mathbb{R}$ with $|M_i| = c_i,\ i = 1, 2, \cdots, r$. Let $d_i = c_i - 1, M = \prod_{i=1,\cdots,r} M_i$ and consider the vector space of functions over $\mathbb{R}$, $V = \{f : M \to \mathbb{R}\}$ with the function addition as the addition operation of the vector space and the scalar product of a real number to the function as the scalar product of the vector space. Then this vector space is of dimension $C = \prod_{i=1,\cdots,r} c_i$ and $\{x_1^{i_1} \cdots x_r^{i_r}\}_{0 \le i_1 \le d_1, \cdots, 0 \le i_r \le d_r}$ forms a basis for it.*

Proof To show that the dimension of the vector space is C, suppose $M = \{m_1, \cdots, m_c\}$ and consider following the isomorphism of vector spaces:

$$I : V \to \mathbb{R}^C,$$

$$f \mapsto (f(m_1), \cdots, f(m_C)).$$

To show that $\{x_1^{i_1} \cdots x_r^{i_r}\}_{0 \le i_1 \le d_i, \cdots, 0 \le i_r \le d_r}$ forms a basis, we only need to show that it is an independent collection since there are exactly C elements in it. We proceed by induction on r. The case $r = 1$ was shown in the above lemma. Suppose we have shown the result for $r - 1$ and we want to show it for r. Assume a linear combination of the basis is equal to zero. We can arrange the terms based on powers of x_r:

$$p_0(x_1, \cdots, x_{r-1}) + x_r p_1(x_1, \cdots, x_{r-1}) + \cdots + x_r^{d_r} p_d(x_1, \cdots, x_{r-1}) = 0, \quad (3.7)$$

$$\forall (x_1, \cdots, x_r) \in M_1 \times \cdots \times M_r.$$

Fix the values of $x_1', \cdots, x_{r-1}' \in M_1 \times \cdots \times M_{r-1}$. Then Equation (3.7) is zero for c_r values of x_r. Hence by Lemma 3.5.5, all the coefficients:

$$p_0(x_1', \cdots, x_{r-1}'), p_1(x_1', \cdots, x_{r-1}'), \cdots, p_d(x_1', \cdots, x_{r-1}'),$$

are zero and we conclude:

$$p_0(x_1, \cdots, x_{r-1}) = 0, p_1(x_1, \cdots, x_{r-1}) = 0, \cdots, p_d(x_1, \cdots, x_{r-1}) = 0,$$

$$\forall (x_1, \cdots, x_{r-1}) \in M_1 \times \cdots \times M_{r-1}.$$

Again by the induction assumption all the coefficients in these polynomials are zero. Hence, all the coefficients in the original linear combination in Equation (3.7) are zero. ■

Corollary 3.5.7 *Suppose X_t is a categorical stochastic process, where X_t takes values in M_t, $|M_t| = c_t = d_t+1 < \infty$. Also assume that the conditional probability*

$$P(X_t = x_t|X_{t-1} = x_{t-1}, \cdots, X_0 = x_0),$$

is well–defined and in (0,1). Fix $m_t^1 \in M_t$. Let $g : \mathbb{R} \to \mathbb{R}^+$ be a bijective transformation, then there are unique parameters

$$\{\alpha^t_{i_0,\cdots,i_t}\}_{t\in\mathbb{N},0\leq i_0\leq d_t-1,0\leq i_1\leq d_{t-1},0\leq i_2\leq d_{t-2},\cdots,0\leq i_t\leq d_0},$$

such that

$$P(X_t = x_t|X_{t-1} = x_{t-1}, \cdots, X_0 = x_0) = P_t(x_0, \cdots, x_t),$$

where

$$P_t(x_0, \cdots, x_{t-1}, m_1^t) = \frac{1}{1 + \sum_{y\in M-\{m_1^t\}} h_t(y)}, \tag{3.8}$$

$$P_t(x_0, \cdots, x_{t-1}, x) = \frac{h(x)}{1 + \sum_{y\in M-\{m_1\}} h_t(y)}, x \neq m_1^t \in M_t, \tag{3.9}$$

for $h_t(x_0, \cdots, x_t) = g \circ g_t(x_0, \cdots, x_{t-1}, x_t)$ and

$$g_t(x_0, \cdots, x_{t-1}, x_t) = \sum_{0\leq i_0\leq d_t-1,0\leq i_1\leq d_{t-1},\cdots,0\leq i_t\leq d_0} \alpha^t_{i_0,\cdots,i_t} x_{t-0}^{i_0} \cdots x_{t-t}^{i_t},$$

$$(x_0, \cdots, x_t) \in M_0 \times \cdots \times M_{t-1} \times M_t'.$$

On the other hand any set of arbitrary parameters $\alpha^t_{i_0,\cdots,i_t}$ gives rise to a unique stochastic process with the above equations.

Corollary 3.5.8 *Suppose that $\{X_t\}$ is an rth–order Markov chain where X_t takes values in M_t a finite subset of real numbers, $|M_t| = c_t = d_t + 1 < \infty$, the conditional probability*

$$P(X_t = x_t|X_{t-1} = x_{t-1}, \cdots, X_0 = x_0),$$

is well–defined and belongs to $(0, 1)$. Fix $m_t^1 \in M_t$, let $M_t' = M_t - \{m_t^1\}$ and suppose $g : \mathbb{R} \to \mathbb{R}^+$ is a given bijective transformation. Then

$$g_t(x_t, \cdots, x_0) = g^{-1}\{\frac{P(X_t = x_t|X_{t-1} = x_{t-1}, \cdots, X_0 = x_0)}{P(X_t = m_t^1|X_{t-1} = x_{t-1}, \cdots, X_0 = x_0)}\},$$

is a function of $t+1$ variables for $t < r$, $(x_t, \cdots, x_0)$ and is a function of $r+1$ variables,$(x_t, \cdots, x_{t-r})$, for $t > r$. Hence there exist parameters

$$\{\alpha^t_{i_0,\cdots,i_t}\}_{0\le i_0\le d_t-1, 0\le i_1\le d_{t-1},\cdots,0\le i_t\le d_0}, \text{ for } t < r$$

and

$$\{\alpha^t_{i_0,\cdots,i_r}\}_{0\le i_0\le d_t-1, 0\le i_1\le d_{t-1},\cdots,0\le i_r\le d_{t-r}}, \text{ for } t \ge r$$

such that for $t < r$:

$$g^{-1}\{\frac{P(X_t = x_t|X_{t-1} = x_{t-1},\cdots,X_0 = x_0)}{P(X_t = m^1_t|X_{t-1} = x_{t-1},\cdots,X_0 = x_0)}\} = \sum_{0\le i_0\le d_t-1, 0\le i_1\le d_{t-1},\cdots,0\le i_t\le d_0} \alpha^t_{i_0,\cdots,i_t} x^{i_0}_{t-0}\cdots x^{i_t}_{t-t},$$
$$(x_0,\cdots,x_t) \in M_0\times\cdots M_{t-1}\times M'_t,$$

and for $t \ge r$:

$$g^{-1}\{\frac{P(X_t = x_t|X_{t-1} = x_{t-1},\cdots,X_0 = x_0)}{P(X_t = m^1_t|X_{t-1} = x_{t-1},\cdots,X_0 = x_0)}\} = \sum_{0\le i_0\le d_t-1, 0\le i_1\le d_{t-1},\cdots,0\le i_r\le d_{t-r}} \alpha^t_{i_0,\cdots,i_r} x^{i_0}_{t-0}\cdots x^{i_r}_{t-r}$$
$$(x_0,\cdots,x_t) \in M_0\times\cdots M_{t-1}\times M'_t.$$

Moreover any collection of arbitrary parameters

$$\{\alpha^t_{i_0,\cdots,i_t}\}_{0\le i_0\le d_t-1, 0\le i_1\le d_{t-1},\cdots,0\le i_t\le d_0}, \text{ for } t < r,$$

and

$$\{\alpha^t_{i_0,\cdots,i_r}\}_{0\le i_0\le d_t-1, 0\le i_1\le d_{t-1},\cdots,0\le i_r\le d_{t-r}}, \text{ for } t \ge r,$$

specify a unique stochastic process (upto distribution) by the above relations. In the case of homogenous Markov chains the $\alpha^t_{i_1,\cdots,i_r}$ do not depend on t for $t > r$.

One might question the usefulness of such a representation. After all we have exactly as many parameters in the model as the values of the original function. In the following, we explain the importance of linear representations of such functions.

1. A vast amount of theory has been developed to deal with linear models. Generalized linear models in the case of independent sequence of random variables is a powerful tool. As we will see in sequel, these ideas can be imported into time series using the concept of partial likelihood.

2. Although we have as many parameters in the model as the values of the original function, the representation gives us a convenient framework for modeling, in particular for making various model reductions by omitting some terms or assuming certain coefficients are equal.

3. Although this is a representation for stationary rth–order Markov chains (or representation for arbitrary locally rth–order chains at time t), this representation allows us to accommodate other explanatory variables simply as additive linear terms and extend the model to non–stationary cases. This cannot be done in the same way if we try to model the original values of the function.

Example As an example consider a categorical response variable Y and r categorical explanatory variables

$$X_1, \cdots, X_r,$$

are given. Suppose the X_i takes values in the M_i which include 0. Our purpose is to model Y based on $X_1, \cdots, X_r$. In order to do that, we consider the conditional probability

$$P(Y = y | X_1 = x_1, \cdots, X_r = x_r).$$

Again, we assume that the conditional probability is well-defined everywhere and takes values in $(0, 1)$. The above theorem shows that after applying a transformation the conditional probability can be written as a linear combination of multiples of powers of the X_i.

Although, the theorem above shows the form of the conditional probability in general and paves the way to the estimation of the conditional probabilities by estimating the parameters, the large number of parameters makes this a challenging task which might be impractical in some cases. In the next section, we introduce some classes of r variable functions that can be useful for some applications.

3.5.3 Special cases of functions of r finite variables

The first class of functions we introduce are obtained by power restrictions. We simply assume that g_t can be represented only by powers less than k. Suppose X_t takes values in $0, 1, \cdots, c_t - 1$. Then for a k-restricted power stationary rth–order Markov chain, the g_t, $t > r$ is given by:

$$\sum_{0 \le i_1 \le d_1, \cdots, 0 \le i_r \le d_r, \sum_j i_j \le k} \alpha_{i_1, \cdots, i_r} X_{t-1}^{i_1} \cdots X_{t-r}^{i_r}.$$

In particular, we can let $k = 1$ and get

$$\beta_0 + \sum_i \beta_i X_{t-i}.$$

This is useful especially for binary Markov chains.

The second class of functions are useful in the case when relationships exist between the states in terms of a semi–metric d. Suppose $\{X_t\}$ is an rth–order Markov chain and X_t takes values in the same finite set $M = \{1, \cdots, m\}$. Also let

$$d : M \times M \to \mathbb{R},$$

be a semi–metric being a mapping on M that satisfies the following conditions:

$$\begin{aligned} d &\ge 0; \\ d(x, z) &\le d(x, y) + d(y, z); \\ d(x, x) &= 0. \end{aligned}$$

Then we introduce the following model:

$$g^{-1}\{\frac{P(X_t = j | X_{t-1}, \cdots, X_{t-r})}{P(X_t = 1 | X_{t-1}, \cdots, X_{t-r})}\} = \alpha_{0,j} + \sum_{i=1}^{k} \alpha_{i,j} d(j, X_{t-i})$$

for $j = 2, \cdots, m$. For this model

$$P(X_t = 1 | X_{t-1}, \cdots, X_{t-r}) = 1 - \sum_{j=2, \cdots, m} P(X_t = j | X_{t-1}, \cdots, X_{t-r}).$$

Finally, we introduce a simple class for the binary Markov chain of order r. For any bijective transformation $g : \mathbb{R} \to \mathbb{R}^+$

$$g^{-1}\{\frac{P(X_t = 1 | X_{t-1}, \cdots, X_{t-r})}{P(X_t = 0 | X_{t-1}, \cdots, X_{t-r})}\} = \alpha_0 + \alpha_1 N_{t-1},$$

where $N_{t-1} = \sum_{j=1}^{r} X_{t-j}$. For example in the 0-1 precipitation process example seen in the Introduction, N_{t-1} counts the number of the days out of r days before today that had some precipitation.

3.6 Generalized linear models for time series

Generalized linear models were developed to extend ordinary linear regression to the case that the response is not normal. However, that extension required the assumption of independently observed responses. The notion of partial likelihood was introduced to generalize these ideas to time series where the data are dependent. What follows in this section is a summary of the first chapter in Kedem and Fokianos [27], which we have included for completeness.

Definition Let $\mathcal{F}_t,\ t = 1, 2, \cdots$ be an increasing sequence of σ–fields, $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2, \cdots$ and let $Y_1, Y_2, \cdots$ be a sequence of random variables such that Y_t is $\mathcal{F}_t$–measurable. Denote the density of Y_t, given $\mathcal{F}_t$, by $f_t(y_t; \theta)$, where $\theta \in \mathbb{R}^p$ is a fixed parameter. The partial likelihood (PL) is given by

$$PL(\theta; y_1, \cdots, y_N) = \prod_{t=1}^{N} f_t(y_t; \theta).$$

Example As an example, suppose Y_t represents the 0-1 PN process in Calgary, while MT_t denotes the maximum daily temperature process. We can define $\mathcal{F}_t$ as follows:

1. $\mathcal{F}_t = \sigma\{Y_{t-1}, Y_{t-2}, \cdots\}$. In this case, we are assuming the information available to us is the value of the process on each of the previous days.

2. $\mathcal{F}_t = \sigma\{Y_{t-1}, Y_{t-2}, \cdots MT_{t-1}, MT_{t-2}, \cdots\}$. In this case, we are assuming we have all the information regarding the 0-1 process of precipitation and maximum temperature for previous days.

3. $\mathcal{F}_t = \sigma\{Y_{t-1}, Y_{t-2}, \cdots MT_t, MT_{t-1}, MT_{t-2}, \cdots\}$. In this case, we add to the information in 2 the knowledge of today's maximum temperature.

The vector θ that maximizes the above equation is called the maximum partial likelihood (MPLE). Wong [48] has studied its properties. Its consistency, asymptotic normality and efficiency can be shown under certain regularity conditions.

In this report, we are mainly interested in the case: $\mathcal{F}_t = \sigma\{Y_{t-1}, Y_{t-2}, \cdots\}$. We assume that the information $\mathcal{F}_t$ is given as a vector of random variables and denote it by Z_t, which we call the covariate process:

$$Z_t = (Z_{t1}, \cdots, Z_{tp})'.$$

Z_t might also include the past values of responses $Y_{t-1}, Y_{t-2}, \cdots$.

Let $\mu_t = E[Y_t|\mathcal{F}_{t-1}]$, be the conditional expectation of the response given the information we have up to the time t.

Kedem and Fokianos in [27] address time series following generalized linear models satisfying certain conditions about the so-called random and systematic components:

- Random components: For $t = 1, 2, \cdots, N$

$$f(y_t; \theta_t, \phi|\mathcal{F}_{t-1}) = \exp\{\frac{y_t\theta_t - b(\theta_t)}{a_t(\phi)} + c(y_t; \phi)\}.$$

- The parametric function $\alpha_t(\phi)$ is of the form ϕ/w_t, where ϕ is the dispersion parameter, and w_t is a known parameter called "weight parameter". The parameter θ_t is called the natural parameter.

- Systematic components: For $t = 1, 2, \cdots, N$,

$$g(\mu_t) = \eta_t = \sum_{j=1}^{p} \beta_j Z_{(t-1)j} = Z'_{t-1}\beta,$$

  for some known monotone function g called the link function.

Example Binary time series: As an example consider $\{Y_t\}$, a binary time series. Let us denote by π_t the probability of success given $\mathcal{F}_{t-1}$. Then for $t = 1, 2, \cdots, N$,

$$f(y_t; \theta_t, \phi|\mathcal{F}_{t-1}) = \exp(y_t \log(\frac{\pi_t}{1-\pi_t}) + \log(1-\pi_t))$$

with $E[Y_t|\mathcal{F}_{t-1}] = \pi_t,\ b(\theta_t) = -\log(1-\pi_t) = \log(1+\exp(\theta_t)),\ V(\pi_t) = \pi_t(1-\pi_t),\ \phi = 1,$ and $w_t = 1$.
The canonical link gives rise to the so–called "logistic model":

$$g(\pi_t) = \theta_t(\pi_t) = \log(\frac{\pi_t}{1-\pi_t}) = \eta_t = Z'_{t-1}\beta.$$

In the notation of Corollary 3.5.4, $Y_t = X_t$, $\pi_t = P(X_t = 1|X_{t-1}, \cdots, X_{t-r})$ and $Z'_{t-1} = (1, X_{t-1}, \cdots, X_{t-r}, X_{t-1}X_{t-2}, \cdots, X_{t-1}\cdots X_{t-r})$. We can also consider other covariate processes such as $Z'_{t-1} = (1, X_{t-1}, \cdots, X_{t-r})$ and so on.

In order to study the asymptotic behavior of the maximum likelihood estimator, we consider the conditional information matrix. To establish large sample properties, the stability of the conditional information matrix and the central limit theorem for martingales are required. Proofs may be found in Kedem and Fokianos [27].

Inference for partial likelihood

The definitions of partial likelihood and exponential family of distributions imply that the log partial likelihood is given by

$$l(\beta) = \sum_{t=1}^{N} \log f(y_t; \theta_t, \phi|\mathcal{F}_{t-1}) = \sum_{t=1}^{N} \{\frac{y_t\theta_t - b(\theta_t)}{\alpha_t(\phi)} + c(y_t, \phi)\} \quad =$$
$$\sum_{t=1}^{N} \{\frac{y_t u(z'_{t-1}\beta) - b(u(z'_{t-1}))}{\alpha_t(\phi)} + c(y_t, \phi)\} \quad = \quad \sum_{t=1}^{N} l_t,$$

where $u(.) = (g \circ \mu(.))^{-1} = \mu^{-1}(g^{-1}(.))$, so that $\theta_t = u(z_{t-1}\beta)$. We introduce the notation,

$$\nabla = (\frac{\partial}{\partial\beta_1}, \cdots, \frac{\partial}{\partial\beta_p})'$$

and call $\nabla l(\beta)$ the partial score. To compute the gradient, we can use the chain rule in the following manner

$$\frac{\partial l_t}{\partial\beta_j} = \frac{\partial l_t}{\partial\beta_j}\frac{\partial\theta_t}{\partial\mu_t}\frac{\partial\mu_t}{\partial\eta_t}\frac{\partial\eta_t}{\partial\beta_j}.$$

Some algebra shows

$$S_N(\beta) = \nabla l(\beta) = \sum_{t=1}^{N} Z_{(t-1)}\frac{\partial\mu_t}{\partial\eta_t}\frac{Y_t - \mu_t(\beta)}{\sigma_t^2(\beta)},$$

where, $\sigma_t^2(\beta) = Var[Y_t|\mathcal{F}_{t-1}]$. The partial score process is defined from the partial sums as

$$S_t(\beta) = \nabla l(\beta) = \sum_{s=1}^{t} Z_{(s-1)}\frac{\partial\mu_s}{\partial\eta_s}\frac{Y_s - \mu_s(\beta)}{\sigma_s^2(\beta)}.$$

One can show the terms in the above sums to be orthogonal:

$$E[Z_{(t-1)}\frac{\partial \mu_t}{\partial \eta_t}\frac{Y_t-\mu_t(\beta)}{\sigma_t^2(\beta)}Z_{(s-1)}\frac{\partial \mu_s}{\partial \eta_s}\frac{Y_s-\mu_s(\beta)}{\sigma_s^2(\beta)}]=0,\ s<t.$$

Also, $E[S_N(\beta)=0]$.

The cumulative information matrix is defined by

$$G_N(\beta)=\sum_{t=1}^{N} Cov[Z_{(t-1)}\frac{\partial \mu_t}{\partial \eta_t}\frac{Y_t-\mu_t(\beta)}{\sigma_t^2(\beta)}|\mathcal{F}_{t-1}].$$

The unconditional information matrix is simply

$$Cov(S_N(\beta))=F_N(\beta)=E[G_N(\beta)].$$

Next let

$$H_N(\beta)=-\nabla\nabla' l(\beta).$$

Kedem and Fokianso [27] show that

$$H_N(\beta)=G_N(\beta)-R_N(\beta),$$

where

$$R_N(\beta)=\frac{1}{\alpha_t(\phi)}\sum_{t=1}^{N} Z_{t-1}d_t(\beta)Z'_{t-1}(Y_t-\mu_t(\beta)),$$

and $d_t(\beta)=[\partial^2 u(\eta_t)/\partial\eta_t^2]$.

S_t satisfies the martingale property:

$$E[S_{t+1}(\beta)|\mathcal{F}_{t-1}]=S_t(\beta).$$

To prove the consistency and other properties of the estimators, we need:

Assumption A:

A1. The true parameter β belongs to an open set $B\subset\mathbb{R}$.

A2. The covariate vector Z_t almost surely lies in a non random compact set Γ of $\mathbb{R}^p$, such that $P[\sum_{t=1}^{N} Z_{t-1}Z'_{t-1}>0]=1$. In addition, $Z'_{t-1}\beta$ lies almost surely in the domain H of the inverse link function $h=g^{-1}$ for all $Z_{t-1}\in\Gamma$ and $\beta\in B$.

A3. The inverse link function h, defined in (A2), is twice continuously differentiable and $|\partial h(\lambda)/\partial\lambda|\neq 0$.

A4. There is a probability measure ν on $\mathbb{R}^p$ such that $\int_{\mathbb{R}^p} zz'\nu(dz)$ is positive definite, and such that for Borel sets $A\subset\mathbb{R}^p$,

$$\frac{1}{N}\sum_{t=1}^{N} I_{[Z_{t-1}\in A]}\to\nu(A).$$

Theorem 3.6.1 *Under assumption A the maximum likelihood estimator is almost surely unique for all sufficiently large N, and*

1. *the estimator is consistent and asymptotically normal,*

$$\hat{\beta} \xrightarrow{p} \beta$$

in probability, and

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N_p(0, G^{-1}(\beta)),$$

in distribution as $N \to \infty$, for some matrix G.

2. *The following limit holds in probability, as $N \to \infty$:*

$$\sqrt{N}(\hat{\beta} - \beta) - \frac{1}{\sqrt{N}} G^{-1}(\beta) S_N(\beta) \xrightarrow{p} 0.$$

We follow Kedem and Fokianos [27], who used similar models, to assume the above conditions for our models. However, we conjecture that the above assumptions hold for the partial likelihood of stationary rth–order Markov chains (with strictly positive joint distribution) in terms of our parametric linear form at least for the binary case. In fact assumptions A1. to A3. are easy to check and only A4. poses some challenge. We leave this for future research and use several simulation studies to check the consistency of the estimators in next section as well as Chapter 4 and Chapter 10. For more discussion regarding the assumptions and consistency see [27].

3.7 Simulation studies

This section presents the results of some simulation studies about the partial likelihood applied to categorical rth–order Markov chains. We also investigate the performance of the BIC to pick the appropriate ("true") model. In particular, we generate samples from a seasonal Markov chain Y_t where,

$$Z_{t-1} = (1, Y_{t-1}, cos(\omega t)),\ \omega = \frac{2\pi}{366}.$$

We consider this Markov chain over 5 years from 2000 to 2005 and assume

$$logit\{P(Y_t = 1|Z_{t-1})\} = \beta' Z_{t-1},$$

where $\beta = (-1, 1, -0.5)$.

To generate samples for this chain, we need an initial value of the past two states, which we take it to be $(1, 1)$. We denote the process Y_{t-k} by Y^k for simplicity.

To check the performance of the partial likelihood and estimates of the variance using G_N, we generate 50 chains with this initial value and then compare the parameter estimates with the true parameters. We also compare the theoretical variances with the experimental variances. Table 3.7 shows that the parameter estimates are fairly close to the true values. Also the experimental and theoretical variances are similar.

			sim. sd			theo. sd		
$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$sd(\hat{\beta}_1)$	$sd(\hat{\beta}_2)$	$sd(\hat{\beta}_3)$	$sd(\hat{\beta}_1)$	$sd(\hat{\beta}_2)$	$sd(\hat{\beta}_3)$
-0.99	1.0	-0.42	0.07	0.10	0.07	0.06	0.12	0.07

Table 3.1: The estimated parameters for the model $Z_{t-1} = (1, Y_{t-1}, cos(\omega t))$ with parameters $\beta = (-1, 1, -0.5)$. The standard deviation for the parameters is computed once using G_N (theo. sd) and once using the generated samples (sim. sd).

In Kedem and Fokianos [27] other simulation studies have been done to check the validity of this method.

To check the normality of the parameter estimates, we plot the three parameter estimates histograms in Figure 3.1. The figure shows that the parameter estimates have a distribution close to Gaussian.

Next we check the performance of the BIC criterion in picking the optimal ("true") model. We use the same model as above and then compute the BIC for a few models to see if BIC picks the right one. We denote Y_{t-k} by Y^k and $cos(\omega t)$ by COS for simplicity. For an assessment, we simulate a few other chains.

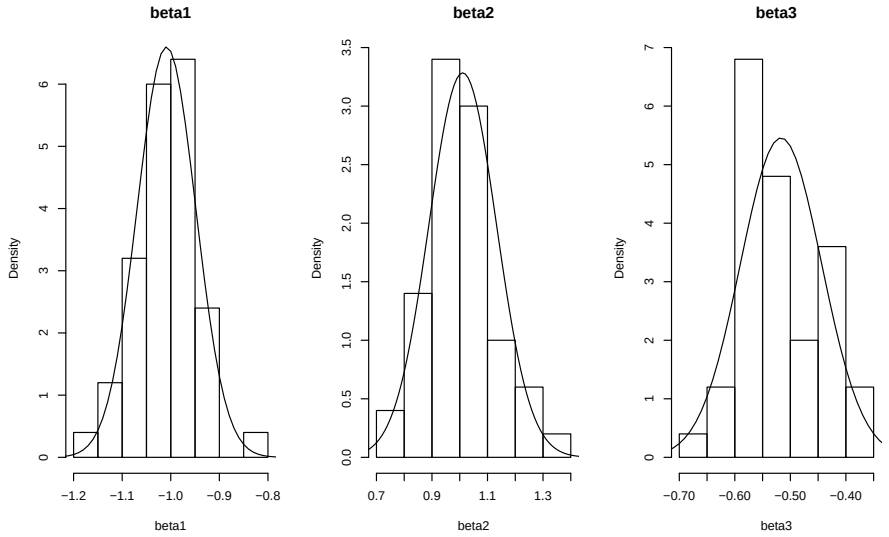

Figure 3.1: The distribution of parameter estimates for the model with the covariate process $Z_{t-1} = (1, Y_{t-1}, cos(\omega t))$ and parameters $(\beta_1 = -1, \beta_2 = 1, \beta_3 = -0.5)$.

Model: Z_{t-1}	BIC	parameter estimates
(1)	2380.0	(-0.605)
$(1, Y^1)$	2267.1	(-1.03, 1.11)
$(1, Y^1, Y^2)$	2273.7	(-1.064, 1.091, 0.101)
$(1, Y^1, COS)$	**2217.7**	(-1.00, 0.970, -0.558)
$(1, Y^1, SIN)$	2274.4	(-1.037, 1.117, 0.026)
$(1, Y^1, COS, SIN)$	2225.1	(-1.00, 0.970, -0.559, 0.028)
$(1, Y^1, Y^2, Y^1Y^2)$	2281.1	(-1.055, 1.0615, 0.0647, 0.077)
$(1, Y^1, Y^2, Y^1Y^2, COS)$	2232.4	(-0.985, 0.943, -0.0870, 0.0915, -0.564)
$(1, Y^1, Y^2, Y^1Y^2, COS, SIN)$	2239.8	(-0.981, 0.957, -0.0946, 0.0723, -0.575, 0.0232)

Table 3.2: BIC values for several models competing for the role of the true model, where $Z_{t-1} = (1, Y^1, COS),\ \beta = (-1, 1, -0.5)$.

As we see in Table 3.2, the true model has the smallest BIC showing it performs well in this case. Also note that models which include the covariates of the true model have accurate estimates for the parameters associated with $(1, Y^1, COS)$, while giving very small magnitude for other parameters.

Model: Z_{t-1}	BIC	parameter estimates
(1)	2537.3	(0.0799)
$(1, Y^1)$	2329.5	(-0.649, 1.417)
$(1, Y^1, Y^2)$	2245.5	(-1.022, 1.144, 0.998)
$(1, Y^1, COS)$	2265.9	(-0.553, 1.236, -0.617)
$(1, Y^1, SIN)$	2336.7	(-0.648, 1.415, -0.0433)
$(1, Y^1, COS, SIN)$	2273.0	(-0.552, 1.235, -0.617, -0.0480)
$(1, Y^1, Y^2, Y^1Y^2)$	2251.3	(-1.08, 1.287, 1.140, -0.278)
$(1, Y^1, Y^2, Y^1Y^2, COS)$	**2213.7**	(-0.936, 1.11, 0.966, -0.175, -0.511)
$(1, Y^1, Y^2, Y^1Y^2, COS, SIN)$	2221.2	(-0.927, 1.101, 0.940, -0.160, -0.549, -0.0441)
$(1, Y^1, Y^2, COS)$	**2206.8**	(-0.899, 1.0263, 0.875, -0.515)

Table 3.3: BIC values for several models competing for the role of true model given by $Z_{t-1} = (1, Y^1, Y^2, COS),\ \beta = (-1, 1, 1, -0.5)$.

Table 3.3 presents the true model in the last row. Ignore that row for a moment. The smallest "BIC" corresponds to $(1, Y^1, Y^2, Y^1Y^2, COS)$, which has an component Y^1Y^2 added to the true model. However, the coefficients of this model are very close to the true model and the coefficient for Y^1Y^2 is relatively small in magnitude. The true model has the smallest BIC again and the parameter estimates are close to the correct values.

3.8 Concluding remarks

In summary, this chapter shows that a categorical discrete–time stochastic process can be represented using a small number of ascending joint distributions

$$P(X_0 = x_0), P(X_0 = x_0, X_1 = x_1), P(X_0 = x_0, X_1 = x_1, X_2 = x_2), \cdots .$$

As a corollary of the above, we showed that a categorical discrete–time stochastic process can be represented using the conditional probabilities

$$P(X_0 = x_0), P(X_1 = x_1 | X_0 = x_0), P(X_2 = x_2 | X_0 = x_0, X_1 = x_1), \cdots .$$

A parametric form was found for the conditional probability distribution of categorical discrete–time stochastic processes. The parameters can be estimated for stationary binary Markov chains using partial likelihood.

Chapter 4

Binary precipitation process

4.1 Introduction

This chapter studies the Markov order of the 0-1 precipitation process (PN from now on). Many authors such as Anderson et al. in [4] and Barlett in [5] have developed techniques to test different assumptions about the order of the Markov chain. For example in [4], Anderson et al. develop a Chi-squared test to test that a Markov chain is of a given order against a larger order. In particular, we can test the hypothesis that a chain is 0th–order Markov against a 1st–order Markov chain, which in this case is testing independence against the usual (1st–order) Markov assumption. (This reduces simply to the well–known Pearson's Chi-squared test.) Hence, to "choose" the Markov order one might follow a strategy of testing 0th–order against 1st–order, testing 1st–order against 2nd–order and so on to rth–order against $(r+1)$th–order, until the test rejects the null hypothesis and then choose the last r as the optimal order. However, some drawbacks are immediately seen with this method:

1. The choice of the significance level will affect our chosen order.
2. The method only works for chains with several independent observations of the same finite chain.
3. We cannot account for some other explanatory variables in the model, for example the maximum temperature.

Issues like this have led researchers to think about other methods of order selection. Akaike in [2], using the information distance and Schwartz in [42] using Bayesian methods develop the AIC and BIC, respectively. Other methods and generalizations of the above methods have been proposed by some authors such as Hannan in [20], Shibita in [44] and Haughton in [22].

Many authors have studied the order of precipitation processes at different locations on Earth. Gabriel et al. in [18] use the test developed in Anderson et al. [4] to show that the precipitation in Tel-aviv is a 1st–order

Markov chain. Tong in [45] used the AIC for Hong Kong, Honolulu and New York and showed that the process is 1st–order in Hong Kong and Honolulu but 0th–order in New York. In a later paper, [46], Tong and Gates use the same techniques for Manchester and Liverpool in England and also re-examined the Tel–aviv data. Chin in [12] studies the problem using AIC over 100 stations (separately) in the United States over 25 years. He concludes that the order depends on the season and geographical location. Moreover, he finds a prevalence of first order conditional dependence in summer and higher orders in winter. Other studies have been done by several authors using similar techniques over other locations. For example, Moon et al. in [35] study this issue at 14 location in South Korea.

This report investigates the Markov order for a cold–climate region. The Markov order of the precipitation in this region might be different due to a large fraction of precipitation being being in the form of snowfall. The report also drops the homogeneity (stationarity) condition usually imposed in studying the Markov order. In fact the model proposed here can accommodate both continuous (here time and potentially geographical location and other explanatory variables) and categorical variables (e.g. precipitation occurred/not occurred on a given day).

An issue with increasing the order of a Markov chain is the exponential increase in number of parameters in the model. Here as a special case, we propose models that increase with the order of Markov chain by adding only 1 parameter. Other authors such as Raftery in [40] and Ching in [13] have proposed other methods to reduce the number of parameters. The dataset used in this study contains more than 110 years of daily precipitation for some stations. This allows us to look at some properties of the precipitation process such as stationarity more closely.

4.2 Models for 0-1 precipitation process

In the light of Categorical Expansion Theorem (Theorem 3.5.6), from the previous chapter, we know all the possible forms of rth–order Markov chains for binary data. Since, this theorem gives us linear forms, time series following generalized linear models (TGLM) provides a method to estimate the parameters. For two reasons it is beneficial to study simpler models rather than a full model:

1. There are a large number of parameters to estimate in the full model.

2. There are better interpretations for the parameters in simpler models.

We introduce a few processes that are useful in modeling precipitation:

- Y_t represents the occurrence of precipitation on day t. Here Y_t is a binary process with 1 denoting precipitation and 0 denoting its absence on day t.
- $N_{t-1}^l = \sum_{j=1}^{l} Y_{t-j}$ represents the number of PN days in the past l days.
- Binary processes for modeling m years, say l_1 to l_2. Here, we define the binary processes A_t^l, $l \in [l_1, l_2]$ by

$$A_t^l = \begin{cases} 1, & \text{if } t \text{ belongs to the year } l \\ 0, & \text{otherwsie} \end{cases}.$$

  This is a binary deterministic process to model the year effect.

- Seasonal processes (deterministic):

$$\cos(\omega t) \text{ and } \sin(\omega t),\ \omega = \frac{2\pi}{366}.$$

  We can also consider higher order terms in the Fourier series $\cos(\omega nt)$ and $\sin(\omega nt)$, where n is a natural number.

Some possibly interesting models present themselves when Z_{t-1} is a covariate process. The probability of precipitation today depends on the value of that covariate process, and those processes might include:

- $Z_{t-1} = (1, N_{t-1}^l)$. This model assumes that the probability of PN today only depends on the number of PN days during l previous days.
- $Z_{t-1} = (1, N_{t-1}^l, Y_{t-1})$. This model assumes that the PN occurrence today depends on the PN occurrence yesterday and the number of PN occurrences during l previous days.
- $Z_{t-1} = (1, \cos(\omega t), \sin(\omega t), N_{t-1}^l, Y_{t-1})$.
- $Z_{t-1} = (1, \cos(\omega t), \sin(\omega t), Y_{t-1})$.
- $Z_{t-1} = (1, Y_{t-1}, \cdots, Y_{t-r})$. This is a special case of Markov chain of order r. No interaction between the days is assumed. In this model increasing the order of Markov chain by one corresponds to adding one parameter to the model.

- $Z_{t-1} = (1, Y_{t-1}, \cdots, Y_{t-r}, Y_{t-1}Y_{t-2})$. In this model, the interaction between the previous day and two days ago is included.
- $Z_{t-1} = (1, \cos(\omega t), \sin(\omega t), Y_{t-1}, \cdots, Y_{t-r})$. In this model, two seasonal terms are added to the previous model.
- $Z_{t-1} = (A_t^1, \cdots, A_t^k, Y_{t-1}, \cdots, Y_{t-r})$. This model has a different intercept for various years (year effect).

4.3 Exploratory analysis of the data

The data includes the daily precipitation for 48 stations over Alberta from 1895 to 2006.

First, we make the plot of transition probabilities for a few locations. We pick Calgary and Banff, which have a rather long period of data available for PN. We have also repeated the procedure for some other locations such as Edmonton and seen similar results. Figures 4.1 to 4.7 show the plots for Banff. For Calgary see plots in Chapter 2. Figure 4.1 plots the estimated 1st–order transition probabilities $\hat{p}_{11}$ (the probability of precipitation if precipitation occurs the day before) and $\hat{p}_{01}$ (the probability of precipitation if it does not occur the day before). These transition probabilities are estimated using the observed data. For example $\hat{p}_{11}$ for January 5th is estimated by $\frac{n_{11}}{n_1}$, where n_{11} is the number of pairs of days (Jan. 4th, Jan. 5th) with precipitation and n_1 is the number of Jan. 5th with precipitation during available years. Figures 4.2 and 4.3 show similar plots for estimated 2nd–order transition probabilities. Figures 4.4 and 4.5 give the estimated annual probability of precipitation for Banff and Calgary computed by dividing the number of wet days of a year by the number of days in that year. The plot of the *logit* function and the transformed estimated probability of precipitation in Banff are shown in Figures 4.6 and 4.7. We summarize the conclusions and conjectures based on the exploratory analysis of the data as followings:

- The binary PN process is not stationary. Figure 4.1 shows that the transition probabilities change over time and depend on the season.
- Figure 4.1 also suggests the transition probabilities change continuously over time. Although a high variation is seen in the higher order probabilities, a generally continuous trend is observed. There is a periodic trend for the transition probabilities over the course of the year

and a simple periodic function should suffice modeling these probabilities.

- Figure 4.1 suggests p_{11} and p_{01} differ over the course of the year, so a 0th–order Markov chain (independent) does not seem appropriate.

- Figure 4.2 plots the curves $\hat{p}_{111}, \hat{p}_{011}$ and Figure 4.3 plots the curves $\hat{p}_{001}, \hat{p}_{101}$. They have considerable overlaps over the course of the year. Therefore a 2nd–order Markov chain does not seem necessary.

- Figures 4.4 and 4.5 show the estimated probability of precipitation for different years, computed by averaging through the days of a given year. The probability of precipitation seems to differ year–to–year. It also seems that consecutive years have similar probability and hence assuming that different years are identically distributed and independent does not seem reasonable. The probability of precipitation has increased over the past century for Calgary, while for Banff the probability of precipitation seems to have been changing with a more irregular pattern.

- Figure 4.6 shows the plot of the *logit* function, while Figure 4.7 shows the result of applying the *logit* function to the estimated probabilities. We observe how the *logit* function transforms the values between 0 and 1 to a wider range in $\mathbb{R}$. Since *logit* is an increasing function the peaks are observed at the same time as the original values.

The Categorical Expansion Theorem (Theorem 3.5.6) shows the general form for binary rth–order Markov processes. Table 4.8 compares all possible 2nd–order Markov chains (including the constant process). We discuss the implications of these possible models and use the following abbreviations: $Y^k = Y_{t-k}$, $COS = cos(\omega t)$, $SIN = sin(\omega t)$, $COS2 = cos(2\omega t)$ and $SIN2 = sin(2\omega t)$.

Some proposed models:

- $Z_{t-1} = 1$:
  The probability of PN's occurrence does not depend on the previous days. In other words days are independent.

- $Z_{t-1} = (1, Y^1)$:
  The probability of PN today depends only on the day before and given the latter's value, it is independent of the other previous days.

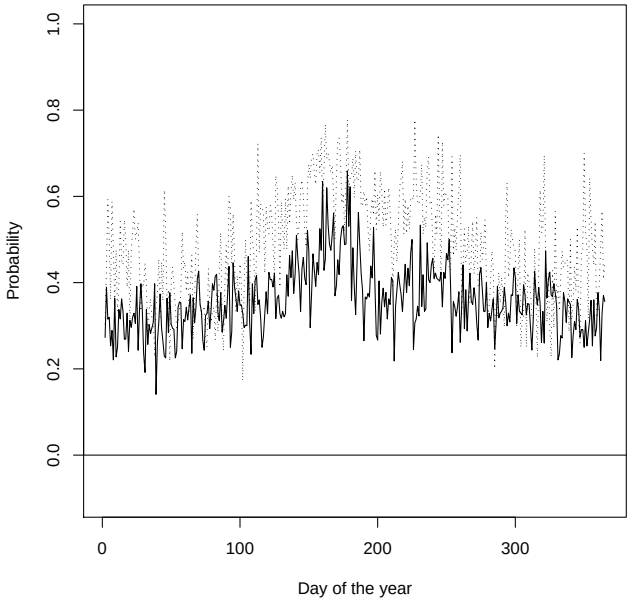

Figure 4.1: The transition probabilities for the Banff site. The dotted line represents $\hat{p_{11}}$ (the estimated probability of precipitation if precipitation occurs the day before) and the dashed represents $\hat{p_{01}}$ (the estimated probability of precipitation if precipitation does not occur the day before.)

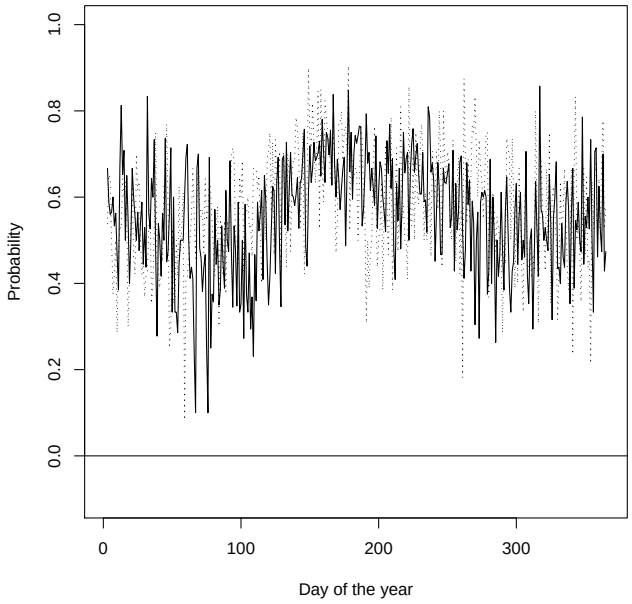

Figure 4.2: The solid curve represents $\hat{p}_{111}$ (the estimated probability of precipitation if during both two previous days precipitation occurs) and the dashed curve represents $\hat{p}_{011}$ (the estimated probability that precipitation occurs if precipitation occurs the day before and does not occur two days ago) for the Banff site.

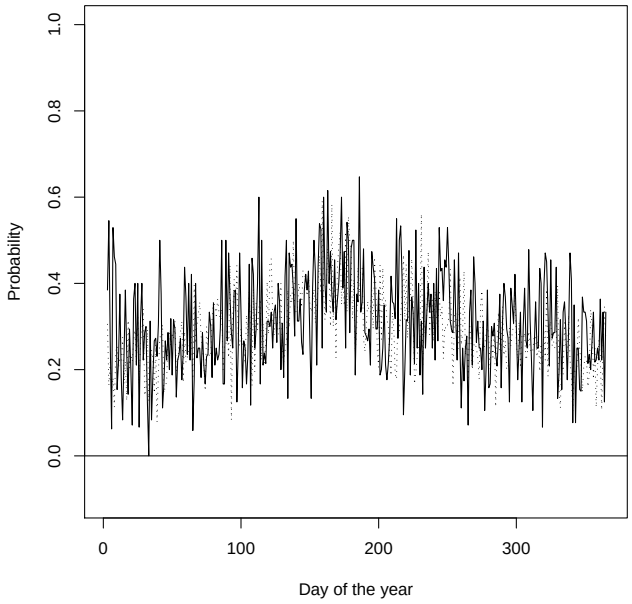

Figure 4.3: The solid curve represents $\hat{p}_{001}$ (the estimated probability of precipitation occurring if it does not occur during the two previous days) and the dotted curve is $\hat{p}_{101}$ (the estimated probability that precipitation occurs if precipitation does not occur the day before but occurs two days ago) for the Banff site.

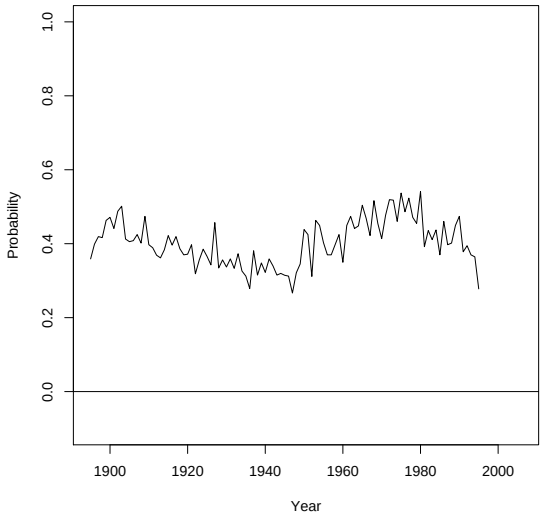

Figure 4.4: Banff's estimated mean annual probability of precipitation calculated from historical data.

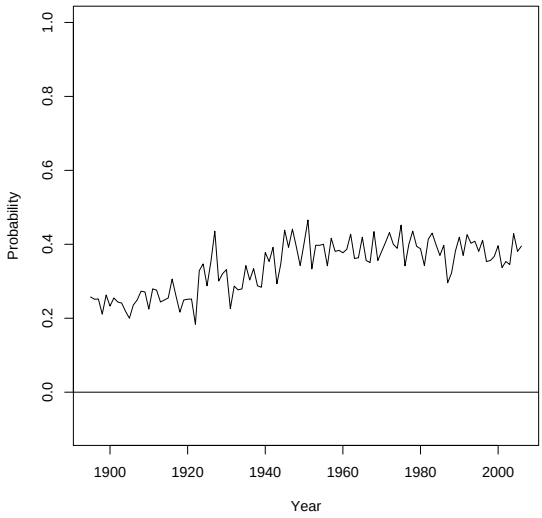

Figure 4.5: Calgary's estimated mean annual probability of precipitation calculated from historical data.

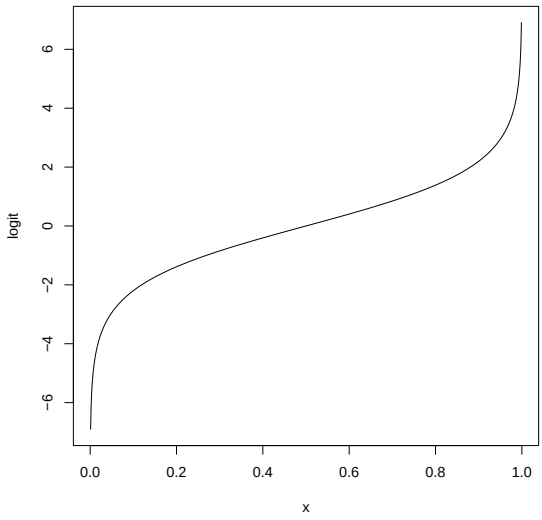

Figure 4.6: The *logit* function: $logit(x) = log(x/(1-x))$.

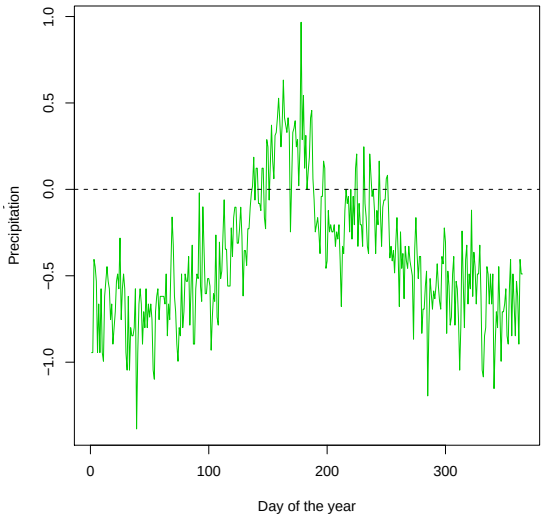

Figure 4.7: The *logit* of the estimated probability of precipitation in Banff for different days of the year.

- $Z_{t-1} = (1, Y^2)$:
  The probability of PN given the information for the day before yesterday is independent of other previous days, in particular yesterday! This does not seem reasonable.

- $Z_{t-1} = (1, Y^1, Y^2)$:
  This model includes both Y^1 and Y^2. One might suspect that it has all the information and therefore is the most general 2nd–order Markov model. However, note that in the model the transformed conditional probability is a linear combination of the past two states:

$$logit\{P(Y = 1|Y^1, Y^2)\} = \alpha_0 + \alpha_1 Y^1 + \alpha_2 Y^2,$$

  which implies,

$$logit\{P(Y = 1|Y^1 = 0, Y^2 = 0)\} = \alpha_0,$$

$$logit\{P(Y = 1|Y^1 = 1, Y^2 = 0)\} = \alpha_0 + \alpha_1,$$

$$logit\{P(Y = 1|Y^1 = 0, Y^2 = 1)\} = \alpha_0 + \alpha_2,$$

  and

$$logit\{P(Y = 1|Y^1 = 1, Y^2 = 1)\} = \alpha_0 + \alpha_1 + \alpha_2.$$

  We conclude that

$$logit\{P(Y = 1|Y^1 = 1, Y^2 = 0)\} - logit\{P(Y = 1|Y^1 = 0, Y^2 = 0)\} =$$
$$logit\{P(Y = 1|Y^1 = 1, Y^2 = 1)\} - logit\{P(Y = 1|Y^1 = 0, Y^2 = 1)\} =$$
$$\alpha_1.$$

  In other words, the model implies that no matter what the value Y^2 has, the differences between the conditional probabilities given $Y^1 = 1$ and given $Y^1 = 0$ (in the *logit* scale) are the same.

- $Z_{t-1} = (1, Y^1 Y^2)$:
  Among other things, this model implies that the conditional probabilities given $(Y^1 = 0, Y^2 = 1)$, $(Y^1 = 1, Y^2 = 0)$ or $(Y^1 = 0, Y^2 = 0)$ are the same.

- $Z_{t-1} = (1, Y^1, Y^1Y^2)$:
  Among other things this model implies that the conditional probabilities given any of the pairs $(Y^1 = 0, Y^2 = 0)$ or $(Y^1 = 0, Y^2 = 0)$ are the same.

- $Z_{t-1} = (1, Y^2, Y^1Y^2)$:
  The interpretation is similar to the previous case.

- $Z_{t-1} = (1, Y^1, Y^2, Y^1Y^2)$:
  This is the full 2nd–order stationary Markov model with no restrictive assumptions as shown by Categorical Expansion Theorem.

The above explanations show that one must be careful about the assumptions made about any proposed model. Including/dropping various covariates can lead to implications that might be unrealistic.

4.4 Comparing the models using BIC

This section uses the methods developed previously to find appropriate models for the 0-1 PN process. We use the PN data for Calgary from 2000 to 2004. We compare several models using the BIC criterion. The partial likelihood is computed and then maximized using the "optim" function in "R".

Using "Time Series Following Generalized Linear Models" as discussed by Kedem et al. in [27], for binary time series with the canonical link function, we have:

$$P(Y_t = 1|Z_{t-1}) = logit^{-1}(\alpha Z_{t-1}),$$

and,

$$P(Y_t = 0|Z_{t-1}) = 1 - logit^{-1}(\alpha Z_{t-1}).$$

We conclude that the log partial likelihood is equal to:

$$\sum_{t=1}^{N} \log P(Y_t|Z_{t-1}) =$$

$$\sum_{1\le t\le N, Y_t=1} \log(logit^{-1}(\alpha Z_{t-1})) + \sum_{1\le t\le N, Y_t=0} \log(1 - logit^{-1}(\alpha Z_{t-1})).$$

To ensure that the maximum picked by "optim" in the R package is close to the actual maximum, several initial values were chosen randomly until stability was achieved.

In order to find an optimal model to describe a binary (0-1) PN process, we can include several factors such as previous values of the process, seasonal terms, previous maximum temperature values and so on. We have done this comparison in several tables. The smallest BIC in the tables is shown by boldface.

Table 4.1 shows the constant process 1 and N^l, the number of wet days during l previous days, as predictors. Note that $N^1 = Y^1$. The BIC criterion in this case picks the simplest model which includes only the previous day. Hence a 1st–order Markov chain is chosen among these particular lth–order chains.

Model: Z_{t-1}	BIC	parameter estimates
$(1, N^1)$	**2268.1**	$(-1.035, 1.268)$
$(1, N^2)$	2294.5	$(-1.097, 0.726)$
$(1, N^3)$	2293.4	$(-1.181, 0.559)$
$(1, N^4)$	2292.7	$(-1.244, 0.462)$
$(1, N^5)$	2296.9	$(-1.281, 0.390)$
$(1, N^6)$	2305.9	$(-1.292, 0.331)$
$(1, N^7)$	2311.3	$(-1.308, 0.291)$
$(1, N^8)$	2317.2	$(-1.317, 0.258)$
$(1, N^9)$	2322.1	$(-1.32, 0.232)$
$(1, N^{10})$	2325.6	$(-1.34, 0.212)$
$(1, N^{11})$	2330.4	$(-1.34, 0.193)$
$(1, N^{12})$	2335.7	$(-1.34, 0.177)$
$(1, N^{13})$	2336.3	$(-1.36, 0.168)$
$(1, N^{14})$	2340.5	$(-1.35, 0.155)$
$(1, N^{15})$	2342.6	$(-1.36, 0.146)$

Table 4.1: BIC values for models including N^l, the number of precipitation days during the past l days for the Calgary site.

Table 4.2 compares models with predictors:

$$1, Y^l \text{ and } N^l,\ l = 1, 2, \cdots, 30.$$

Since $Y^1 = N^1$ the first row is obviously an over–parameterized model. The smallest BIC corresponds to the model $(1, Y^1, N^{28})$. Even the model $(1, Y^1, N^4)$ shows an improvement over $(1, Y^1)$. Hence by adding the number of PN days to the simple model $(1, Y^1)$, an improvement is achieved.

Model: Z_{t-1}	BIC	parameter estimates
$(1, Y^1, N^1)$	2275.6	(-1.04, -0.40, 1.67)
$(1, Y^1, N^2)$	2270.2	(-1.10, 0.94, 0.255)
$(1, Y^1, N^3)$	2258.3	(-1.21, 0.88, 0.279)
$(1, Y^1, N^4)$	2250.6	(-1.28, 0.88, 0.254)
$(1, Y^1, N^5)$	2247.5	(-1.32, 0.91, 0.221)
$(1, Y^1, N^6)$	2248.2	(-1.34, 0.95, 0.187)
$(1, Y^1, N^7)$	2247.1	(-1.37, 0.97, 0.167)
$(1, Y^1, N^8)$	2247.5	(-1.39, 0.99, 0.149)
$(1, Y^1, N^9)$	2247.6	(-1.40, 1.01, 0.136)
$(1, Y^1, N^{10})$	2247.4	(-1.42, 1.02, 0.126)
$(1, Y^1, N^{11})$	2248.3	(-1.43, 1.04, 0.115)
$(1, Y^1, N^{12})$	2249.6	(-1.43, 1.05, 0.105)
$(1, Y^1, N^{13})$	2248.1	(-1.46, 1.06, 0.102)
$(1, Y^1, N^{14})$	2249.7	(-1.46, 1.07, 0.0945)
$(1, Y^1, N^{15})$	2249.5	(-1.47, 1.07, 0.0905)
$(1, Y^1, N^{16})$	2249.0	(-1.49, 1.08, 0.0872)
$(1, Y^1, N^{17})$	2245.3	(-1.51, 1.08, 0.0853)
$(1, Y^1, N^{18})$	2246.8	(-1.53, 1.08, 0.0831)
$(1, Y^1, N^{19})$	2246.8	(-1.55, 1.08, 0.0820)
$(1, Y^1, N^{20})$	2245.6	(-1.56, 1.08, 0.0787)
$(1, Y^1, N^{21})$	2246.0	(-1.56, 1.08, 0.0749)
$(1, Y^1, N^{22})$	2247.6	(-1.55, 1.09, 0.0703)
$(1, Y^1, N^{23})$	2245.9	(-1.58, 1.09, 0.0701)
$(1, Y^1, N^{24})$	2246.0	(-1.58, 1.09, 0.0678)
$(1, Y^1, N^{25})$	2246.8	(-1.58, 1.10, 0.0647)
$(1, Y^1, N^{26})$	2246.6	(-1.59, 1.10, 0.0632)
$(1, Y^1, N^{27})$	2246.2	(-1.60, 1.10, 0.0618)
$(1, Y^1, N^{28})$	**2244.7**	(-1.62, 1.10, 0.0615)
$(1, Y^1, N^{29})$	2245.4	(-1.62, 1.10, 0.0593)
$(1, Y^1, N^{30})$	2246.2	(-1.622, 1.11, 0.0571)

Table 4.2: BIC values for models including N^l, the number of wet days during the past l days and Y^1, the precipitation occurrence of the previous day for the Calgary site.

Table 4.3 compares models with predictors $(1, N^l, COS, SIN)$. We have added (COS, SIN) to capture the seasonality in the precipitation over a year. $(1, N^1, COS, SIN)$ (which is the same as $(1, Y^1, COS, SIN)$) is the winner. Note that this model is better than the simpler model $(1, Y^1)$ or the model $(1, Y^1, N^{28})$.

Model: Z_{t-1}	BIC	parameter estimates
$(1, N^1, COS, SIN)$	**2222.5**	(-1.00, 1.10, -0.588, 0.0999)
$(1, N^2, COS, SIN)$	2254.6	(-1.02, 0.592, -0.564, 0.0977)
$(1, N^3, COS, SIN)$	2260.1	(-1.07, 0.443, -0.538, 0.0961)
$(1, N^4, COS, SIN)$	2264.1	(-1.11, 0.359, -0.518, 0.0959)
$(1, N^5, COS, SIN)$	2270.8	(-1.12, 0.295,-0.508, 0.0971)
$(1, N^6, COS, SIN)$	2280.5	(-1.11, 0.240, -0.510, 0.0999)
$(1, N^7, COS, SIN)$	2286.7	(-1.11, 0.205, -0.508, 0.101)
$(1, N^8, COS, SIN)$	2293.0	(-1.09, 0.176, -0.511, 0.103)
$(1, N^9, COS, SIN)$	2293.1	(-1.08, 0.153, -0.513, 0.105)
$(1, N^{10}, COS, SIN)$	2302.2	(-1.07, 0.136, -0.516, 0.107)

Table 4.3: BIC values for models including N^l, the number of wet days during the past l days and seasonal terms for the Calgary site.

Table 4.4 includes Y^1, seasonal terms and N^l for $l = 1, 2, \cdots, 10$ as predictors. The model with predictors

$$(1, Y^1, N^5, COS, SIN),$$

which includes a combination of seasonal terms and number of precipitation days has the smallest BIC so far. Note that both the seasonal terms and the number of precipitation days prior to the day we are looking at, are indicators of "weather conditions". There are natural cycles throughout the year that can inform us about the weather conditions of a particular day of the year. These natural cycles are modeled by the periodic functions COS and SIN. Also by looking at a short period prior to the current day (short–term past), we might be able to determine the weather conditions. Precipitation may not follow a very regular seasonal pattern similar to temperature as shown in the exploratory analysis. Which one of these variables (seasonal or short–term past) is more important or necessary might depend on the location and other factors.

Model: Z_{t-1}	BIC	parameter estimates
$(1, Y^1, N^1, COS, SIN)$	2230.0	(-1.00, -2.31, 3.41, -0.589, 0.0999)
$(1, Y^1, N^2, COS, SIN)$	2229.2	(-1.03, 0.977, 0.0997, -0.576, 0.0985)
$(1, Y^1, N^3, COS, SIN)$	2224.8	(-1.10, 0.895, 0.156, -0.546, 0.0946)
$(1, Y^1, N^4, COS, SIN)$	2222.1	(-1.14, 0.89, 0.147, -0.525, 0.0941)
$(1, Y^1, N^5, COS, SIN)$	**2221.7**	(-1.16, 0.922, 0.124, -0.515, 0.0934)
$(1, Y^1, N^6, COS, SIN)$	2223.3	(-1.16, 0.959, 0.0954, -0.517, 0.0946)
$(1, Y^1, N^7, COS, SIN)$	2223.7	(-1.17, 0.978, 0.0822, -0.513, 0.0947)
$(1, Y^1, N^8, COS, SIN)$	2224.7	(-1.16, 0.997, 0.0682, -0.515, 0.0945)
$(1, Y^1, N^9, COS, SIN)$	2225.5	(-1.16, 1.0129, 0.0582, -0.515, 0.0961)
$(1, Y^1, N^{10}, COS, SIN)$	2226.0	(-1.16, 1.026, 0.0502, -0.517, 0.0958)

Table 4.4: BIC values for models including N^l, the number of PN days during the past l days, Y^1, the precipitation occurrence of the previous day and seasonal terms for the Calgary site.

Table 4.5 compares models with different number of predictors from $(1, Y^1)$ to

$$(1, Y^1, \cdots, Y^7).$$

The first model is a 1st–order Markov chain and the last one is a 7th–order chain. The optimal model picked is: $(1, Y^1, Y^2, Y^3)$. Comparing this table to Table 4.2, we see that $(1, Y^1, N^3)$ is superior to $(1, Y^1)$, $(1, Y^1, Y^2)$ and $(1, Y^1, Y^2, Y^3)$. Note that $(1, Y^1, N^3)$ is equivalent to $(1, Y^1, Y^2 + Y^3)$. Hence, including Y^2 and Y^3 and giving them the same weight is better than not including them, including one of them or including both of them.

Model: Z_{t-1}	BIC	parameter estimates
$(1, Y^1)$	2268.1	(-1.034, 1.27)
$(1, Y^1, Y^2)$	2270.2	(-1.11, 1.20, 0.23)
$(1, Y^1, Y^2, Y^3)$	**2263.3**	(-1.21, 1.19, 0.140, 0.410)
$(1, Y^1, \cdots, Y^4)$	2263.9	(-1.28, 1.16, 0.133, 0.334, 0.281)
$(1, Y^1, \cdots, Y^5)$	2268.5	(-1.32, 1.15, 0.121, 0.328, 0.232, 0.192)
$(1, Y^1, \cdots, Y^6)$	2335.4	(-1.34, 1.15, 0.0837, 0.357, 0.213, 0.135, 0.115)
$(1, Y^1, \cdots, Y^7)$	2286.7	(-1.51, 1.33, -0.113, 0.378, 0.418, 0.204, -0.0050, 0.214)

Table 4.5: BIC values for Markov models of different order with small number os parameters for the Calgary site.

Table 4.6 compares models with different Markov orders plus the seasonal terms. The model $(1, Y^1, COS, SIN)$ is the winner. Hence, whether we include the seasonal terms or not, the model that only depends on the previous day is the winner.

Model: Z_{t-1}	BIC	parameter estimates
$(1, COS, SIN, Y^1)$	**2222.6**	(-1.0, -0.5, 0.1, 1.1)
$(1, COS, SIN, Y^1, Y^2)$	2229.1	(-1.0, -0.5, 0.1, 1.0, 0.1)
$(1, COS, SIN, Y^1, Y^2, Y^3)$	2230.4	(-1.1, -0.5, 0.1, 1.0, 0.02, 0.3)
$(1, COS, SIN, Y^1, \cdots, Y^4)$	2247.3	(-1.1, -0.5, 0.1, 1.0, 0.03, 0.2, 0.15)
$(1, COS, SIN, Y^1, \cdots, Y^5)$	2243.4	(-1.3, -0.4, 0.2, 1.4, -0.4, -0.1, 1.0, -0.15)
$(1, COS, SIN, Y^1, \cdots, Y^6)$	2501.6	(-1.2, -1.5, 0.4, 0.2, 0.8, 0.9, 0.9, -0.6, -0.2)
$(1, COS, SIN, Y^1, \cdots, Y^7)$	2447.3	(-1.1, -0.2, 0.07, 0.8, -0.02, 0.3, 0.4, -0.07, 0.4, -0.3)

Table 4.6: BIC values for Markov models with different order plus seasonal terms for the Calgary site.

Table 4.7 studies seasonality more. We consider the possibility that there are more/less terms of the Fourier series of a periodic function over the year. It turns out that the model with $(1, Y^1, COS)$ is the optimal model so far. Hence, only one term seem to suffice modeling the seasonal nature of the process.

Model: Z_{t-1}	BIC	parameter estimates
$(1, COS)$	2322.7	(-0.556, -0.717)
$(1, SIN)$	2424.3	(-0.523, 0.115)
$(1, COS, SIN)$	2327.3	(-0.568, -0.738, 0.119)
$(1, Y^1, COS)$	**2216.9**	(-1.00 , 1.10, -0.587)
$(1, Y^1, SIN)$	2273.9	(-1.03, 1.26, 0.0933)
$(1, Y^1, COS, SIN)$	2222.6	(-1.004, 1.102, -0.589, 0.100)
$(1, Y^1, COS, SIN, COS2)$	2229.7	(-1.00, 1.10, -0.586, 0.0998, 0.0247)
$(1, Y^1, COS, SIN, SIN2)$	2230.0	(-1.00, 1.10, -0.590, 0.101, 0.0125)
$(1, Y^1, COS, SIN, COS2, SIN2)$	2237.2	(-1.01, 1.11, -0.575, 0.0978, 0.0236, -0.0101)

Table 4.7: BIC values for models including seasonal terms and the occurrence of precipitation during the previous day for the Calgary site.

Table 4.8 compares all stationary 2nd–order Markov models. The smallest BIC corresponds to $(1, Y^1)$.

Model: Z_{t-1}	BIC	parameter estimates
(1)	2419.6	(-0.528)
$(1, Y^1)$	**2268.0**	(-1.04, 1.27)
$(1, Y^2)$	2392.8	(-0.756, 0.590)
$(1, Y^1, Y^2)$	2270.2	(-1.110, 1.197, 0.256)
$(1, Y^1 Y^2)$	2335.5	(-0.779, 1.134)
$(1, Y^1, Y^1 Y^2)$	2272.7	(-1.040, 1.113, 0.282)
$(1, Y^2, Y^1 Y^2)$	2342.3	(-0.757, -0.113, 1.225)
$(1, Y^1, Y^2, Y^1 Y^2)$	2277.7	(-1.103, 1.177, 0.234, 0.048)

Table 4.8: BIC values for 2nd–order Markov models for precipitation at the Calgary site.

Table 4.9 compares all 2nd–order Markov chains with a seasonal COS term. The model $(1, Y^1, COS)$ is the winner.

Model: Z_{t-1}	BIC	parameter estimates
$(1, COS)$	2322.7	(-0.567, -0.738)
$(1, COS, Y^1)$	**2216.8**	(-1.005, -0.587, 1.106)
$(1, COS, Y^2)$	2317.4	(-0.708, -0.679, 0.372)
$(1, COS, Y^1 Y^2)$	2223.5	(-0.760, -0.618, 0.905)
$(1, COS, Y^1, Y^2)$	2276.1	(-1.033, -0.575, 1.080, 0.103)
$(1, COS, Y^1, Y^1 Y^2)$	2223.9	(-1.004, -0.580, 1.041, 0.120)
$(1, COS, Y^2, Y^1 Y^2)$	2280.9	(-0.709, -0.632, -0.244, 1.093)
$(1, COS, Y^1, Y^2, Y^1 Y^2)$	2231.0	(-1.028, -0.575, 1.065, 0.085, 0.037)

Table 4.9: BIC values for 2nd–order Markov models for precipitation at the Calgary site plus seasonal terms.

Table 4.10 also includes the maximum and minimum temperature of the day before, as predictors of some of the models which performed better in the above tables. We have also included the annual processes $A^1, \cdots, A^5$ to one of the models. Finally, we have included the model $(1, Y^1, N^5, COS)$. This model has a combination of the seasonal term COS and the short–term past process N^5 which did the best when combined with the seasonal terms and Y^1 in Table 4.4. It turns out that including MT and mt does not improve the BIC as well as does the annual terms. However, $(1, Y^1, N^5, COS)$ has the smallest BIC in all the models, which is a seasonal Markov chain of order 5 with only 4 parameters. Also the simpler model,

$$(1, Y^1, COS),$$

has a close BIC to $(1, Y^1, N^5, COS)$.

Model: Z_{t-1}	BIC	parameter estimates
$(1, COS, Y^1)$	**2216.8**	(-1.005, -0.587, 1.106)
$(1, Y^1, COS, MT^1)$	2221.7	(-0.84, 1.0, -0.74, -0.012)
$(1, Y^1, COS, mt^1)$	2224.2	(-1.0, 1.0, -0.65, -0.0055)
$(1, Y^1, COS, MT^1, mt^1)$	2227.4	(-0.65, 0.99, -0.67, -0.025, 0.022)
$(1, Y^1, COS, A^1, \cdots, A^5)$	2241.2	(1.1, -0.5, -0.9, -1.2, -1.1, -1.0, -0.7)
$(1, Y^1, N^5, COS, MT^1)$	2297.3	(-2.13, 0.9, 0.4, 0.6, 0.2, 0.04)
$(1, Y^1, N^5, COS, SIN, MT^1, mt^1)$	2516.8	(1.4, 0.04, 0.2, 0.7, 0.8, -0.2, 0.3)
$(1, Y^1, N^5, COS, MT^1, mt^1)$	2393.9	(1.4, 0.7, -0.1, -0.5, 0.5, -0.1, 0.2)
$(Y^1, N^5, COS, MT^1, A^1, \cdots, A^5)$	2697.1	(1.23, -0.64, -2.0, -0.10, 2.0, 1.2, 2.2, 1.2, 1.8)
$(Y^1, N^5, COS, A^1, \cdots, A^5)$	2447.1	(0.1, 0.1, -0.7, -0.39, -0.01, -0.2, -0.9, -1)
$(1, Y^1, MT^1)$	2251.5	(-1.2, 1.3, 0.021)
$(1, Y^1, N^5, COS)$	**2215.8**	(-1.1, 0.9, 0.1, -0.5)
$(1, Y^1, N^5, COS, MT^1)$	2223.8	(-1.2, 0.9, 0.1, -0.4, 0.0)

Table 4.10: BIC values for models including several covariates as temperature, seasonal terms and year effect for precipitation at the Calgary site.

4.5 Changing the location and the time period

This section compares various models for a different time period and location. Table 4.11 compares various models for the 0-1 PN process in Calgary between 1990 and 1994 which is a 5–year period. In Table 4.12, we have compared several models for 0-1 PN process over Medicine Hat site between 2000 and 2004.

Table 4.11 shows that among the compared models $(1, Y^1, COS)$ has the smallest BIC. In particular the BIC for this model is smaller than the BIC for $(1, Y^1, N^5, COS)$ which has the smallest BIC for Calgary 2000–2004. However $(1, Y^1, COS)$ was the second optimal model also for Calgary 2000–2004 with a close BIC to the optimal. Including the maximum and minimum temperature to the model increases the BIC again.

Model: Z_{t-1}	BIC	parameter estimates
$(1, Y^1)$	2312.7	(-0.931, 1.275)
$(1, Y^1, Y^2)$	2318.8	(-0.967, 1.238, 0.126)
$(1, Y^1, COS)$	**2228.8**	(-0.858, 1.036, -0.712)
$(1, Y^1, N^5)$	2303.3	(-1.168, 1.012, 0.168)
$(1, Y^1, N^{10})$	2287.9	(-1.581, 1.015, 0.132)
$(1, Y^1, N^{15})$	2282.7	(-1.486, 1.045, 0.105)
$(1, Y^1, COS, SIN)$	2231.9	(-0.855, 1.026, -0.715 , 0.152)
$(1, Y^1, N^5, COS)$	2236.4	(-0.864, 1.032, 0.004, -0.709)
$(1, Y^1, N^5, SIN)$	2307.8	(-1.160, 1.011, 0.164, 0.125)
$(1, Y^1, N^5, COS, SIN)$	2239.4	(-0.849, 1.031, -0.004, -0.718, 0.152)
$(1, Y^1, N^{10}, COS)$	2236.4	(-0.847, 1.030, -0.002, -0.721, 0.153)
$(1, Y^1, N^{10}, COS, SIN)$	2239.4	(-0.847, 1.030, -0.002, -0.721 , 0.153)
$(1, Y^1, N^5, COS, MT^1)$	2244.3	(-0.433, 1.046, -0.096, -1.078, -0.021)
$(1, Y^1, N^5, COS, mt^1)$	2244.1	(-0.910, 1.011, 0.031, -0.584, 0.006)

Table 4.11: BIC values for several models for the binary process of precipitation in Calgary, 1990–1994

Table 4.12 shows that the smallest BIC corresponds to $(1, Y^1, COS)$. However, several models have similar BIC values. Also, including the maximum and minimum temperature increases the BIC here.

Model: Z_{t-1}	BIC	parameter estimates
$(1, Y^1)$	2202.9	(-1.138, 1.094)
$(1, Y^1, Y^2)$	2207.9	(-1.183, 1.051, 0.181)
$(1, Y^1, N^5)$	2203.6	(-1.275, 0.921, 0.119)
$(1, Y^1, N^{10})$	2228.9	(-0.858, 1.036, -0.712)
$(1, Y^1, N^{15})$	2200.5	(-1.420, 0.980, 0.065)
$(1, Y^1, N^{20})$	2202.5	(-1.421, 1.008, 0.048)
$(1, Y^1, COS)$	**2201.2**	(-1.134, 1.067, -0.224)
$(1, Y^1, COS, SIN)$	2202.9	(-1.132, 1.052, -0.225, 0.177)
$(1, Y^1, N^5, COS)$	2203.9	(-1.252, 0.924, 0.101, -0.201)
$(1, Y^1, N^5, SIN)$	2206.6	(-1.263, 0.922, 0.109, 0.158)
$(1, Y^1, N^5, COS, SIN)$	2206.6	(-1.239, 0.925, 0.091, -0.204, 0.163)
$(1, Y^1, N^{10}, COS)$	2201.9	(-1.336, 0.958, 0.073, -0.183)
$(1, Y^1, N^{10}, COS, SIN)$	2205.1	(-1.311, 0.958, 0.065, -0.187, 0.151)
$(1, Y^1, N^5, COS, MT^1)$	2306.5	(-1.455, 2.099, -0.130, 0.041, 0.004)
$(1, Y^1, N^5, COS, mt^1)$	2211.1	(-1.238, 0.937, 0.087, -0.267, -0.005)
$(1, Y^1, N^{15}, COS)$	2202.7	(-1.363, 0.981, 0.053, -0.175)

Table 4.12: BIC values for several models for precipitation occurrence in Medicine Hat, 2000-2004

In summary, in all the three cases

$$(1, Y^1, COS),$$

is either optimal or the second to the optimal (using BIC). We have also tried BIC for Calgary with a long time period of close to 100 years and surprisingly the same simple model $(1, Y^1, COS)$ was the optimal.

Chapter 5

On the definition of "quantile" and its properties

5.1 Introduction

This chapter points out deficiencies in the classical definition (as well as some other widely used definitions) of the median and more generally the quantile and the so-called quantile function. Moreover redefining it appropriately gives us a basis on which we can find necessary and sufficient conditions for the sample quantiles to converge for arbitrary distribution functions. In the next chapter, we define a "degree of separation" function to measure the goodness of the approximation (or estimation). We argue that this function can be viewed as a natural loss function for assessing estimations and approximations. One characteristic of this loss function is its invariance under strictly monotonic transformations of the random variable, in particular re-scaling.

In this chapter, we have used the terms data vector, approximation, estimation, exact and true quantiles repeatedly. To clarify what we mean by these terms, we give the following explanations:

- Data vector: A vector of real numbers. We do not consider these values as random in general. We use the term random vector or random sample for a vector of random variables. We define the quantile for data vectors, but the same definition applies to a random sample.
- Approximation and exact value: Suppose a very large data vector is given. We can compute the exact mean/median of such a vector by using all the data and the definition of mean/median. One can approximate the mean/median using various techniques. Note that both approximation and exact terms are used for data vectors of (non-random) numbers.
- Estimation and true value: Estimation means finding functions of the random sample to estimate parameters of the underlying distribution.

The parameters are called the true values.

The sample definition of quantiles varies in different text books. In [24], Hyndman et al. point out many different definitions in statistical packages for quantiles of a sample. In [17], Freund et al. point out various definitions for quartiles of data and propose a new definition using the concept of "hinge".

The traditional definition of quantiles for a random variable X with distribution function F,

$$lq_X(p) = \inf\{x|F(x) \geq p\},$$

appears in classic works as [38]. We call this the "left quantile function". In some books (e.g. [41]) the quantile is defined as

$$rq_X(p) = \sup\{x|F(x) \leq p\},$$

this is what we call the "right quantile function". Also in robustness literature people talk about the upper and lower medians which are a very specific case of these definitions. However, we do not know of any work that considers both definitions, explore their relation and show that considering both has several advantages.

A physical motivation is given for the right/left definition of quantiles. It is widely claimed that (e.g. Koenker in [29] or Hao and Naiman in [21]) the traditional quantile function is invariant under monotonic transformations. We show that this does not hold even for strictly increasing functions. However, we prove that the traditional quantile function is invariant under non-decreasing left continuous transformations. We also show that the right quantile function is invariant under non-decreasing right continuous transformations. A similar neat result is found for continuous decreasing transformations using the Quantile Symmetry Theorem also proved in this chapter.

Suppose we know that a data point is larger than a known number of other data points and smaller than another known number of data points. Of interest are the quantiles to which this data point corresponds. Lemma 5.2.4 gives a result about this. We will use this lemma later to establish the precision of our proposed algorithm for approximating quantiles of large datasets.

Quantiles are often used as the inverse of distribution functions. In general neither the distribution function nor the quantile function are invertible. However Lemma 5.5.1 shows how quantiles can be used to characterize sets of the form $\{x|F(x) < p\}$, a case that is equivalent to $(-\infty, lq_F(p))$.

Lemma 5.7.1 shows the left continuity of the left quantile function and the right continuity of the right quantile function.

Section 5.8 finds necessary and sufficient conditions for the left and right quantile functions to be equal at $p \in [0, 1]$. We also find out that the left and right quantile functions coincide except for at most a countable number of values in [0,1]. Then we characterize the image of the the left and right quantile functions and show that the image corresponds to "heavy" points (heavy point is a point that the probability of being in a neighborhood around that point is positive).

Section 5.9 shows that given any of lq, rq and F uniquely determines the other two and formulas are given in order to find them. We also show that if one of lq and rq is two-sided continuous then so is the other one. Lemma 5.10.1 shows that the strict monotonicity of the distribution function F on its "real domain" $\{x|0 < F(x) < 1\}$ is equivalent to two-sided continuity of lq/rq. Conversely, strict monotonicity of lq/rq corresponds to continuity of F.

Section 5.12 presents the desirable "Quantile Symmetry Theorem", a result that could be only obtained by considering both left and right quantiles. This relation can help us prove several other useful results regarding quantiles. Also using the quantile symmetry theorem, we find a relation for the equivariance property of quantiles under non-increasing transformations.

Section 5.14 studies the limit properties of left and right quantile functions. In Theorem 5.14.7, we show that if left and right quantiles are equal, i.e. $lq_F(p) = rq_F(p)$, then both sample versions lq_{F_n}, rq_{F_n} are convergent to the common distribution value. We found an equivalent statement in Serfling [43] with a rather similar proof. The condition for convergence there is said to be $lq_F(p)$ being the unique solution of $F(x-) < p \leq F(x)$ which can be shown to be equivalent to $lq_F(p) = rq_F(p)$. Note how considering both left and right quantiles has resulted in a cleaner, more comprehensible condition for the limits. In a problem Serfling asks to show with an example that this condition cannot be dropped. We show much more by proving that if $lq_F(p) \neq rq_F(p)$ then both $rq_{F_n}(p)$ and $rq_{F_n}(p)$ diverge almost surely. The almost sure divergence result can be viewed as an extension to a well-known result in probability theory which says that if $X_1, X_2, \cdots$ an i.i.d sequence from a fair coin with -1 denoting tail and 1 denoting head and $Z_n = \sum_{i=1}^{n} X_i$ then $P(Z_n = 0 \ i.o.) = 1$. The proof in [9] uses the Borel–Cantelli Lemma to get around the problem of dependence of Z_n. This is equivalent to saying for the fair coin both $lq_{F_n}(1/2)$ and $rq_{F_n}(1/2)$ diverge almost surely. For the general case, we use the Borel–Cantelli Lemma again. But we also need a lemma (Lemma 5.14.10) which uses the Berry–Esseen Theorem in

its proof to show the deviations of the sum of the random variables can become arbitrarily large, a result that is easy to show as done in [9] for the simple fair coin example. Finally, we show that even though in the case that $lq_F(p) \neq rq_F(p)$, lq_{F_n}, rq_{F_n} are divergent; for large ns they will fall in

$$(lq_F(p) - \epsilon, lq_F(p)] \cup [rq_F(p), rq_F(p) + \epsilon).$$

In fact we show that

$$\liminf_{n\to\infty} lq_{F_n}(p) = \liminf_{n\to\infty} rq_{F_n}(p) = lq_F(p)$$

and

$$\limsup_{n\to\infty} lq_{F_n}(p) = \limsup_{n\to\infty} rq_{F_n}(p) = rq_F(p).$$

The proof is done by constructing a new random variable Y from the original random variable X with distribution function F_X by shifting back all the values greater than $rq_X(p)$ to $lq_X(p)$. This makes $lq_Y(p) = rq_Y(p)$ in the new random variable. Then we apply the convergence result to Y.

5.2 Definition of median and quantiles of data vectors and random samples

This section presents a way to define quantiles of data vectors and random samples. We confine our discussion to data vectors since the definition for random samples is merely a formalistic extension. Suppose, we are given a very long data vector. The goal is to find the median of this vector. Let us denote the data vector by $x = (x_1, \cdots, x_n)$. Suppose $y = (y_1, \cdots, y_n)$ is an increasing sorted vector of elements of $x = (x_1, \cdots, x_n)$. Then usually the median of x is defined to be $y_{(n+1)/2}$ if n is odd and $\frac{y_{n/2}+y_{(n+2)/2}}{2}$ if n is even.

Essentially the median is defined so that half data lies below it and half lies above it. However, when n is even, any value between $y_{n/2}$ and $y_{(n+2)/2}$ serves this purpose and taking the average of the two values seems arbitrary.

Intuitively, the quantile should have the following properties:

1. It should be a member of the data vector. In other words if $x = (x_1, \cdots, x_n)$ is the data vector then the quantile should be equal to one of $x_i,\ i = 1, \cdots, n$.

2. Equivariance: If we transform the data using an increasing continuous transformation of $\mathbb{R}$, find the quantile and transform back, we should get the same result, had we found the quantile of the original data.

More formally, if we denote the quantile of a data vector x for $p \in (0,1)$ by $q_x(p)$ then for any $\phi : \mathbb{R} \to \mathbb{R}$ strictly increasing and bijective

$$q_x(p) = y \Leftrightarrow q_{\phi(x)}(p) = \phi(y).$$

3. Symmetry: The p-th quantile of the data vector $x = (x_1, \cdots, x_n)$ should be the negative of $(1-p)$-th quantile of data vector $-x = (-x_1, \cdots, -x_n)$:

$$q_x(p) = -q_{-x}(1-p).$$

Particularly, the median of x should be the image of the median of the image of x with respect to 0.

4. The "amount" of data between $q_x(p_1)$ and $q_x(p_2)$ should be $p_2 - p_1$ of the the "data amount" of the whole vector if $p_1 < p_2$.

5. If we "cut" a sorted data vector up until the p_1-th quantile and compute the p_2-th quantile for the new vector, we should get the p_1p_2-th quantile of the original vector. For example the median of a sorted vector upto its median should be the first quartile.

This chapter develops a definition for quantiles that satisfies the first three conditions. We will address the last two conditions in later chapters and develop a framework in which they are satisfied.

Consider the example $x = (0, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 10)$. We see that the median by the usual definition is 1.5 not apparent in the observed data. Also if we take bijective, increasing and continuous transformation $\phi(x) = x^3$, we see that the classic definition does not satisfy the second property.

The median and quantiles can be defined both for distributions and data vectors (and random samples). For a random variable X having a distribution function F, the p-th quantile is traditionally defined as

$$q_F(p) = \inf\{x | F(x) \geq p\}. \tag{5.1}$$

This can be used to define the quantiles of a data vector using the empirical (sample) distribution function F_n,

$$F_n(x) = \sum_{i=1}^{n} 1_{(-\infty, x_i]}(x).$$

With this definition of the quantile, the equivariance property holds and the result is a realizable data value. This definition faces another issue however. Consider flipping a fair coin with outcomes: 0,1. Then the distribution of X is given by

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1/2 & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

Hence by definition 5.1, $q_F(p) = 0,\ p \leq 1/2$ and $q_F(p) = 1,\ p > 1/2$. This all seems to be reasonable other than $q_F(p) = 0,\ p = 1/2$. Based on the symmetry of the distribution there should not be any advantage for 0 over 1 to be the median. For the quantiles of the data vectors the same issue occurs. For example, consider $x = (1, 2, 3, 4, 5, 6)$ and apply definition 5.1 to F_n corresponding to this data vector. We will get 3 as the median but in fact 4 should to be as eligible by symmetry.

Before to get to our definition of quantile we provide the following motivating examples.

Example A student decided to buy a new memory chip for his computer. He needed to choose between the available RAM sizes (1 GB, 2GB etc) in his favorite store. In a trade-off between price and speed, he decided to get a RAM chip that is at least as large as 2/3 RAMs bought in the store during the day before. He could access the information regarding all RAMs bought the day before, in particular their size. He entered the size data into the R package he had recently downloaded for free. He had heard about the quantiles in his elementary statistics course so he decided to compute the quantile of the data for $p = 2/3$. When he computed that he got 2.666 (GB). He knew a RAM of size 2.666 does not exist and concluded this must be a result of an interpolation procedure in R. Since the closest integer to 2.666 is 3 he concluded that 3 GB is the size he is looking for. He went back to the store asking for 3 GB RAM and was told they have never sold such a RAM in that store! He thought there must be an error in the dataset so he looked the data again

$$1, 1, 1, 1, 2, 2, 2, 2, 4, 4, 4, 4$$

Surprisingly there was no 3. R had interpolated 2 and 4 to give 2.66 and mislead the student.

Example A supervisor asked 2 graduate students to summarize the following data regarding the intensity of the earthquakes in a specific region:

row number	M_L (Richter)	A (shaking amplitude)
1	4.21094	1.62532×10^4
2	4.69852	4.99482×10^4
3	4.92185	8.35314×10^4
4	5.12098	13.21235×10^4
5	5.21478	16.39759×10^4
6	5.28943	19.47287×10^4
7	5.32558	21.16313×10^4
8	5.47828	30.08015×10^4
9	5.59103	38.99689×10^4
10	5.72736	53.37772×10^4

Table 5.1: Earthquakes intensities

Earthquake intensity is usually measured in M_L scale, which is related to A by the following formula:

$$M_L = \log_{10} A.$$

In the data file handed to the students (Table 5.1), the data is sorted with respect to M_L in increasing order from top to bottom. Hence the data is arranged decreasingly with respect to A from top to bottom.

The supervisor asked two graduate students to compute the center of the intensity of the earthquakes using this dataset. One of the students used A and the usual definition of median and so obtained

$$(16.39759 \times 10^4 + 19.47287 \times 10^4)/2 = 17.93523 \times 10^4.$$

The second student used the M_L and the usual definition of median to find

$$(5.21478 + 5.28943)/2 = 5.252105.$$

When the supervisor saw the results he figured that the students must have used different scales. Hence he tried to make the scales the same by transforming one of the results

$$10^{5.252105} = 17.86920 \times 10^4.$$

To his surprise the results were not quite the same. He was bothered to notice that the definition of median is not invariant under the change of scale which is continuous strictly increasing.

Example A scientist asked two of his assistants to summarize the following data regarding the acidity of rain:

row number	pH	aH
1	4.7336	18.4672×10^{-6}
2	4.8327	14.6994×10^{-6}
3	4.8492	14.1514×10^{-6}
4	5.0050	9.8855×10^{-6}
5	5.0389	9.1432×10^{-6}
6	5.2487	5.6403×10^{-6}
7	5.2713	5.3543×10^{-6}
8	5.2901	5.1274×10^{-6}
9	5.5731	2.6724×10^{-6}
10	5.6105	2.4519×10^{-6}

Table 5.2: Rain acidity data

pH is defined as the cologarithm of the activity of dissolved hydrogen ions (H^+).

$$pH = -\log_{10} aH.$$

In the data file handed to the students (Table 5.2) the data is sorted with respect to pH in increasing order from top to bottom. Hence the data is arranged decreasingly with respect to aH from top to bottom.

The scientist asked the two assistant to compute the 20th and 80th percentile of the data to get an idea of the variability of the acidity. First assistant used the pH scale and the traditional definition of the quantile

$$q_F(p) = \inf\{x | F(x) \geq p\},$$

where F is the empirical distribution of the data. He got the following two numbers

$$q_F(0.2) = 4.8327 \text{ and } q_F(0.8) = 5.2901 \tag{5.2}$$

these values are positioned in row 2 and 8 respectively.

The second assistant also used the traditional definition of the quantiles and the aH scale to get

$$q_F(0.2) = 2.6724 \times 10^{-6} \text{ and } q_F(0.8) = 14.1514 \times 10^{-6}, \tag{5.3}$$

which correspond to row 9 and 3.

The scientist noticed the assistants used different scales. Then he thought since one of the scales is in the opposite order of the other and 0.2 and 0.8 are the same distance from 0 and 1 respectively, he must get the other assistant's result by transforming one. So he transformed the second assistant's results given in Equation 5.3 (or by simply looking at the corresponding rows, 9 and 3 under pH), to get

$$5.5731 \text{ and } 4.8492,$$

which are not the same as the first assistants result in Equation 5.2. He noticed the position of these values are only one off from the previous values (being in row 9 and 3 instead of 8 and 2).

Then he tried the same himself for 25th and 75th percentile using both scales

$$pH : q_F(0.25) = 4.8492 \text{ and } q_F(0.75) = 5.2901,$$

which are positioned at 3rd and 8th row.

$$aH : q_F(0.25) = 5.1274 \times 10^{-6} \text{ and } q_F(0.25) = 14.1514 \times 10^{-6},$$

which are positioned at row 8th and 3rd. This time he was surprised to observe the symmetry he expected. He wondered when such symmetry exist and what is true in general. He conjectured that the asymmetric definition of the traditional quantile is the reason of this asymmetry. He also thought that the symmetry property is off at most by one position in the dataset.

To define the quantile, we perform a thought experiment and use our intuition to decide how it should be defined. Suppose a data vector $x = (x_1, \cdots, x_n)$ is given. Define the *sort* operator which permutes the components of a vector to give a vector with non-decreasing coordinates by

$$sort(x) = (y_1, \cdots, y_n).$$

In statistics y_i defined as above is called the i–th order statistics of x and is usually denoted by $x_{(i)}$ or $x_{i:n}$. [This definition extends to random vectors $(X_1, \cdots, X_n)$ as well.] The concept of quantile should only depend on $sort(x)$. Let $z = (z_1, \cdots, z_r)$ be the non–decreasing subvector of all distinct elements of x. If z_i is repeated m_i times, we say z_i has multiplicity m_i and therefore $\sum_{i=1}^{r} m_i = n$. Now imagine, a uniform bar of length 1. Cut the bar from left to right to r parts of lengths $\frac{m_1}{n}, \cdots, \frac{m_r}{n}$ proportional to

the multiplicity of the z_i. Assign a unique color to every $z_i,\ i = 1, \cdots, r$ and color its piece with that color. Then reassemble the stick from left to right in the original order. To define the p-th quantile measure a length p from the left hand of the bar (whose total length is one). Determine the reassembled bar's color at that point. However, this protocol fails at the end points as well as the points where two colors meet. Since each color is an equally eligible choice, we are led to the idea in defining the quantiles of a two–state solution at these points, giving us the left and right quantiles. But proceeding with our bar analogy, the intersection points and boundary points are:

$$0, \frac{m_1}{n}, \frac{m_1 + m_2}{n}, \cdots, \frac{m_1 + \cdots + m_{r-1}}{n}, 1.$$

By the above discussion, if p is not an intersection/boundary point both left and right quantiles, which we denote by lq_x and rq_x respectively should be the same and equal to

$$lq_x(p) = rq_x(p) = \begin{cases} z_1 & 0 < p < \frac{m_1}{n} \\ z_i & \frac{m_1+\cdots+m_{i-1}}{n} < p < \frac{m_1+\cdots+m_i}{n} \\ z_r & \frac{m_1+\cdots+m_{r-1}}{n} < p < 1 \end{cases}$$

For the intersection points, if $p = \frac{m_1+\cdots+m_{i-1}}{n}$ then

$$lq_x(p) = z_{i-1} \ \text{ and } \ rq_x(p) = z_i.$$

For the boundary points we define

$$lq_x(0) = -\infty, rq_x(0) = z_1, lq_x(1) = z_r, rq_x(1) = \infty.$$

As a convention, for a sorted vector y of length n, we define $y_0 = -\infty$ and $y_{n+1} = \infty$.

Lemma 5.2.1 *Suppose x is a data vector of length n and $y = sort(x) = (y_1, \cdots, y_n)$. Also let $y_0 = -\infty$ and $y_{n+1} = \infty$. For $0 < p < 1$, let $[np]$ denote the integer part of np. Then*

$$a)\ np = [np] \Rightarrow lq_x(p) = y_{[np]}, \quad rq_x(p) = y_{[np]+1}.$$

$$b)\ np > [np] \Rightarrow lq_x(p) = y_{[np]+1}, \quad rq_x(p) = y_{[np]+1}.$$

c) $y = sort(x)$ and $p_i = i/n,\ i = 0, 1, \cdots, n$, implies

$$y = (lq_x(p_1), \cdots, lq_x(p_n)) = (rq_x(p_0), \cdots, rq_x(p_{n-1})).$$

Proof
a) Let $np = h \in \mathbb{N}$. There are four cases:

1. For $h = 0$ and $h = n$ the result is trivial by the definition of y_0 and y_{n+1}.

2. $0 < h < m_1 \Rightarrow 0 < p < m_1/n$ and by definition $lq_x(p) = rq_x(p) = z_1$. But $y_h = y_{h+1} = z_1$.

3. There exists $1 < i \leq r$ such that $m_1 + \cdots + m_{i-1} < h < m_1 + \cdots + m_i \Rightarrow \frac{m_1+\cdots+m_{i-1}}{n} < p < \frac{m_1+\cdots+m_i}{n}$ and by definition $lq_x(p) = rq_x(p) = z_i$. But $y_h = y_{h+1} = z_i$ since $m_1 + \cdots + m_{i-1} < h < m_1 + \cdots + m_i$.

4. $h = m_1 + \cdots + m_i, i < r \Rightarrow p = \frac{m_1+\cdots+m_i}{n}$, $i < r$. By definition since this is an intersection point $lq_x(p) = z_i$ and $rq_x(p) = z_{i+1}$. But $z_i = y_h$ and $z_{i+1} = y_{h+1}$.

b) Let $h = [np] \Rightarrow \frac{h}{n} < p < \frac{h+1}{n}$. Since h and $h + 1$ differ exactly by one unit, there exists an i such that

$$\frac{m_1 + \cdots + m_{i-1}}{n} \leq \frac{h}{n} < p < \frac{h+1}{n} \leq \frac{m_1 + \cdots + m_i}{n}.$$

Then by definition $lq_x(p) = rq_x(p) = z_i$. But since

$$m_1 + \cdots + m_{i-1} < h + 1 \leq m_1 + \cdots + m_i,$$

$y_{h+1} = z_i$.
c) Straightforward consequence of the definition. ■

Suppose $y' \in \{y_1, \cdots, y_n\}$, for future reference, we define some additional notations for data vectors.

Definition The minimal index of y', $m(y')$ and the maximal index of y', $M(y')$ are defined as below:

$$m(y') = \min\{i | y_i = y'\},\ M(y') = \max\{i | y_i = y'\}.$$

It is easy to see that in $y = sort(x) = (y_1, \cdots, y_n)$ all the coordinates between $m(y')$ and $M(y')$ are equal to y'. Also note that if $y' = z_i$ then $M(y') - m(y') + 1 = m_i$ is the multiplicity of z_i. We use the notation m_x and M_x whenever we want to emphasize that they depend on the data vector x.

Lemma 5.2.2 *Suppose $x = (x_1, \cdots, x_n)$, $y = sort(x)$ and z a non–decreasing vector of all distinct elements of x. Then*
a) $m(z_{i+1}) = M(z_i) + 1, \;\; i = 0, \cdots, r-1$.
b) Suppose ϕ is a bijective increasing transformation over $\mathbb{R}$,

$$m_\phi(x)(\phi(z_i)) = m_x(z_i),$$

and

$$M_{\phi(x)}(\phi(z_i)) = M_x(z_i),$$

for $i = 1, \cdots, r$.

Proof a) is straightforward.
b) Note that

$$m_x(y') = \min\{i | y_i = y'\} = \min\{i | \phi(y_i) = \phi(y')\} = m_{\phi(x)}(\phi(y')).$$

A similar argument works for M_x. ■

We also define the position and standardized position of an element of a data vector.

Definition Let $x = (x_1, \cdots, x_n)$ be a vector and $y = sort(x) = (y_1, \cdots, y\_n)$. Then for $y' \in \{y_1, \cdots, y_n\}$, we define

$$pos_x(y') = \{m_x(y'), m_x(y') + 1, \cdots, M_x(y')\},$$

where *pos* stands for position. Then we define the standardized position of y' to be

$$spos_x(y') = (\frac{m_x(y') - 1}{n}, \frac{M_x(y')}{n}).$$

In the following lemma we show that for every $p \in spos(y')$ (and only $p \in spos(y')$), we have $rq(p) = lq(p) = y'$. For example if $1/2 \in spos(y')$ then y' is the (left and right) median.

Lemma 5.2.3 *Suppose $x = (x_1, \cdots, x_n)$, $y = sort(x) = (y_1, \cdots, y_n)$ and $y' \in \{y_1, \cdots, y_n\}$. Then*

$$p \in spos_x(y') \Leftrightarrow lq_x(p) = rq_x(p) = y'.$$

Proof Let $z = (z_1, \cdots, z_r)$ be the reduced vector with multiplicities $m_1, \cdots, m_r$. Then $y' = m_i$ for some $i = 1, \cdots, r$.

case I: If $i = 2, \cdots, r$, then

$$m(y') = m_1 + \cdots + m_{i-1} + 1,$$

and

$$M(y') = m_1 + \cdots + m_i.$$

case II: If $i = 1$, then $m(y') = 1$ and $M(y') = m_1$.

In any of the above cases for $p \in (\frac{m(y')-1}{n}, \frac{M(y')}{n})$ and only $p \in (\frac{m(y')-1}{n}, \frac{M(y')}{n})$

$$rq_x(p) = lq_x(p) = z_i,$$

by definition. ■

Now we prove a lemma that will become useful later on. It is easy to see that if $u \in pos(y')$ then

$$(\frac{u-1}{n}, \frac{u}{n}) \subset spos(y').$$

We conclude that

$$\cup_{u \in pos(y')} (\frac{u-1}{n}, \frac{u}{n}) \subset spos(y').$$

In fact $spos(y')$ can possibly have a few points on the edge of the intervals not in $\cup_{u \in pos(y')} (\frac{u-1}{n}, \frac{u}{n})$.

Lemma 5.2.4 *Suppose x is a data vector of length n and y' is an element of this vector. Also assume*

$$y' \geq x_i, \;\; i \in I, \quad y' \leq x_j, \;\; j \in J,$$

$$I \cap J = \phi, \quad I, J \subset \{1, 2, \cdots, n\}.$$

Then there exist a p in $(\frac{|I|-1}{n}, 1 - \frac{|J|}{n})$ that belongs to $spos(y')$. In other words $lq(p) = rq(p) = y'$.

Proof From the assumption, we conclude that $pos(y')$ includes a number between $|I|$ and $n - |J|$. Let us call it u_0. Hence $(\frac{u_0-1}{n}, \frac{u_0}{n}) \subset spos(y')$. Since $|I| \leq u_0 \leq n - |J|$, we conclude that $spos(y')$ intersects with

$$\cup_{|I| \leq u \leq n-|J|} (\frac{u-1}{n}, \frac{u}{n}) \subset (\frac{|I|-1}{n}, 1 - \frac{|J|}{n}).$$

■

5.3 Defining quantiles of a distribution

So far, we have only defined the quantile for data vectors. Now we turn to defining the quantile for distribution functions.

The p-th quantile for a random variable X with distribution function F as pointed out above is traditionally defined to be

$$q(p) = \inf\{u|F(u) \geq p\}.$$

We showed by an example above the asymmetry issue to which that definition can lead. We show that the issue arises due to the flatness of F in an interval. To get around this problem as the case of data vectors, we define the left and right quantile for the distribution F as follows:

$$lq_F(p) = \inf\{u|F(u) \geq p\},$$

and

$$rq_F(p) = \inf\{u|F(u) > p\}.$$

If there are more than one random variables in the discussion, to avoid confusion, we use the notations lq_{F_X}, rq_{F_X}. Also when there is no chance of confusion, we simply use lq, rq. The reason for this definition should become clear soon. First let us apply this definition to the fair coin example. If $p \neq 1/2$ then both $lq_F(p)$ and $rq_F(p)$ will be the same and give us the same value. However, $lq_F(1/2) = 0$ and $rq_F(1/2) = 1$. This is exactly what one would hope for. To see the consequences of this definition, we prove the following lemma:

Lemma 5.3.1 *(Quantile Properties Lemma) Suppose X is a random variable on the probability space (Ω, Σ, P) with distribution function F:*

a) $F(lq_F(p)) \equiv P(X \leq lq_F(p)) \geq p$.

b) $lq_F(p) \leq rq_F(p)$.

c) $p_1 < p_2 \Rightarrow rq_F(p_1) \leq lq_F(p_2)$. *This and (b) imply that*

$$lq_F(p_1) \leq rq_F(p_1) \leq lq_F(p_2) \leq rq_F(p_2).$$

d) $rq_F(p) = \sup\{x|F(x) \leq p\}$.

e) $P(lq_F(p) < X < rq_F(p)) = 0$. *In other words if* $lq_F(p) < rq_F(p)$ *then* F *is flat in the interval* $(lq_F(p), rq_F(p))$.

f) $P(X < rq_F(p)) \leq p$.

g) If $lq_F(p) < rq_F(p)$ *then* $F(lq_F(p)) = p$ *and hence* $P(X \geq rq_F(p)) = 1-p$.

h) $lq_F(1) > -\infty, rq_F(0) < \infty$ *and* $P(rq_F(0) \leq X \leq lq_F(1)) = 1$.

i) $lq_F(p)$ *and* $rq_F(p)$ *are non–decreasing functions of* p.

j) Suppose F *has a jump at* x, *in other words* $P(X = x) > 0$, *which is equivalent to* $\lim_{y \to x^-} F(y) < F(x)$. *Then* $lq_F(F(x)) = x$.

k) $x < lq_F(p) \Rightarrow F(x) < p$ *and* $x > rq_F(p) \Rightarrow F(x) > p$.

Proof

a) Take a strictly decreasing sequence $\{x_n\}$ in $\mathbb{R}$ that tends to $lq(p)$. For every x_n, $F(x_n) \geq p$ since $x_n > lq(p)$. Otherwise

$$F(x_n) < p \Rightarrow F(y) < p, \ \ \forall y \leq x_n.$$

Hence $(-\infty, x_n] \cap \{y | F(y) \geq p\} = \emptyset$. We conclude that

$$lq(p) = \inf\{y | F(y) \geq p\} \geq x_n > lq(p),$$

which is a contradiction. Now since F is right continuous

$$\lim_{n \to \infty} F(x_n) = F(lq(p)).$$

But $F(x_n) \geq p, \ \ \forall n \in \mathbb{N}$. Hence $\lim_{n \to \infty} F(x_n) \geq p$.

b) Note that $\{u | F(u) > p\} \subset \{u | F(u) \geq p\}$.

c) Note that $\{x | F(x) \geq p_2\} \subset \{x | F(x) > p_1\}$ if $p_2 > p_1$.

d) Suppose $p \in [0,1]$ is given. Let $A = \{x|F(x) > p\}$ and $B = \{x|F(x) \leq p\}$. We want to show that $\inf A = \sup B$.

Consider two cases:

1) Suppose $\inf A < \sup B$. Then pick $\inf A < y < \sup B$. We get a contradiction as follows:
$\inf A < y \Rightarrow F(y) > p$. Otherwise, since F is increasing $F(y) \leq p \Rightarrow y < x, \ \forall x \in A \Rightarrow y \leq \inf A$.
$y < \sup B \Rightarrow F(y) \leq p$. Otherwise, since F is increasing $F(y) > p \Rightarrow y > x, \ \forall x \in B \Rightarrow y \geq \sup B$.
We conclude $F(y) > p$ and $F(y) \leq p$, a contradiction.

2) Suppose $\sup B < \inf A$. Take $\sup B < y < \inf A$.
$\sup B < y \Rightarrow F(y) > p$. Otherwise, $F(y) \leq p \Rightarrow y \in B \Rightarrow y \leq \sup B$.
$y < \inf A \Rightarrow F(y) \leq p$. Otherwise $F(y) > p \Rightarrow y \in A \Rightarrow y \geq \inf A$.
Once more $F(y) > p$ and $F(y) \leq p$ which is a contradiction.

e) Suppose F is not flat in that interval. $\exists v_1 < v_2 \in (lq(p), rq(p))$ such that $F(v_2) > F(v_1)$.

$$F(v_2) > F(v_1) \geq F(lq(p)) \geq p.$$

This is a contradiction since $v_2 < rq(p)$.

f) Take an increasing sequence $x_n \uparrow rq_F(p)$, then note that $P(X \leq x_n) \leq p$ since $x_n < rq_F(p)$. Let $A_n = \{X \leq x_n\}$ and $A = \{X < rq_F(p)\}$ then $\lim_{n\to\infty} A_n = A$, by continuity of the probability (See [9]):

$$P(X < rq_F(p)) = P(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} P(A_n) \leq p.$$

g) By a) $F(lq_F(p)) = P(X \leq lq_F(p)) \geq p$. Suppose $P(X \leq lq_F(p)) > p$. This implies that $lq_F(p) \geq rq_F(p)$. By b) we get $lq_F(p) = rq_F(p)$, which is a contradiction.

h) Note that

$$lq_F(0) = \inf\{x|F(x) \geq 0\} = \inf \mathbb{R} = -\infty.$$

Suppose $rq_F(0) = \infty$. Then

$$\{x|F(x) > 0\} = \emptyset \Rightarrow \forall x \in \mathbb{R}, \ \ F(x) = 0,$$

a contradiction to the properties of a distribution function F.

Also note that

$$rq_F(1) = \inf\{x|F(x) > 1\} = \inf \emptyset = \infty.$$

Suppose $lq_F(1) = -\infty$. Then

$$\inf\{x|F(x) \geq 1\} = -\infty \Rightarrow$$

$$\forall x \in \mathbb{R},\ F(x) \geq 1 \Rightarrow \forall x \in \mathbb{R},\ \ F(x) = 1,$$

a contradiction. For the second part note that $rq_F(0) \leq lq_F(1)$ by (c). Then

$$\begin{aligned} &P(rq_F(0) \leq X \leq lq_F(1)) = \\ &1 - P(lq_F(1) < X < rq_F(1)) \quad - \quad P(lq_F(0) < X < rq_F(0)) = \\ &\qquad 1 - 0 - 0, \end{aligned}$$

by part (e).

i) Trivial.

j) Suppose $P(X = x) > 0$ then $\lim_{y \to x^-} F(y) = P(X < x) < P(X < x) + P(X = x) = F(x)$. Now assume that $\lim_{y \to x^-} F(y) < F(x)$, then $P(X < x) < F(x) \Rightarrow P(X = x) > 0$.

   To prove that in this case $lq_F(F(x)) = x$, let $p = F(x)$ we want to show $lq_F(p) = x$. Note that $F(x) = p$ gives $lq_F(p) \leq x$. On other hand for any $y < x$, we know that $F(y) < p$, by a) y cannot be $lq_F(p)$. Hence $x = lq_F(F(x))$.

k) First part follows from the definition of lq and the second part from part (d).

■

The following lemma is useful in proving that a specific value is the left or right quantile for a given p.

Lemma 5.3.2 *(Quantile value criterion)*

a) $lq_F(p)$ is the only a satisfying (i) and (ii), where
(i) $F(a) \geq p$,
(ii) $x < a \Rightarrow F(x) < p$.

b) $rq_F(p)$ *is the only* a *satisfying (i) and (ii), where*
(i) $x < a \Rightarrow F(x) \leq p$,
(ii) $x > a \Rightarrow F(x) > p$.

Proof

a) Both properties hold for $lq_F(p)$ by previous lemma. If both $a < b$ satisfy them, then $F(a) \geq p$ by (i). But since b satisfies the properties and $a < b$, by (ii), $F(a) < p$ which is a contradiction.

b) Both properties hold for $rq_F(p)$ by previous lemma. If both $a < b$ satisfy them, then we can get a contradiction similar to above.

■

5.4 Left and right extreme points

In Lemma 5.3.1, we showed these properties about $rq_X(0)$ and $lq_X(1)$:

$$rq_X(0) < \infty, \quad lq_X(1) > -\infty,$$
$$rq_X(0) \leq lq_X(1),$$

and

$$P(rq_X(0) \leq X \leq lq_X(1)) = 1.$$

The above states that all the mass is between these two values. We will show in the next lemma that these values are also the minimal values to satisfy this property. This is the motivation for the following definition.

Definition We call $rq_F(0)$ the "left extreme" and $lq_F(1)$ the "right extreme" of the distribution function F.

Lemma 5.4.1 *(Left and right extreme points property)*
Suppose X *is a random variable with distribution function* F.
a) The right extreme $lq_F(1)$ *is the smallest* a *satisfying*

$$P(X \leq a) = 1.$$

In other words

$$\min_a \{P(X \leq a) = 1\} = lq_F(1).$$

b) The left extreme $rq_F(0)$ is the biggest a satisfying

$$P(X \geq a) = 1.$$

$$\max_a \{P(X \geq a) = 1\} = rq_F(0).$$

c) Consider the following subset of $\mathbb{R}^2$

$$I^2 = \{(a,b) \in \mathbb{R}^2 | P(X \in [a,b]) = 1\}.$$

Then

$$\cap_{(a,b)\in I^2}[a,b] = [rq_X(0), lq_X(1)].$$

Proof a) In Lemma 5.3.1, we showed $F(lq_F(1)) = 1$. Also $F(a) < 1$ for $a < lq_F(1)$ by the definition of lq_F.
b) In Lemma 5.3.1, we showed $P(X \geq rq_X(0)) = 1$. Suppose $a > rq_X(0)$. Then since $rq_X(p) = \inf\{x|F(x) > 0\}$,

$$\begin{aligned} \exists c \in \{x|F(x) > 0\}, c < a &\Rightarrow \\ \exists c < a, F(c) > 0 &\Rightarrow \\ \exists c,\ P(X < a) \geq F(c) > 0 &\Rightarrow \\ P(X \geq a) = 1 - P(X < a) &< 1. \end{aligned}$$

c) This is straightforward from a) and b). ■

5.5 The quantile functions as inverse

The following lemma shows that lq_X and rq_X can be considered as the inverse of the distribution function in some sense.

Lemma 5.5.1 *(Quantile functions as inverse of the distribution function)*
a) $F(x) < p \Leftrightarrow x < lq_X(p)$. *(i.e.* $\{x|F(x) < p\} = (-\infty, lq_F(p))$.*)*
b) $\{x|F(x) \leq p\} = (-\infty, rq_X(p)]$ *or* $(-\infty, rq_X(p))$.
c) If F *is continuous at* $rq_X(p)$ *then* $\{x|F(x) \leq p\} = (-\infty, rq_X(p)]$.
d) $\{x|F(x) \geq p\} = [lq_X(p), \infty)$.
e) $\{x|F(x) > p\} = (rq_X(p), \infty)$ *or* $[rq_X(p), \infty)$.
f) If F *is continuous then* $\{x|F(x) > p\} = (rq_X(p), \infty)$.

Proof

a) ($\Rightarrow$) is true because otherwise if $x \geq lq_X(p) \Rightarrow F(x) \geq F(lq_X(p)) \geq p$, which is a contradiction. To show ($\Leftarrow$) note that by the definition of $lq_X(p)$, if $F(x) \geq p$ then $x \leq lq_X(p)$.

b) We need to show that (1) $(-\infty, rq_X(p)) \subset \{x|F(x) \leq p\}$ and (2) $\{x|F(x) \leq p\} \subset (-\infty, rq_X(p)]$. For (1), suppose $x < rq_X(p)$. We claim $F(x) \leq p$. Otherwise if $F(x) > p$ by the definition of $rq_X(p)$, $rq_X(p) \leq x$. For (2), suppose $F(x) \leq p$. Then since $rq_X(p) = \sup\{x|F(x) \leq p\}$, we conclude $x \leq rq_X(p)$.

c) By Part (b), it suffices to show $F(rq_X(p)) = p$. This is shown in the next lemma.

d) R.H.S $\subset$ L.H.S by Lemma 5.3.1 part (a). L.H.S $\subset$ R.H.S by the definition of lq.

e) Note that $x > rq_F(p)$ then $F(x) > p$ by Lemma 5.3.1 part (k). Also $F(x) > p \Rightarrow rq_F(p) \leq x$ by definition of rq.

f) This is a consequence of part (e) and next lemma.

■

For the continuous distribution functions, we have the following lemma.

Lemma 5.5.2 *(Continuous distributions inverse) If F is continuous $F(x) = p \Leftrightarrow x \in [lq_X(p), rq_X(p)]$.*

Proof If $x < lq_X(p)$ then we already showed that $F(x) < p$. Also if $x > lq_X(p)$ then $rq_X(p) = \sup\{x|F(x) \leq p\} \Rightarrow F(x) > p$. (Because otherwise if $F(x) \leq p \Rightarrow rq_X(p) \geq p$.) It remains to show that $F(lq_X(p)) = F(rq_X(p)) = p$. But by Lemma 5.3.1, we have

$$F(lq_X(p)) \geq p.$$

Hence it suffices to show that $F(rq_X(p)) \leq p$. But by Part (f) of Lemma 5.3.1 and continuity of F

$$F(rq_F(x)) = P(X \leq rq_F(x)) = P(X < rq_F(x)) \leq p.$$

■

5.6 Equivariance property of quantile functions

Example (Counter example for Koenker–Hao claim) Suppose X is distributed uniformly on [0,1]. Then $lq_X(1/2) = 1/2$. Now consider the following strictly increasing transformations

$$\phi(x) = \begin{cases} x & -\infty < x < 1/2 \\ x+5 & x \geq 1/2 \end{cases}.$$

Let $T = \phi(X)$ then the distribution of T is given by

$$P(T \leq t) = \begin{cases} 0 & t \leq 0 \\ t & 0 < t \leq 1/2 \\ 1/2 & 1/2 < t \leq 5 + 1/2 \\ t-5 & 5+1/2 < t \leq 5+1 \\ 1 & t > 5+1 \end{cases}.$$

It is clear form above that $lq_T(1/2) = 1/2 \neq \phi(lq_X(1/2)) = \phi(1/2) = 5 + 1/2$.

We start by defining

$$\phi^{\leq}(y) = \{x | \phi(x) \leq y\},\ \phi^{\star}(y) = \sup \phi^{\leq}(y),$$

and

$$\phi^{\geq}(y) = \{x | \phi(x) \geq y\},\ \phi_{\star}(y) = \inf \phi^{\geq}(y).$$

Then we have the following lemma.

Lemma 5.6.1 *Suppose ϕ is non-decreasing.*

a) If ϕ is left continuous then

$$\phi(\phi^{\star}(y)) \leq y.$$

b) If ϕ is right continuous then

$$\phi(\phi_{\star}(y)) \geq y.$$

Proof

a) Suppose $x_n \uparrow \phi^\star(y)$ a strictly increasing sequence. Then since $x_n < \phi^\star(y)$, we conclude $x_n \in \phi^{\leq}(y) \Rightarrow \phi(x_n) \leq y$. Hence $\lim_{n\to\infty} \phi(x_n) \leq y$. But by left continuity $\phi(x_n) \uparrow \phi(\phi^\star(y))$.

b) Suppose $x_n \downarrow \phi_\star(y)$ a strictly decreasing sequence. Then since $x_n > \phi^\star(y)$, we conclude $x_n \in \phi^{\geq}(y) \Rightarrow \phi(x_n) \geq y$. Hence $\lim_{n\to\infty} \phi(x_n) \geq y$. But by right continuity $\phi(x_n) \downarrow \phi(\phi_\star(y))$.

■

Theorem 5.6.2 *(Quantile Equivariance Theorem) Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is non-decreasing.*

a) If ϕ is left continuous then

$$lq_{\phi(X)}(p) = \phi(lq_X(p)).$$

b) If ϕ is right continuous then

$$rq_{\phi(X)}(p) = \phi(rq_X(p)).$$

Proof

a) We use Lemma 5.3.2 to prove this. We need to show (i) and (ii) in that lemma for $\phi(lq_X(p))$. First note that (i) holds since

$$F_{\phi(X)}(\phi(lq_X(p))) = P(\phi(X) \leq \phi(lq_X(p))) \leq P(X \leq lq_X(p)) \geq p.$$

For (ii) let $y < \phi(lq_X(p))$. Then we want to show that $F_{\phi(X)}(y) < p$. It is sufficient to show $\phi^\star(y) < lq_X(p)$. Because then

$$P(\phi(X) \leq y) \leq P(X \leq \phi^\star(y)) < p.$$

To prove $\phi^\star(y) < lq_X(p)$, note that by the previous lemma

$$\phi(\phi^\star(y)) \leq y < \phi(lq_X(p)).$$

b) We use Lemma 5.3.2 to prove this. We need to show (i) and (ii) in that lemma for $\phi(rq_X(p))$. To show (i) note that if $y < \phi(rq_X(p))$,

$$P(\phi(X) \leq y) \leq P(\phi(X) < \phi(rq_X(p))) \leq P(X < rq_X(p)) \leq p.$$

To show (ii), suppose $y > \phi(rq_X(p))$. We only need to show $\phi_\star(y) > rq_X(p)$ because then

$$P(\phi(X) \leq y) \geq P(X < \phi_\star(y)) > p.$$

But by previous lemma $\phi(\phi_\star(y)) \geq y > \phi(rq_X(p))$. Hence $\phi_\star(y) > rq_X(p)$.

■

5.7 Continuity of the left and right quantile functions

Lemma 5.7.1 *(Continuity of quantile functions) Suppose F is a distribution function. Then*
a) lq_F is left continuous.
b) rq_F is right continuous.

Proof
a) Suppose $p_n \uparrow p$ be a strictly increasing sequence in [0,1]. Then since lq_F is increasing, $lq_F(p_n)$ is increasing and hence has a limit we call y. We need to show $y = lq_F(p)$. We show this in two steps:

1. $y \leq lq_F(p)$: Let $A = \{x | F(x) \geq p\}$. Then for any $x \in A$:

$$F(x) \geq p \Rightarrow F(x) \geq p_n \Rightarrow x \geq lq_F(p_n) \Rightarrow x \geq \sup_{n\in\mathbb{N}} lq_F(p_n) \Rightarrow x \geq y.$$

   Hence $lq_F(p) = \inf A \geq y$.

2. $y \geq lq_F(p)$: We only need to show that $F(y) \geq p$. But

$$y \geq lq_F(p_n),\ \forall n \Rightarrow F(y) \geq F(lq_F(p_n)) \geq p_n,\ \ \forall n \Rightarrow F(y) \geq p.$$

b) Take a strictly decreasing sequence $p_n \downarrow p$, we need to show $rq_F(p_n) \to rq(p)$. The limit of $rq_F(p_n)$ exists since rq is non–decreasing. Let $y = \inf_{n\in\mathbb{N}} rq_F(p_n)$. We proceed in two steps:

1. $rq_F(p) \leq y$:

$$rq_F(p) \leq rq_F(p_n),\ \ \forall n \in \mathbb{N} \Rightarrow rq_F(p) \leq \inf_{n\in\mathbb{N}} rq_F(p_n) = y.$$

2. $rq_F(p) \geq y$: Since $rq_F(p) = \sup\{x|F(x) \leq p\}$ by Lemma 5.3.1, we only need to show $z < y \Rightarrow F(z) \leq p$. But if $F(z) > p$ then

$$F(z) > p_n \text{ for some } n \in \mathbb{N} \Rightarrow z \geq rq_F(p_n) \text{ for some } n \in \mathbb{N}.$$

Hence,

$$y > z \geq rq(p_n) \text{ for some } n \in \mathbb{N},$$

which is a contradiction to $y = \inf_{n\in\mathbb{N}} rq(p_n)$.

■

F_X is a function that ranges over $[0, 1]$. Once F hits 1 it will remain one. Similarly before F becomes positive it is always zero. This is the motivation for the following definition.

Definition Suppose F is a distribution function. We define the real domain of F to be $RD(F) = \{x|0 < F(x) < 1\}$.

Lemma 5.7.2 *Suppose F is a distribution function. Then*

$$RD(F) = (rq(0), lq(1)) \quad or \quad RD(F) = [rq(0), lq(1)).$$

Proof We proceed in two steps (a),(b).
(a) $RD(F) \subset [rq(0), lq(1))$:
Note that (a) $\Leftrightarrow [rq(0), lq(1))^c \subset RD(F)^c$, where ^{c} stands for taking the compliment of a set in $\mathbb{R}$. If $x \in [rq(0), lq(1))^c$ then $x < rq(0)$ or $x \geq lq(1)$.
$x < rq(0)$ then $F(x) = 0$ by the definition of $rq(0)$.
$x \geq lq(1)$ then $F(x) \geq F(lq(1)) \geq 1 \Rightarrow F(x) = 1$.
(b) $(rq(0), lq(1)) \subset RD(F)$:
$x > rq(0) \Rightarrow F(x) > 0$. (This is because $rq(0) = \sup\{x|F(x) \leq 0\}$.)
$x < lq(1) \Rightarrow F(x) < 1$. (This is because $lq(1) = \inf\{x|\ F(x) = 1\}$.) ■

Definition For a random variable X with distribution function F, we define the L-quantile and R-quantile functions on $\mathbb{R}$:

$$LQ_F : \mathbb{R} \to \mathbb{R}, \quad LQ_F = lq_F \circ F,$$

$$RQ_F : \mathbb{R} \to \mathbb{R}, \quad RQ_F = rq_F \circ F.$$

Lemma 5.7.3 *(Properties of LQ and RQ)*
a) LQ_F, RQ_F *are non–decreasing.*
b)$LQ_F(x) \leq x \leq RQ_F(x)$.
c) LQ_F, RQ_F *are left continuous and right continuous, respectively.*
d) $lq_F(F(x)) = rq_F(F(x)) \Rightarrow LQ_F(x) = RQ_F(x) = x$.
e) We have the following equalities:

$$LQ_F(v) = \inf\{u|F(u) = F(v)\}, \quad RQ_F(v) = \sup\{u|F(u) = F(v)\}.$$

f) $P(LQ_F(x) < X < RQ_F(x)) = 0$.

Proof
a) This result follows from the fact that lq_F, rq_F and F are non–decreasing.
b) $LQ_F(x) = \inf\{y|F(y) \geq F(x)\}$. Since $x \in \{y|F(y) \geq F(x)\}$, $x \geq LQ_F(x)$.
$RQ_F(x) = \sup\{y|F(y) \leq F(x)\}$. Since $x \in \{y|F(y) \leq F(x)\}$, $RQ_F(x) \geq x$.
c) Suppose $x_n \downarrow x$ is a strictly decreasing sequence, then $F(x_n) \downarrow F(x)$ since F is right continuous. Hence $rq_F(F(x_n)) \downarrow rq_F(F(x))$ since rq_F is right continuous by Lemma 5.7.1.
To prove LQ_F is left continuous, let $x_n \uparrow x$ be a strictly increasing sequence and let $p_n = F(x_n)$. Then since $\{p_n\}$ is an increasing and bounded sequence, $p_n \to p'$. Also let $F(x) = p$. We consider two cases:

1. $p = p'$. In this case $p_n \uparrow p$ is a strictly increasing sequence. Since lq_F is left continuous, $\lim_{n\to\infty} LQ_F(x_n) = \lim_{n\to\infty} lq_F(p_n) = lq_F(p) = LQ_F(x)$.

2. $p' < p$. This means F has a jump at x. By Lemma 5.3.1 j), $LQ_F(x) = lq_F(F(x)) = x$. Let $y = \lim_{n\to\infty} lq_F(F(x_n))$. We claim $y \geq x$. Otherwise since $F(x) = p$ and F has a jump at p, $F(y) < p \Rightarrow F(y) < p_n$, for some $n \in \mathbb{N}$. But $y = \sup_{n\in\mathbb{N}} lq(F(x_n))$. Hence $y \geq lq(F(x_n))$ and $F(y) \geq F(lq(p_n)) \geq p_n > p$ a contradiction. Thus $y = \lim_{n\to\infty} lq_F(F(x_n)) \geq x$.

   Also note that $lq_F(p_n) \leq lq_F(F(x)) = x, \ \forall n \Rightarrow y = \sup_{n\in\mathbb{N}} lq_F(p_n) \leq lq_F(F(x)) = x$. We conclude $y = x$. In other words $y = \lim_{n\to\infty} LQ_F(x_n) = LQ_F(x)$.

d) This result is a straightforward consequence of b).

e) This result follows immediately from the definition of these quantiles.

f) $P(LQ_F(x) < X < RQ_F(x)) = P(lq_F(F(x)) < X < rq_F(F(x))) = 0$, by Lemma 5.3.1. ■

Example Suppose the distribution function F depicted in Figure 5.1 is given as follows

$$F(x) = \begin{cases} \frac{\frac{2}{\pi}\arctan(x)+1}{5} & x \leq 0 \\ 1/5 & 0 \leq x \leq 1 \\ x/5 & 1 \leq x < 2 \\ 3/5 & 2 \leq x < 3 \\ \frac{\frac{2}{\pi}\arctan(x-3)+4}{5} & x \geq 3 \end{cases}.$$

Then $lq_F(0.2) = 0$, $rq_F(0.2) = 1$, $lq_F(0.5) = rq_F(0.5) = 2$ and $lq_F(0.55) = rq_F(0.55) = 2$. We have also plotted lq, rq, LQ, RQ in Figures 5.2 to 5.5.

If we are given a data vector, we can compute the sample distribution and then compute the left and right quantile functions. In the sequel, we show that we get the same definition as we gave for left and right quantile for a vector.

Lemma 5.7.4 *Suppose a data vector x is given and F_n is its sample distribution. Then $lq_x(p) = lq_{F_n}(p)$ and $rq_x(p) = rq_{F_n}(p)$.*

Proof

We show this for non–intersection points. Similar arguments work for intersection points. If p is not an intersection point, then $\frac{m_1+\cdots+m_{i-1}}{n} < p < \frac{m_1+\cdots+m_i}{n}$ and $rq_x(p) = lq_x(p) = z_i$. We want to show that $\inf\{u|F_n(u) \geq p\}$ is also z_i, where

$$F_n(u) = \sum_{i=1}^{n} I_{(-\infty,x_i]}(u).$$

But it follows that:

$$F_n(z_i) = \frac{m_1 + \cdots + m_i}{n};$$

$$lq_{F_n}(p) = \inf\{u|F_n(u) \geq p\};$$

$$rq_{F_n}(p) = \inf\{u|F_n(u) > p\}.$$

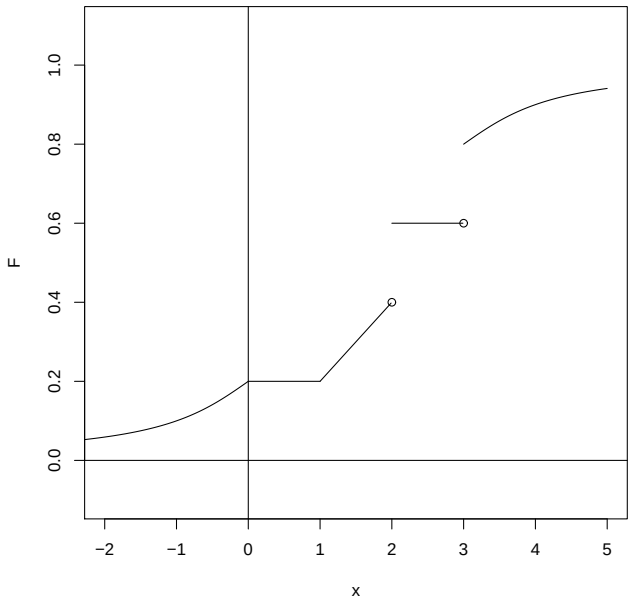

Figure 5.1: An example of a distribution function with discontinuities and flat intervals.

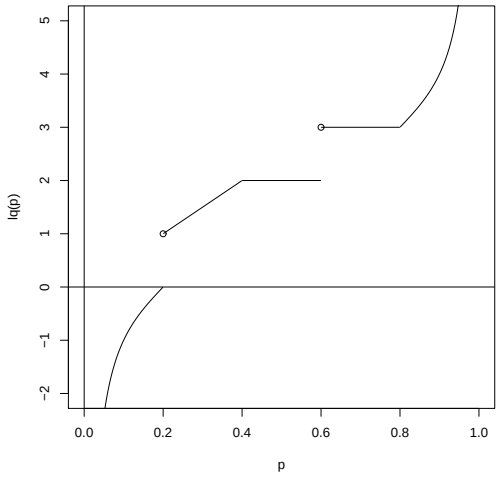

Figure 5.2: The left quantile (lq) function for the distribution function given in Example 5.7. Notice that this function is left continuous and increasing.

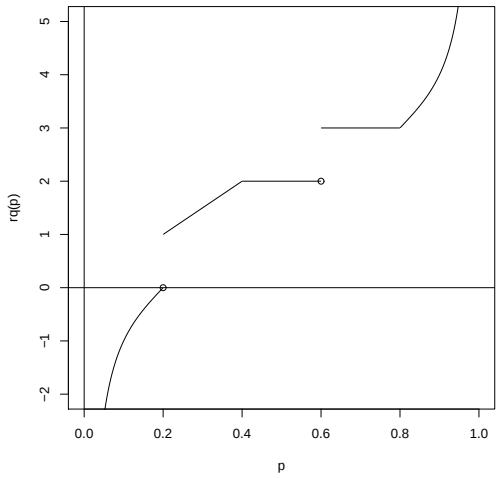

Figure 5.3: The right quantile (rq) function for the distribution function given in Example 5.7. Notice that this function is right continuous and increasing.

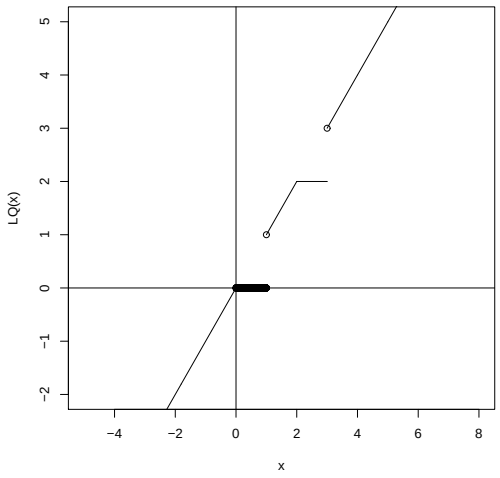

Figure 5.4: LQ function for Example 5.7. Notice that this function is increasing and left continuous.

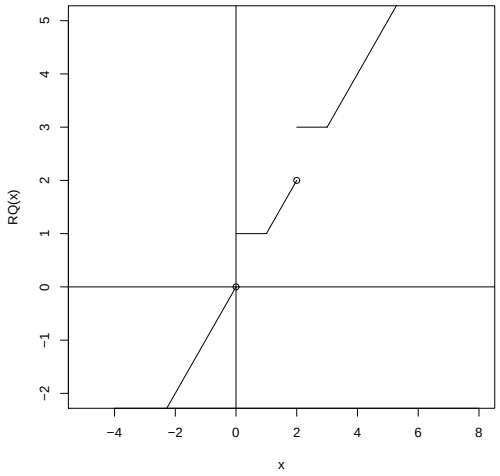

Figure 5.5: RQ function for Example 5.7, notice that this function is increasing and right continuous.

Since F_n is a step function the right hand side of the two above equations can only be one of $-\infty, z_1, \cdots, z_r, \infty$. The first u that makes F_n greater than or equal to p is z_i, proving the assertion.

■

Lemma 5.7.4 guarantees that our definition of quantile for data vectors is consistent with the definition for distributions.

Lemma 5.3.1 shows that if a distribution function F is flat then rq and lq might differ. To study this further when rq and lq are equal, we define the concept of heavy and weightless points in the next section.

5.8 Equality of left and right quantiles

This section finds necessary and sufficient conditions for the left and right quantiles to be equal. We start with some definitions.

Definition Suppose X is a random variable with the distribution function F. $x \in \mathbb{R}$ is called a weightless point of a distribution function F if there exist a neighborhood (an open interval) around x such that F is flat in that neighborhood. We call a point heavy if it is not weightless. Denote the set of all heavy points by H.

Definition A point $x \in \mathbb{R}$ is called a super heavy point if

$$P(X \in (x-\epsilon, x]) > 0, P(X \in [x, x+\epsilon)) > 0, \ \ \forall \epsilon > 0.$$

We denote the set of super heavy points by SH. Obviously any super heavy point is heavy.

We can also define right heavy points and left heavy points.

Definition A point $x \in \mathbb{R}$ is called a right heavy point if

$$P(X \in [x, x+\epsilon)) > 0, \ \ \forall \epsilon > 0.$$

We show the set of all right heavy points by RH. A point $x \in \mathbb{R}$ is called a left heavy point if

$$P(X \in (x-\epsilon, x]) > 0, \ \ \forall \epsilon > 0.$$

We denote the set of all such points by LH. Obviously any heavy point is either right heavy or left heavy. Also a super heavy point is both right heavy and left heavy.

Lemma 5.8.1 *Suppose X is a random variable with distribution function F. Also suppose that $u_1 < u_2$ are heavy points and F is flat on $[u_1, u_2]$ i.e. $F(u_1) = F(u_2)$. Then $lq(p) = u_1$ and $rq(p) = u_2$, where $p = F(u_1) = P(X \leq u_1)$.*

Proof

1. $lq(p) = u_1$: Since $F(u_1) = p$, $lq(p) \leq u_1$. Suppose $lq(p) < u_1$. Then

$$P(lq(p) < X < u_2) > 0,$$

since u_1 is a heavy point. We can rewrite above as

$$P(lq(p) < X \leq u_1) + P(u_1 < X < u_2) > 0,$$

the second term is zero by the flatness assumption. Hence

$$P(lq(p) < X \leq u_1) > 0.$$

But then

$$P(X \leq lq(p)) = p(X \leq u_1) - P(lq(p) < X \leq u_1) < p,$$

which is a contradiction to Lemma 5.3.1 a).

2. $rq(p) = u_2$: From $F(u) = p$ for all $u_1 \leq u < u_2$, we conclude $rq(p) \geq u_2$. To prove the inverse, note that for any $u_1 < u_3 < u_2$, $F(u_3) = p$ since F is flat on $[u_1, u_2]$. Since $rq(p) = \sup\{x|F(x) \leq p\}$ by Lemma 5.3.1, $rq(p) \geq u_2$. Now note that since u_2 is heavy, for any $u_3 > u_2$,

$$P(u_1 < X < u_3) > 0 \Rightarrow F(u_3) = F(u_1) + P(u_1 < X < u_3) > p.$$

Hence only values less than or equal to u_2 are in $\{x|F(x) \leq p\}$. We conclude the sup is at most u_2. In other words $rq(p) \leq u_2$.

■

Lemma 5.8.2 *Suppose X is a random variable with distribution function F. Then*
v is a weightless point $\Leftrightarrow v \in (LQ_F(v), RQ_F(v))$.

Proof ($\Leftarrow$): This is trivial by Lemma 5.7.3 part (f).
($\Rightarrow$): If $v \notin (LQ_F(v), RQ_F(v)) \Rightarrow LQ_F(v) = RQ_F(v) = v$ by Lemma 5.7.3.

$$\begin{aligned} RQ_F(v) = v \Rightarrow \inf\{x|\ F(x) > F(v)\} = v \Rightarrow \\ F(x) > F(v),\ \forall x > v \Rightarrow P(v < X \leq x) > 0,\ \forall x > v \Rightarrow \\ P(v < X < x) > 0, \forall x > v, \end{aligned}$$

where the last ($\Rightarrow$) is because for any $x > v$, we can take $v < x' < x$ and note that $P(v < X < x) \geq P(x < X \leq x') > 0$. We conclude v is a right heavy point which is a contradiction. ■

For a weightless point v, there is an interval (a, b) such that $v \in (a, b)$ and F is flat in that interval. It is useful to consider the flat interval around v. This is the motivation for the following definition.

Definition Suppose X is a random variable with distribution function F and v is a weightless point of F. Then we define the weightless interval of v, $I(v)$ by

$$I(v) = \cup_{a<b, F(a)=F(b)=F(v)}(a, b).$$

Lemma 5.8.3 *Suppose F is a distribution function and v is a weightless point of this distribution function. Then*

$$I(v) = (LQ_F(v), RQ_F(v)).$$

Proof (L.H.S $\subset$ R.H.S): $x \in (a, b)$ for some a, b where $F(a) = F(b) = F(v)$ then $F(x) = F(v)$. Take x_1, x_2 such that $a < x_1 < x < x_2 < b$ then

$$\begin{aligned} F(x_1) = F(x_2) = F(v) \Rightarrow \\ LQ_F(v) \leq x_1 < x < x_2 \leq RQ_F(v) \Rightarrow \\ x \in (LQ_F(v), RQ_F(v)). \end{aligned}$$

(R.H.S $\subset$ L.H.S): This is trivial since if v is weightless then $LQ_F(v) < RQ_F(v)$. Let $a = LQ_F(v)$ and $b = RQ_F(v)$ then $(a, b) \subset I(v)$ by definition of $I(v)$. ■

Corollary 5.8.4 *For any weightless point v, its weightless interval is indeed an open interval.*

Lemma 5.8.5 *Suppose X is a random variable with a distribution function F and v, v' are weightless points then $I(v) = I(v')$ or $I(v) \cap I(v') = \emptyset$.*

Proof Suppose $I(v) \cap I(v') \neq \emptyset$. Fix $u \in I(v) \cap I(v')$. But

$$F(u) = F(v) \Rightarrow$$

$$LQ_F(u) = lq_F(F(u)) = lq_F(F(v)) = LQ_F(v)$$

and

$$RQ_F(u) = rq_F(F(u)) = rq_F(F(v)) = RQ_F(v).$$

Hence by the previous lemma

$$I(u) = (LQ_F(u), RQ_F(u)) = (LQ_F(v), RQ_F(v)) = I(v).$$

A similar argument shows that $I(u) = I(v')$ and this completes the proof. ■

Theorem 5.8.6 *Suppose X is a random variable with distribution function F, then*

a) Let N be the set of all weightless points. Then N is measurable and of probability zero.

b) The ranges of rq_F and lq_F do not intersect N. In other words

$$range(rq_F) \cup range(lq_F) \subset H.$$

c) Any heavy point is either $lq_F(p)$ or $rq_F(p)$ (or both) for some $p \in [0,1]$. In other words

$$H \subset range(rq_F) \cup range(lq_F).$$

More precisely, if x is right heavy then $x \in range(rq_F)$ and if x is left heavy then $x \in range(lq_F)$.

d) $x = lq(p) = rq(p)$ for some $p \in [0,1]$ if and only if x is a super heavy point. Also $H - SH$ is countable.

Proof

a) Suppose v is a weightless point and consider $I(v) = (LQ_F(v), RQ_F(v))$. Then by Lemma 5.7.3, all the points in $I(v)$ are weightless. We showed that $I(v) \cap I(v') \neq \emptyset$, then $I(v) = I(v')$. Hence N can be written as a disjoint union of the form:

$$N = \cup_{v \in N'} I(v),$$

for some $N' \subset N$. Pick a rational number $q_v \in I(v),\ v \in N'$ ("the Axiom of choice" from set theory is not needed to pick a rational number from an interval (a, b) because one can take a rational number by comparing the expansion of a and b in the base 10). But

$$I(v) \cap I(v') = \emptyset,\ v \neq v' \in N' \Rightarrow q_v \neq q_{v'}.$$

This shows N' is countable since the set of rational numbers is countable. Hence, N is a countable union of intervals and is measurable. Moreover,

$$P(N) = P(\cup_{v \in N'} I(v)) = \sum_{v \in N'} P(X \in I(v)) = 0.$$

b) Suppose $z \in N$. Then there exist a, b such that $a < z < b$ and $P(a < X < b) = 0$. Take a', b' such that $a < a' < z < b' < b$. Suppose $z = lq_F(p)$ for some p. Then $P(X \leq z) \geq p$ and also $P(X \leq a') = P(X \leq z) \geq p$. This is a contradiction since z is the left quantile. Similarly, suppose $z = rq_F(p)$ for some p. Then since $z < b'$, $F(b') > p$ while $a' < z$ gives $F(a') \leq p$. Hence $P(a' \leq X \leq b') > 0$, a contradiction.
c) Assume x is right heavy. Then let $p = F(x)$. We claim that $rq_F(p) = x$. Suppose $rq_F(p) = x' < x$ then $F(x) = p$ is a contradiction to $rq_F(p) = \sup\{y | F(y) \leq p\}$. On the other hand for any $x' > x$, pick $x < x'' < x'$. We have $F(x'') > p$ since x is right heavy. Since $rq_F(p) = \inf\{y | F(y) > p\}$ and $F(x'') > p$ then $x' > rq_F(p)$. We conclude that $rq_F(p) = x$.

Now suppose x is left heavy. Let $p = F(x)$. We claim $lq_F(p) = x$. First note that for any $x' < x$, $F(x') < F(x) = p$ since x is left heavy. Hence $lq_F(p) \geq x$. But $F(x) = p$ and since $lq_F(p) = \inf\{y | F(y) \geq p\}$ we are done.
d) The necessary and sufficient conditions follow immediately from c). To show that $H - SH$ is countable, we prove $LH - SH$ and $RH - SH$ are countable. To that end, for any $x \in LH - SH$ consider $I_x = (LQ(x), RQ(x))$. Since x is not super heavy this interval has positive length. Also note that $x < y, x, y \in H$ implies $I_x \cap I_y = \emptyset$. To prove this, note that since x, y are left heavy, $LQ(x) = x$ and $LQ(y) = y$. We conclude

$$I_x = (x, RQ(x))$$

$$I_y = (y, RQ(y)).$$

If $I_x \cap I_y$ is nonempty then we conclude $x < y < RQ(x)$. Then

$$0 = P(X \in (x, RQ(x))) \leq P(X \in (x, y)) > 0.$$

$(P(X \in (x,y)) > 0$ since y is left heavy.) This is a contradiction and hence $I_x \cap I_y = \emptyset$. Now pick a rational number $q_x \in I_x$. Then

$$I_x \cap I_y = \emptyset \Rightarrow q_x \neq q_y.$$

Since the set of rational numbers is countable $LH - SH$ is countable. A similar argument works for $RH - SH$. ■

Lemma 5.8.7 *Suppose X is a random variable with distribution function F. Then the set $A = \{p|\ p \in [0,1],\ lq_F(p) \neq rq_F(p)\}$ is countable.*

Proof For every $p \in A$ let $J(p) = (lq_F(p), rq_F(p))$. Then for every $x \in J(p)$, $F(x) = p$. ($F(x) \geq F(lq_F(p)) \geq p$. Now if $F(x) > p$, we get a contradiction to $x < lq_X(p)$.) We conclude

$$p, p' \in A, p \neq p' \Rightarrow J(p) \cap J(p') = \emptyset.$$

The intervals are disjoint, every interval has a positive length and their union is a subset of $[0,1]$. Hence there are only countable number of such intervals. We conclude A is countable. ■

The following lemma gives sufficient and necessary conditions for $lq_X = rq_X,\ \forall p \in (0,1)$.

Lemma 5.8.8 *$lq_X(p) = rq_X(p),\ p \in (0,1)$ iff F_X is strictly increasing.*

Proof ($\Rightarrow$)

$$\begin{aligned} lq_X(p) = \inf\{x|F_X(x) \geq p\} = \\ \inf\{x|x \geq F_X^{-1}(p)\} = \\ \inf\{x|x > F_X^{-1}(p)\} = rq_X(p) \end{aligned} \quad .$$

($\Leftarrow$): If F_x is not strictly increasing then $\exists x_2 < x_1$ s.t $F_X(x_1) = F_X(x_2)$. Then let $p = F_X(x_1)$. We also have $p = F_X(x_2)$. Hence

$$lq_X(p) = \inf\{F_X(x) \geq p\} \leq x_1,$$

and

$$rq_X(p) = \sup\{F_X(x) \leq p\} \geq x_2,$$

which is a contradiction. ■

5.9 Distribution function in terms of the quantile functions

It is interesting to understand the connections amongst lq, rq and F. We answer the following question:
Question: Given one of lq, rq or F, are the other two uniquely determined? The answer to this question is affirmative and the following theorem says much more.

Theorem 5.9.1 *Suppose F is a distribution function. Then*

a) For $p_0 \in (0,1)$, $lq(p_0) = \lim_{p \to p_0^-} rq(p_0)$. Hence, the function rq uniquely determines lq.

b) For $p_0 \in (0,1)$, $rq(p_0) = \lim_{p \to p_0^+} lq(p_0)$. Hence lq uniquely determines rq.

c) lq or rq continuous at $p_0 \in (0,1) \Rightarrow lq(p_0) = rq(p_0)$.

d) $lq(p_0) = rq(p_0) \Rightarrow lq$ and rq are continuous at p_0.

e) lq is continuous at $p \Leftrightarrow rq$ is continuous at p.

f) $F(x) = \inf\{p | lq(p) > x\}$.

g) $F(x) = \inf\{p | rq(p) > x\}$.

Proof

a) Take a strictly increasing sequence $p_n \uparrow p_0$ in $[0,1]$. Then

$$p_{n-1} < p_n < p_{n+1} \Rightarrow$$

$$lq(p_{n-1}) < rq(p_n) < lq(p_{n+1}), \tag{5.4}$$

by Lemma 5.3.1, part (c). By the left continuity of lq, $lq(p_n) \to lq(p_0)$. Applying the Sandwich Theorem about the limits from elementary calculus to the Equation (5.4), we conclude that $rq(p_n) \to lq(p_0)$.

b) Take a strictly decreasing sequence $p_n \downarrow p_0$ in $[0,1]$. Then

$$p_{n-1} > p_n > p_{n+1} \Rightarrow$$

$$rq(p_{n-1}) > lq(p_n) > rq(p_{n+1}), \tag{5.5}$$

again by Lemma 5.3.1, part (c). By the right continuity of rq, $rq(p_n) \to rq(p_0)$. Applying the Sandwich Theorem for limits to Equation (5.5), we conclude that $lq(p_n) \to rq(p_0)$.

c) Suppose lq is continuous at p_0. Then $\lim_{p \to p_0^+} lq(p) = lq(p_0)$. But by the previous parts of this theorem, we also have $\lim_{p \to p_0^+} = rq(p_0)$. Similar arguments work if rq is continuous at p_0.

d) To prove lq is continuous at p_0 note that

$$\lim_{p \to p_0^-} lq(p_0) = lq(p_0) = rq(p_0) = \lim_{p \to p_0^+} lq(p_0),$$

where the first equality comes from the left continuity of lq and the last one comes from (b). Similar arguments work for rq.

e) This result follows immediately from the previous two parts.

f) Let $A = \{p | lq(p) > x\}$. We want to show that $F(x) = \inf A$.
To do that we first show that $F(x) \leq \inf A$.
By Lemma 5.7.3,

$$lq(F(x)) \leq x \Rightarrow F(x) \leq a, \ \forall a \in A \Rightarrow F(x) \leq \inf A.$$

It remains to show that $\inf A \leq F(x)$. Suppose to the contrary that $F(x) < \inf A$. Then take $F(x) < p_0 < \inf A$ to get

$$lq(p_0) \leq x, p_0 > F(x)$$
$$\Rightarrow F(lq(p_0)) \leq F(x), p_0 > F(x).$$

But by Lemma 5.3.1 part (a), $p_0 \leq F(lq(p_0))$. Hence

$$p_0 \leq F(lq(p_0)) \leq F(x), p_0 > F(x),$$

which is a contradiction.

g) Let $B = \{p|rq(p) > x\}$ and A be as the previous part. Then $F(x) = \inf A \leq \inf B$.
It only remains to show that $\inf B \leq F(x)$. Otherwise, we can pick p_0, $F(x) < p_0 < \inf B$ so that

$$rq(p_0) \leq x, p_0 > F(x) \Rightarrow$$

$$p_0 \leq F(rq(p_0)) \leq F(x), p_0 > F(x),$$

which is a contradiction.

■

5.10 Two-sided continuity of lq/rq

Lemma 5.10.1 *Suppose F is a distribution function for the random variable X and lq, rq are its corresponding left and right quantile functions. Then*
a) F is continuous $\Leftrightarrow$ lq is strictly increasing on $(0,1)$.
b) F is strictly increasing on $RD(F) = \{x|0 < F(x) < 1\} = (rq(0), lq(1))$ or $[rq(0), lq(1)) \Leftrightarrow$ lq is continuous on $(0,1)$.

Proof a)
($\Rightarrow$): F is continuous iff $P(X = x) = 0$, $\forall x \in \mathbb{R}$. If the R.H.S does not hold then $x = lq(p_1) = lq(p_2)$, $p_1 < p_2$. Then for every $y < x$, we have $F(y) < p_1$. Hence

$$P(X < x) = \lim_{y \to x^-} P(X \leq y) \leq p_1 < p_2.$$

But $F(x) \geq p_2$ since $lq(p_2) = x$ and we conclude $P(X = x) \geq p_2 - p_1$, a contradiction.
($\Leftarrow$): If F is not continuous then $P(X = x) = \epsilon > 0$ for some $x \in \mathbb{R}$. Let $p = F(x)$ then $P(X < x) = p - \epsilon$. Pick $p_1 < p_2$ in the interval $(p - \epsilon, p)$ then $lq(p_1) = lq(p_2) = x$.
b)
($\Rightarrow$): lq is left continuous. Hence if it is not continuous then

$$\lim_{p \to p_0^+} lq(p) = rq(p_0) \neq lq(p_0).$$

Hence F is flat on $(lq(p_0), rq(p_0)) \neq \emptyset$, which is a contradiction to F being increasing.

($\Leftarrow$): Suppose F is not continuous on $RD(F)$, then there exist $a, b \in \mathbb{R}$ such that F is flat on $[a, b]$:

$$F(a) = F(b) = p \in (0, 1).$$

But then $lq(p) \leq a$ and $rq(p) \geq b$. Hence $lq(p) \neq rq(p)$, which means lq is not continuous. ■

Remark. We can replace lq is the above lemma by rq. A similar argument can be done for the proof.

5.11 Characterization of left/right quantile functions

The characterization of the distribution function is a well–known result in probability. Here we characterize the left and right quantile functions of a distribution. We start by some simple lemmas which we need in the proof.

Lemma 5.11.1 *Suppose $A_n \subset \mathbb{R}, n \in N$. Then*

$$\inf \cup_{n\in\mathbb{N}} A_n = \inf_{n\in\mathbb{N}} (\inf A_n)$$

Proof

a) $\inf \cup_{n\in\mathbb{N}} A_n \geq \inf_{n\in\mathbb{N}} (\inf A_n)$:

$$a \in \cup_{n\in\mathbb{N}} A_n \Rightarrow \exists m \in \mathbb{N},\ a \in A_m \Rightarrow \exists m \in \mathbb{N},\ a \geq \inf A_m \Rightarrow a \geq \inf_{n\in\mathbb{N}} (\inf A_n).$$

Hence, $\inf \cup_{n\in\mathbb{N}} A_n \geq \inf_{n\in\mathbb{N}} (\inf A_n)$.

b) $\inf \cup_{n\in\mathbb{N}} A_n \leq \inf_{n\in\mathbb{N}} (\inf A_n)$:

$$\inf \cup_{n\in\mathbb{N}} A_n \leq \inf A_m,\ \forall m \in \mathbb{N} \Rightarrow \inf \cup_{n\in\mathbb{N}} A_n \leq \inf_{n\in\mathbb{N}} (\inf A_n).$$

■

Lemma 5.11.2 *Suppose $h : (0, 1) \to \mathbb{R}$ is a non–decreasing function. Then $G(x) = \inf\{p \in (0, 1) | h(p) > x\}$ is a distribution function.*

Proof a) We claim G is non–decreasing. Suppose $x_1 < x_2$ then let $A = \{p | h(p) > x_1\}$ and $B = \{p | h(p) > x_1\}$. Then $G(x_1) = \inf A$ and $G(x_2) = \inf B$. But clearly $B \subset A$ hence $G(x_1) \leq G(x_2)$.

b) $\lim_{x\to\infty} G(x) = 1$: First note that such a limit exist and is bounded by 1. (Because the domain of h is (0,1)). Assume $\lim_{x\to\infty} G(x) = q < 1$, take $q < q' < 1$ then take $x_0 > h(q')$. Let $A = \inf\{p|h(p) > x_0\}$ such that $G(x_0) = \inf A$. Then

$$(p \in A \Rightarrow h(p) > x_0 > h(q') \Rightarrow p > q') \Rightarrow G(x_0) = \inf A \geq q' > q.$$

We have shown there is an x_0 such that $G(x_0) > q$ this is a contradiction to $\lim_{x\to\infty} G(x) = q$ since G is non-decreasing.
c) Suppose that $\lim_{x\to-\infty} G(x) = q > 0$ then take $0 < q' < q$ and $x_0 < h(q')$. Let $A = \inf\{p|h(p) > x_0\}$ such that $G(x_0) = \inf A$. We have

$$h(q') > x_0 \Rightarrow q' \in A \Rightarrow \inf A \leq q' \Rightarrow G(x_0) \leq q' < q$$

This contradicts $\lim_{x\to-\infty} G(x) = q > 0$ since G is non–decreasing.
d) G is right continuous: $\lim_{x\to x_0^+} G(x) = x_0$. Suppose $x_n \downarrow x_0$. In the previous lemma, let $A_n = \{p|h(p) > x_n\}$ and $A = \cup_{n\in\mathbb{N}} A_n = \{p|h(p) > x_0\}$. Then

$$G(x_0) = \inf A = \inf \cup_{n\in\mathbb{N}} A_n = \inf_{n\in\mathbb{N}} (\inf A_n) = \inf_{n\in\mathbb{N}} G(x_n) = \lim_{x\to x_0^+} G(x).$$

■

Theorem 5.11.3 *(Quantile function characterization theorem) Suppose a function $h : (0,1) \to \mathbb{R}$ is given. Then*
(a) h is a left quantile function for some random variable X iff h is left continuous and non–decreasing.
(b) h is a right quantile function for some random variable X iff h is right continuous and non–decreasing.

Proof If h is a left quantile function, then h is left continuous and non–decreasing as we showed in previous sections. Also if h is right continuous function then h is non–decreasing and right continuous. For the inverse of both a) and b) define G as in the above lemma. We will prove that h is lq_G in a) and rq_G in b).

(a) Let $A = \{x|G(x) \geq p_0\}$, we want to show $h(p_0) = \inf A$.

(i) $\inf A \leq h(p_0)$: Otherwise if $\inf A > y > h(p_0)$, then:

$$\begin{aligned}\inf A > y \Rightarrow \inf\{x|G(x) \geq p_0\} > y \Rightarrow G(y) < p_0 \Rightarrow \\ \inf\{p \in (0,1)|h(p) > y\} < p_0 \Rightarrow \\ \exists p \in (0,1),\ h(p) > y,\ p < p_0 \Rightarrow \\ \exists p \in (0,1) h(p_0) \geq h(p) > y,\end{aligned}$$

which is a contradiction.
(ii) $\inf A \geq h(p_0)$:

$$x \in A \Rightarrow G(x) \geq p_0 \Rightarrow \inf\{p \in (0,1)|h(p) > x\} \geq p_0.$$

Hence,

$$\forall p < p_0,\ h(p) \leq x \Rightarrow \lim_{p \to p_0^-} h(p) \leq x \Rightarrow h(p_0) \leq x,$$

by left continuity of h. Hence

$$\forall x \in A,\ h(p_0) \leq x \Rightarrow h(p_0) \leq \inf A.$$

(b) Let $A = \{x|G(x) > p_0\}$, we want to show $h(p_0) = \inf A$.
(i) $\inf A \leq h(p_0)$: Otherwise if $\inf A > y > h(p_0)$, then

$$\begin{aligned}\inf A > y \Rightarrow y \notin A \Rightarrow G(y) \leq p_0 \Rightarrow \\ \inf\{p' \in (0,1)|h(p') > y\} \leq p_0 \Rightarrow \\ \forall p > p_0, \inf\{p'|h(p') > y\} < p \Rightarrow \\ \forall p > p_0,\ \exists p' \in (0,1),\ h(p') > y, p' < p \Rightarrow \\ \forall p > p_0,\ \exists p' \in (0,1),\ h(p) \geq h(p') > y \Rightarrow \\ h(p_0) \geq y\end{aligned}$$

which is a contradiction.
(ii) $\inf A \geq h(p_0)$:

$$x \in A \Rightarrow G(x) > p_0 \Rightarrow \inf\{p \in (0,1)|h(p) > x\} > p_0 \Rightarrow$$

$$p_0 \notin \{p \in (0,1)|h(p) > x\} \Rightarrow h(p_0) \leq x.$$

Hence $h(p_0) \leq \inf A$. ■

Now we characterize the quantile functions of data vectors. See Figure 5.6 for an example of quantile functions for the vector

$$x = (-2, -2, 2, 2, 2, 2, 4, 4, 4, 4).$$

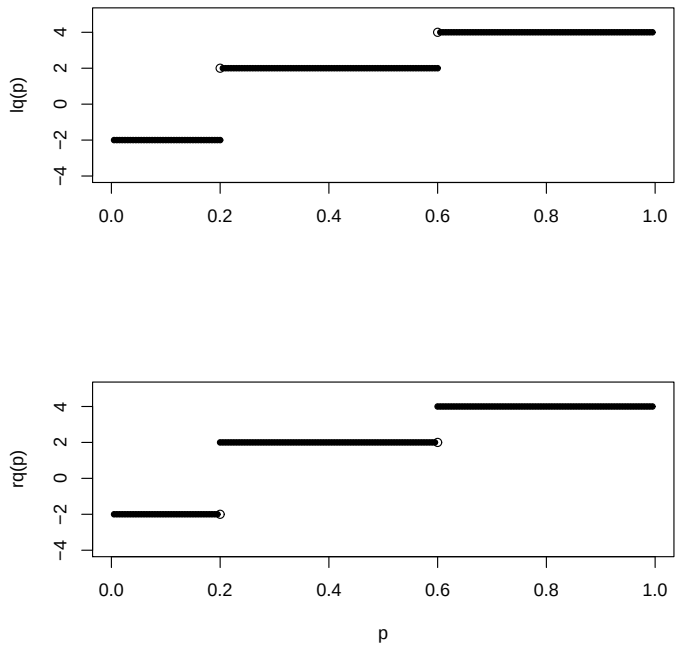

Figure 5.6: For the vector $x = (-2, -2, 2, 2, 4, 4, 4, 4)$ the left (top) and right (bottom) quantile functions are given.

Theorem 5.11.4 *(Data vector quantile function characterization theorem) a) $h : (0,1) \to \mathbb{R}$ is a left quantile function for a data vector x iff h is a left continuous step function with no steps (jumps) or a finite number of steps (jumps) at some points $0 < a_1 < a_2 < \cdots < a_k < 1$ where $a_i = \frac{1}{n} n_i$, for some $n, n_i \in \mathbb{N}$.*
b) $h : (0,1) \to \mathbb{R}$ is a right quantile function for a data vector x iff h is a right continuous step function with no steps (jumps) or finite number of steps (jumps) at some points $0 < a_1 < a_2 < \cdots < a_k < 1$ where $a_i = \frac{1}{n} n_i$ for some $n, n_i \in \mathbb{N}$.

Proof

We only prove a) and b) is obtained either by repeating a similar argument or using the Quantiles Symmetry Theorem (Theorem 5.12.3), which we prove in next sections.

a) ($\Rightarrow$) For $x = (x_1, \cdots, x_n)$, it is clear that lq_x is a step function with jumps at points proportional to $1/n$ and we proved the left continuity before.

a) ($\Leftarrow$) The result is easy to show if h has no jumps. Let $h' = \lim_{x \to +\infty} h(x)$ and suppose h is given with jumps at $a_1 < a_2 < \cdots < a_k, a_1 = n_1(1/n), \cdots, a_k = n_k(1/n)$. Let $b_1 = a_1, b_2 = a_2 - a_1, \cdots, b_k = a_k - a_{k-1}, b_{k+1} = 1 - a_k$. Then $b_i = \frac{1}{n} m_i, i = 1, 2, \cdots, k+1$ with $m_1 = n_1, m_2 = n_2 - n_1, \cdots, m_k = n_k - n_{k-1}$ and finally $m_{k+1} = n - \sum_{i=1}^{k} m_i$. Then let x be a data vector with $h(a_i)$ repeated m_i times. We claim that $h = lq_x$. First note that x is of length n. For $0 < p \le a_1$, we have $lq_x(p) = h(a_1) = h(p)$. For $a_{i-1} < p \le a_i, i \le k$, we have $\frac{n_{i-1}}{n} = \frac{\sum_{j=1}^{i-1} m_j}{n} < p \le \frac{\sum_{j=1}^{i} m_j}{n} = \frac{n_i}{n}$. Hence

$$lq_x(p) = h(a_i) = h(p), \quad a_{i-1} < p \le a_i, i \le k.$$

For $a_k < p < 1$, we have $\frac{n_k}{n} = \frac{\sum_{j=1}^{k} m_j}{n} < p < 1$,

$$lq_x(p) = h' = h(p), \quad a_k < p < 1.$$

■

5.12 Quantile symmetries

This section studies the symmetry properties of distribution functions and quantile functions. Symmetry is in the sense that if X is a random variable with left/right quantile function, some sort of symmetry between the

quantile functions of X and $-X$ should exist. We only treat the quantile functions for distributions here but the results can readily be applied to data vectors by considering their empirical distribution functions.

Here consider different forms of distribution functions. The usual one is defined to be $F_X^c(x) = P(X \leq x)$. But clearly one could have also considered $F_X^o(x) = P(X < x)$, $G_X^c(x) = P(X \geq x)$ or $G_X^o(x) = P(X > x)$ to characterize the distribution of a random variable. We call F^c the left–closed distribution function, F^o the left–open distribution function, G^c the right–closed and G^o the right–open distribution function. Like the usual distribution function these functions can be characterized by their limits in infinity, monotonicity and right continuity.

First note that

$$F_{-X}^c(x) = P(-X \leq x) = P(X \geq -x) = G_X^c(-x).$$

Since the left hand side is right continuous, G_X^c is left continuous. Also note that

$$F_X^c(x) + G_X^o(x) = 1 \Rightarrow G_X^o(x) = 1 - F_X^c(x),$$
$$F_X^o(x) + G_X^c(x) = 1 \Rightarrow F_X^o(x) = 1 - G_X^c(x).$$

The above equations imply the following:
a) G^o and F^c are right continuous.
b) F^o and G^c are left continuous.
c) G^o and G^c are non–decreasing.
d) $lim_{x\to\infty}F(x) = 1$ and $lim_{x\to-\infty}F(x) = 0$ for $F = F^o, F^c$.
e) $lim_{x\to\infty}G(x) = 0$ and $lim_{x\to-\infty}G(x) = 1$ for $G = G^o, G^c$.

It is easy to see that the above given properties for F^o, G^o, G^c characterize all such functions. The proof can be given directly using the properties of the probability measure (such as continuity) or by using arguments similar to the above.

Another lemma about the relation of F^c, F^o, G^o, G^c is given below.

Lemma 5.12.1 *Suppose F^o, F^c, G^o, G^c are defined as above. Then*
a) if any of F^c, F^o, G^o, G^c are continuous, all of other are continuous too.
b) F^c being strictly increasing is equivalent to F^o being strictly increasing.
c) if F^c is strictly increasing, G^o is strictly decreasing.
d) G^c being strictly decreasing is equivalent to G^o being strictly increasing.

Proof a) Note that $\lim_{y\to x^-} F^c(x) = \lim_{y\to x^-} F^o(x)$ and $\lim_{y\to x^+} F^c(x) = \lim_{y\to x^+} F^o(x)$. If these two limits are equal for either F^c or F^o they are equal for the others as well.
b) If either F^c or F^o are not strictly increasing then they are constant on $[x_1, x_2], x_1 < x_2$. Take $x_1 < y_1 < y_2 < x_2$. Then

$$F^o(x_1) = F^o(x_2) \Rightarrow P(y_1 \leq X \leq y_2) = 0 \Rightarrow F^c(y_1) = F^c(y_2).$$

Also we have

$$F^c(x_1) = F^c(x_2) \Rightarrow P(y_1 \leq X \leq y_2) = 0 \Rightarrow F^o(y_1) = F^o(y_2).$$

c) This is trivial since $G^o = 1 - F^c$.
d) If G^c is strictly decreasing then F^o is strictly increasing since $G^c = 1 - F^o$. By part b), F^c strictly is increasing. Hence $G^o = 1 - F^c$ is strictly decreasing. ■

The relationship between these distribution functions and the quantile functions are interesting and have interesting implications. It turns out that we can replace F^c by F^o in some definitions.

Lemma 5.12.2 *Suppose X is a random variable with open and closed left distributions F^o, F^c as well as open and closed right distributions G^o, G^c. Then*
a) $lq_X(p) = \inf\{x | F^o_X(x) \geq p\}$. In other words, we can replace F^c by F^o in the left quantile definition.
b) $rq_X(p) = \inf\{x | F^o_X(x) > p\}$. In other words, we can replace F^c by F^o in the right quantile definition.

Proof a) Let $A = \{x | F^o_X(x) \geq p\}$ and $B = \{x | F^c_X(x) \geq p\}$. We want to show that $\inf A = \inf B$. Now

$$A \subset B \Rightarrow \inf A \geq \inf B.$$

But

$$\inf B < \inf A \Rightarrow \exists x_0, y_0, \;\; \inf B < x_0 < y_0 < \inf A.$$

Then

$$\inf B < x_0 \Rightarrow \exists b \in B,\ b < x_0 \Rightarrow \exists b \in \mathbb{R},\ p \leq P(X \leq b) \leq P(X \leq x_0)$$
$$\Rightarrow P(X \leq x_0) \geq p \Rightarrow P(X < y_0) \geq p.$$

On the other hand

$$y_0 < \inf A \Rightarrow y_0 \notin A \Rightarrow P(X < y_0) < p,$$

which is a contradiction, thus proving a).
b) Let $A = \{x | F_X^o(x) > p\}$ and $B = \{x | F_X^c(x) > p\}$. We want to show $\inf A = \inf B$. Again,

$$A \subset B \Rightarrow \inf A \geq \inf B.$$

But

$$\inf B < \inf A \Rightarrow \exists x_0, y_0,\ \inf B < x_0 < y_0 < \inf A.$$

Then

$$\inf B < x_0 \Rightarrow \exists b \in B,\ b < x_0 \Rightarrow \exists b \in \mathbb{R},\ p < P(X \leq b) \leq P(X \leq x_0)$$
$$\Rightarrow P(X \leq x_0) > p \Rightarrow P(X < y_0) > p.$$

On the other hand,

$$y_0 < \inf A \Rightarrow y_0 \notin A \Rightarrow P(X < y_0) \leq p,$$

which is a contradiction. ■

Using the above results, we establish the main theorem of this section which states the symmetry property of the left and right quantiles.

Theorem 5.12.3 *(Quantile Symmetry Theorem) Suppose X is a random variable and $p \in [0, 1]$. Then*

$$lq_X(p) = -rq_{-X}(1 - p).$$

Remark. We immediately conclude

$$rq_X(p) = -lq_{-X}(1 - p),$$

by replacing X by $-X$ and p by $1 - p$.

Proof

$$\begin{aligned} R.H.S = -\sup\{x|P(-X \leq x) \leq 1-p\} = \\ \inf\{-x|P(X \geq -x) \leq 1-p\} = \\ \inf\{x|P(X \geq x) \leq 1-p\} = \\ \inf\{x|1-P(X \geq x) \geq p\} = \\ \inf\{x|1-G^c(x) \geq p\} = \\ \inf\{x|F^o(x) \geq p\} = lq_X(p). \end{aligned}$$

■

Now we show how these symmetries can become useful to derive other relationships/definitions for quantiles.

Lemma 5.12.4 *Suppose X is a random variable with distribution function F. Then*

$$lq_X(p) = \sup\{x|F^c(x) < p\}.$$

Proof

$$\begin{aligned} lq_X(p) = -rq_{-X}(1-p) = -\inf\{x|F^o_{-X}(x) > 1-p\} = \\ -\inf\{x|1-G^c_{-X}(x) > 1-p\} = \sup\{-x|G^c_{-X}(x) < p\} = \\ \sup\{-x|P(-X \geq x) < p\} = \sup\{x|P(X \leq x) < p\} = \\ \sup\{x|F^c(x) < p\}. \end{aligned}$$

■

In the previous sections, we showed that both lq_X and rq_X are equivariant under non-decreasing continuous transformations:

$$lq_{\phi(X)}(p) = \phi(lq_X(p)),$$

where ϕ is non-decreasing left continuous. Also

$$rq_{\phi(X)}(p) = \phi(rq_X(p)),$$

for $\phi : \mathbb{R} \to \mathbb{R}$ non-decreasing right continuous. However, we did not provide any results for decreasing transformations. Now we are ready to offer a result for this case.

Theorem 5.12.5 *(Decreasing transformation equivariance)*
a) Suppose ϕ is non-increasing and right continuous on $\mathbb{R}$. Then

$$lq_{\phi(X)}(p) = \phi(rq_X(1-p)).$$

b) Suppose ϕ is non-increasing and left continuous on $\mathbb{R}$. Then

$$rq_{\phi(X)}(p) = \phi(lq_X(1-p)).$$

Proof a) By the Quantile Symmetry Theorem, we have

$$lq_{\phi(X)}(p) = -rq_{-\phi(X)}(1-p).$$

But $-\phi$ is non-decreasing right continuous, hence the above is equivalent to

$$-(-\phi(rq_X(1-p))) = \phi(rq_X(1-p)).$$

b) By the Quantile symmetry Theorem

$$rq_{\phi(X)}(p) = -lq_{-\phi(X)(1-p)} = - - \phi(lq_X(1-p)) = \phi(lq_X(p)),$$

since $-\phi$ is non-decreasing and left continuous. ■

Lemma 5.12.6 *Suppose X is a random variable and F^c, F^o, G^c, G^o are the corresponding distribution functions. Then we have the following inequalities:*

a) $F^c(lq(p)) \geq p$. (Hence $F^c(rq(p)) \geq p$.)

b) $F^o(rq(p)) \leq p$. (Hence $F^o(lq(p)) \leq p$.)

c) $G^o(lq(p)) \leq 1-p$. (Hence $G^o(rq(p)) \leq 1-p$.)

d) $G^c(rq(p)) \geq 1-p$. (Hence $G^c(lq(p)) \geq 1-p$.)

Proof We already showed a).
b) Suppose there $F^o(rq(p)) = p+\epsilon$ for some positive ϵ. Then since F^o is left continuous

$$\lim_{x \to rq(p)^+} F^o(x) = p + \epsilon.$$

Hence there exist $x_0 < rq(p)$ such that $F(x_0) \geq F^o(x_0) > p + \epsilon/2$. This is a contradiction to $rq(p)$ being the inf of the set $\{x | F(x) > p\}$.
c) and d) are straightforward consequence of a) and b) since $F^c + G^o = 1$ and $F^o + G^c = 1$. ■

The quantile functions as the inverse of an open distribution function

Lemma 5.12.7 *Suppose X is a random variable with distribution function F and open distribution function F^o.*

a) $\{x|F^o(x) < p\} = (-\infty, lq_F(p))$ *or* $(-\infty, lq_F(p)]$.

b) $\{x|F^o(x) \leq p\} = (-\infty, rq_F(p)]$.

c) If F^o is continuous then $\{x|F^o(x) < p\} = (-\infty, lq_F(p)]$.

d) $\{x|F^o(x) > p\} = (rq_F(p), \infty)$.

e) $\{x|F^o(x) \geq p\} = (lq_F(p), \infty)$ *or* $[lq_F(p), \infty)$

Proof The proof is very similar to Lemma 5.5.1 and we skip the details. ■

5.13 Quantiles from the right

So far, we have defined left/right quantiles using the classic distribution function F^c. We also showed that in quantile definitions F^c can be replaced by F^o. $F^c_X(x) = P(-\infty < X \leq x)$ measures the probability from minus infinity. When we define left/right quantiles, we seek to find points where this probability from minus infinity reaches (passes) a certain value. One could also consider $G^c_X(x) = P(x \leq X < \infty)$ and define another version of quantile functions which seek points where the probability from plus infinity reaches or passes a point. This is a motivation to define the "left/right quantile functions from the right". By indicating from the right we clarify that the probability is compute from the right hand side i.e. plus infinity. The previously defined left and right quantile functions should be called "left/right quantile functions from the left".

Definition Suppose X is a random variable with closed right distribution function $G^c_X(x) = P(X \geq x)$. Then we define the "left quantile function from the right" as follows

$$lqfr_X(p) = \sup\{x|G^c_X(x) > p\}.$$

Definition Suppose X is a random variable with closed right distribution function $G^c_X(x) = P(X \geq x)$. Then we define the right quantile function from the right as follows

$$rqfr_X(p) = \sup\{x|G_X^c(x) \geq p\}.$$

Using the symmetries in the definition of these quantities, we will show that we have already characterized left/right from the right quantile functions. We need the following lemma.

Lemma 5.13.1 *Suppose X is a random variable with quantile functions lq_X, rq_X. Then*
a) $rq_X(p) = \sup\{x|F^o(x) \leq p\}$.
b) $lq_X(p) = \sup\{x|F^o(x) < p\}$

Proof a) Let $A = \{x|F^c(x) \leq p\}$ and $B = \{x|F^o(x) \leq p\}$. First note that

$$A \subset B \Rightarrow \sup A \leq \sup B.$$

To show that the sups are indeed equal, note

$$\sup A < \sup B \Rightarrow \exists x_0, y_0,\ \sup A < x_0 < y_0 < \sup B.$$

Then

$$\sup A < x_0 \Rightarrow F^c(x_0) > p,$$

and

$$y_0 < \sup B \Rightarrow \exists b \in B,\ y_0 < b \Rightarrow \exists b,\ F^o(b) \leq p, y_0 < b \Rightarrow F^o(y_0) \leq p.$$

But

$$F^c(x_0) > p, F^o(y_0) \leq p,$$

which is a contradiction.
b) Let $A = \{x|F^c(x) < p\}$ and $B = \{x|F^o(x) < p\}$. First note that

$$A \subset B \Rightarrow \sup A \leq \sup B.$$

To show that the sups are indeed equal, note

$$\sup A < \sup B \Rightarrow \exists x_0, y_0,\ \sup A < x_0 < y_0 < \sup B.$$

Then

$$\sup A < x_0 \Rightarrow F^c(x_0) \geq p,$$

and

$$y_0 < \sup B \Rightarrow \exists b \in B,\ y_0 < b \Rightarrow \exists b,\ F^o(b) < p, y_0 < b \Rightarrow F^o(y_0) < p.$$

But

$$F^c(x_0) \geq p, F^o(y_0) < p,$$

which is a contradiction.

■

Lemma 5.13.2 *(Quantile functions from the right)*
a) $lqfr_X(p) = rq_X(1-p)$.
b) $rqrf_X(p) = lq_X(1-p)$.

Proof

a)

$$lqrf_X(p) = \sup\{x|G_X^c(x) > p\} = \sup\{x|F_X^o(x) \leq p\} = rq_X(1-p).$$

b)

$$rqrf_X(p) = \sup\{x|G_X^c(x) \geq p\} = \sup\{x|F_X^o(x) < 1-p\} = lq_X(1-p).$$

■

5.14 Limit theory

To prove limit results, we need some limit theorems from probability theory that we include here for completeness and without proof. Their proofs can be found in standard probability textbooks and appropriate references are given below. If we are dealing with two samples, $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_n$, to avoid confusion we use the notation $F_{n,X}$ and $F_{n,Y}$ to denote their empirical distribution functions respectively.

Definition Suppose $X_1, X_2, \cdots$, is a discrete–time stochastic process. Let $\mathcal{F}(X)$ be the σ-algebra generated by the process and $\mathcal{F}(X_n, X_{n+1}, \cdots)$ the σ-algebra generated by $X_n, X_{n+1}, \cdots$. Any $E \in \mathcal{F}(X)$ is called a tail event if $E \in \mathcal{F}(X_n, X_{n+1}, \cdots)$ for any $n \in \mathbb{N}$.

Definition Let $\{A_n\}_{n\in\mathbb{N}}$ be any collection of sets. Then $\{A_n \; i.o.\}$, read as A_n happens infinitely often is defined by:

$$\{A_n \; i.o.\} = \cap_{i\in\mathbb{N}} \cup_{j=i}^{\infty} A_j.$$

Theorem 5.14.1 *(Kolmogorov 0–1 law):*
E being a tail event implies that $P(E)$ is either 0 or 1.

Proof See [9]. ■

Theorem 5.14.2 *(Glivenko–Cantelli Theorem):*
Suppose, $X_1, X_2, \cdots$, i.i.d, has the sample distribution function F_n. Then

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0, \quad a.s..$$

Proof See [7]. ■

Here, we extend the Glivenko–Cantelli Theorem to F^o, G^o and G^c.

Lemma 5.14.3 *Suppose X is a random variable and consider the associated distribution functions F^o_X, G^o_X and G^c_X with corresponding sample distribution functions $F^o_{X,n}, G^o_{X,n}$ and $G^c_{X,n}$. Then*

$$\sup_{x\in\mathbb{R}} |G^o_{X,n} - G^o_X| \to 0, \ a.s.,$$

$$\sup_{x\in\mathbb{R}} |F^o_{X,n} - F^o_X| \to 0, \ a.s.,$$

and

$$\sup_{x\in\mathbb{R}} |G^c_{X,n} - G^c_X| \to 0, \ a.s..$$

Proof Note that

$$F^c_X + G^o_X = 1 \Rightarrow G^o_X = 1 - F^c_X,$$

and

$$F^c_{X,n} + G^o_{X,n} = 1 \Rightarrow G^o_{X,n} = 1 - F^c_{X,n}.$$

Since Glivenko–Cantelli Theorem holds for F^c_X it also holds for G^o_X.

To show the result for F^o_X, note that $F^o_X(x) = G^o_{-X}(-x)$ and $F^o_{X,n}(x) = G^o_{-X,n}(-x)$. Also to show the result for G^c_X note that $G^c_X = 1 - F^o_X$ and $G^c_{X,n} = 1 - F^o_{X,n}$.

■

Theorem 5.14.4 *(Borel–Cantelli lemma):*
Suppose $(\Omega, \mathcal{F}, P)$ is a probability space. Then

1. *$A_n \in \mathcal{F}$ and $\sum_1^\infty P(A_n) < \infty \Rightarrow P(A_n \ i.o) = 0$.*
2. *$A_n \in \mathcal{F}$ independent events with $\sum_1^\infty P(A_n) = \infty \Rightarrow P(A_n \ i.o) = 1$, where i.o. stands for infinitely often.*

Proof See [9]. ■

Theorem 5.14.5 *(Berry–Esseen bound): Let $X_1, X_2, \cdots$, be i.i.d with $E(X_i) = 0 < \infty$, $E(X_i^2) = \sigma$ and $E(|X_i|^3) = \rho$. If G_n is the distribution of*

$$X_1 + \cdots + X_n / \sigma\sqrt{n}$$

and $\Phi(x)$ is the distribution function of a standard normal random variables then

$$|G_n(x) - \Phi(x)| \leq 3\rho/\sigma^3\sqrt{n}.$$

Corollary 5.14.6 *Let $X_1, X_2, \cdots$, be i.i.d with $E(X_i) = \mu < \infty$, $E(|X_i - \mu|^2) = \sigma$ and $E(|X_i - \mu|^3) = \rho$. If G_n is the distribution of $(X_1 + \cdots + X_n - n\mu)/\sigma\sqrt{n} = \sqrt{n}(\frac{\bar{X}_n - \mu}{\sigma})$ and $\Phi(x)$ is the distribution function of a standard normal random variable then*

$$|G_n(x) - \Phi(x)| \leq 3\rho/\sigma^3\sqrt{n}.$$

Proof This corollary is obtained by applying the theorem to $Y_i = X_i - \mu$. ■

Now let $A_n = (X_1 + \cdots + X_n - n\mu)/\sigma\sqrt{n}$. Then

$$|P(A_n > x) - (1 - \Phi(x))| = |P(A_n \leq x) - \Phi(x)| = |G_n(x) - \Phi(x)| < 3\rho/\sigma^3\sqrt{n}.$$

Also

$$|P(x < A_n \leq y) - (\Phi(y) - \Phi(x)))| \leq |G_n(y) - \Phi(y)| + |G_n(x) - \Phi(x)| \leq 6\rho/\sigma^3\sqrt{n}.$$

These inequalities show that for any $\epsilon > 0$ there exist N such that $n > N$,

$$\Phi(z_2) - \Phi(z_1) - \epsilon < P(z_1 < \sqrt{n}(\frac{\bar{X}_n - \mu}{\sigma}) \leq z_2) < \Phi(z_2) - \Phi(z_1) + \epsilon,$$

for $z_1 < z_2 \in \mathbb{R} \cup \{-\infty, \infty\}$.

It is interesting to ask under what conditions lq_{F_n} and rq_{F_n} tend to lq_F and rq_F as $n \to \infty$. Theorem 5.14.7 gives a complete answer to this question.

Theorem 5.14.7 *(Quantile Convergence/Divergence Theorem)*

a) Suppose $rq_F(p) = lq_F(p)$ *then*

$$rq_{F_n}(p) \to rq_F(p), \ \ a.s.,$$

and

$$lq_{F_n}(p) \to lq_F(p), \ \ a.s..$$

b) When $lq_F(p) < rq_F(p)$ *then both* $rq_{F_n}(p), lq_{F_n}(p)$ *diverge almost surely.*

c) Suppose $lq_F(p) < rq_F(p)$*. Then for every* $\epsilon > 0$ *there exists* N *such that* $n > N$,

$$lq_{F_n}(p), rq_{F_n}(p) \in (lq_F(p) - \epsilon, lq_F(p)] \cup [rq_F(p), rq_F(p) + \epsilon).$$

d)

$$\limsup_{n\to\infty} lq_{F_n}(p) = \limsup_{n\to\infty} rq_{F_n}(p) = rq_F(p), \ a.s.,$$

and

$$\liminf_{n\to\infty} lq_{F_n}(p) = \liminf_{n\to\infty} rq_{F_n}(p) = rq_F(p), \ a.s..$$

Proof

a) Since, $lq_F(p) = rq_F(p)$, we use $q_F(p)$ to denote both. Suppose $\epsilon > 0$ is given. Then

$$F(q_F(p) - \epsilon) < p \Rightarrow F(q_F(p) - \epsilon) = p - \delta_1, \ \delta_1 > 0,$$

and

$$F(q_F(p) + \epsilon) > p \Rightarrow F(q_F(p) + \epsilon) = p + \delta_2, \ \delta_2 > 0.$$

By the Glivenko–Cantelli Theorem,

$$F_n(u) \to F(u) \ \ a.s.,$$

uniformly over $\mathbb{R}$. We conclude that

$$F_n(q_F(p) - \epsilon) \to F(q_F(p) - \epsilon) = p - \delta_1, \ \ a.s.,$$

and

$$F_n(q_F(p)+\epsilon) \to F(q_F(p)+\epsilon) = p+\delta_2, \;\; a.s..$$

Let $\epsilon' = \frac{\min(\delta_1,\delta_2)}{2}$. Pick N such that for $n > N$:

$$\begin{aligned} p-\delta_1-\epsilon' < F_n(q_F(p)-\epsilon) < p-\delta_1+\epsilon', \\ p+\delta_2-\epsilon' < F_n(q_F(p)+\epsilon) < p+\delta_2+\epsilon'. \end{aligned}$$

Then

$$\begin{aligned} F_n(q_F(p)-\epsilon) < p-\delta_1+\epsilon' < p \quad \Rightarrow \\ lq_{F_n}(p) \geq q_F(p)-\epsilon \quad \text{and} \quad rq_{F_n}(p) \geq q_F(p)-\epsilon. \end{aligned}$$

Also

$$\begin{aligned} p < p+\delta_2-\epsilon' < F_n(q_F(p)+\epsilon) \quad \Rightarrow \\ lq_{F_n}(p) \leq q_F(p)+\epsilon \quad \text{and} \quad rq_{F_n}(p) \leq q_F(p)+\epsilon. \end{aligned}$$

Re-arranging these inequalities we get:

$$q_F(p)-\epsilon \leq lq_{F_n}(p) \leq q_F(p)+\epsilon,$$

and

$$q_F(p)-\epsilon \leq rq_{F_n}(p) \leq q_F(p)+\epsilon.$$

b) This needs more development in the sequel and the proof follows.

c) This also needs more development in the sequel and the proof follows.

d) If $lq_F(p) = rq_F(p)$ the result follows immediately from (a). Otherwise suppose $lq_F(p) < rq_F(p)$. Then by (b) $lq_{F_n}(p)$ diverges almost surely. Hence $\limsup lq_{F_n}(p) \neq \liminf lq_{F_n}(p), \;\; a.s.$. But by (c), $\forall \epsilon > 0, \; \exists N, \; n > N$

$$lq_{F_n}(p) \in (lq_F(p)-\epsilon, lq_F(p)] \cup [rq_F(p), rq_F(p)+\epsilon).$$

This means that every convergent subsequence of $lq_{F_n}(p)$ has either limit $lq_F(p)$ or $rq_F(p), \;\; a.s.$. Since $\limsup lq_{F_n}(p) \neq \liminf lq_{F_n}(p), \;\; a.s.$, we conclude $\limsup lq_{F_n}(p) = rq_F(p)$ and $\liminf lq_{F_n}(p) = lq_F(p), \;\; a.s.$.
A similar argument works for $rq_{F_n}(p)$.

■

To investigate the case $lq_F(p) \neq rq_F(p)$ more, we start with the simplest example namely a fair coin. Suppose $X_1, X_2, \cdots$ an i.i.d sequence with $P(X_i = -1) = P(X_i = 1) = \frac{1}{2}$ and let $Z_n = \sum_{i=1}^{n} X_i$. Note that

$$Z_n \leq 0 \Leftrightarrow lq_{F_n}(1/2) = -1, \qquad Z_n > 0 \Leftrightarrow lq_{F_n}(1/2) = 1,$$

and

$$Z_n < 0 \Leftrightarrow rq_{F_n}(1/2) = -1, \qquad Z_n \geq 0 \Leftrightarrow rq_{F_n}(1/2) = 1.$$

Hence in order to show that $lq_{F_n}(1/2)$ and $lq_{F_n}(1/2)$ diverge almost surely, we only need to show that $P((Z_n < 0 \ i.o.) \cap (Z_n > 0 \ i.o.)) = 1$. We start with a theorem from [9].

Theorem 5.14.8 *Suppose X_i is as above. Then $P(Z_n = 0 \ i.o.) = 1$.*

Proof The proof of this theorem in [9] uses the Borel–Cantelli Lemma part 2. ■

Theorem 5.14.9 *Suppose, $X_1, X_2, \cdots$ i.i.d. and $P(X_i = -1) = P(X_i = 1) = 1/2$. Then $lq_{F_n}(1/2)$ and $rq_{F_n}(1/2)$ diverge almost surely.*

Proof Suppose, $A = \{Z_n = -1 \ i.o.\}$ and $B = \{Z_n = 1 \ i.o.\}$. It suffices to show that

$$P(A \cap B) = 1.$$

But $\omega \in A \cap B \Rightarrow lq_{F_n}(p)(\omega) = -1, \ i.o.$ and $lq_{F_n}(p)(\omega) = 1, \ i.o.$ Hence $lq_{F_n}(p)(\omega)$ diverges.

Note that $P(A) = P(B)$ by the symmetry of the distribution. Also it is obvious that both A and B are tail events and so have probability either zero or one. To prove $P(A \cap B) = 1$, it only suffices to show that $P(A \cup B) > 0$. Because then at least one of A and B has a positive probability, say A.

$$P(A) > 0 \Rightarrow P(A) = 1 \Rightarrow P(B) = P(A) = 1 \Rightarrow P(A \cap B) = 1.$$

Now let $C = \{Z_n = 0, \ i.o.\}$. Then $P(C) = 1$ by Theorem 5.14.8. If $Z_n(\omega) = 0$ then either $Z_{n+1}(\omega) = 1$ or $Z_{n+1}(\omega) = -1$. Hence if $Z_n(\omega) = 0, \ i.o.$ then at least for one of $a = 1$ or $a = -1$, $Z_n(\omega) = a, \ i.o.$. We conclude that

$\omega \in A \cup B$. This shows $C \subset A \cup B \Rightarrow P(A \cup B) = 1$. ■

To generalize this theorem, suppose $X_1, X_2, \cdots$, arbitrary *i.i.d* process and $lq_F(p) < rq_F(p)$. Define the process

$$Y_i = \begin{cases} 1 & X_i \geq rq_F(p) \\ 0 & X_i \leq lq_F(p). \end{cases}$$

(Note that $P(lq_X(p) < X < rq_X(p)) = 0$.) Then the sequence $Y_1, Y_2, \cdots$ is *i.i.d.*, $P(Y_i = 0) = p$ and $P(Y_i = 1) = 1 - p$. Also note that

$$lq_{F_{n,Y}}(p) \text{ diverges a.s. } \Rightarrow lq_{F_{n,X}}(p) \text{ diverges a.s.}$$

Hence to prove the theorem in general it suffices to prove the theorem for the Y_i process. However, we first prove a lemma that we need in the proof.

Lemma 5.14.10 *Let $Y_1, Y_2, \cdots$ i.i.d with $P(Y_i = 0) = p = 1 - q > 0$ and $P(Y_i = 1) = 1 - p = q > 0$. Let $S_n = \sum_{i=1}^{n} Y_i$, $0 < \alpha$, $k \in \mathbb{N}$. Then there exists a transformation $\phi(k)$ (to $\mathbb{N}$) such that*

$$P(S_{\phi(k)} - \phi(k)q < -k) > 1/2 - \alpha,$$

$$P(S_{\phi(k)} - \phi(k)q > k) > 1/2 - \alpha.$$

Remark. For $\alpha = 1/4$, we get

$$P(S_{\phi(k)} - \phi(k)q < -k) > 1/4,$$

$$P(S_{\phi(k)} - \phi(k)q > k) > 1/4.$$

Proof Since the first three moments of Y_i are finite ($E(Y_i) = q, E(|Y_i - q|^2) = q(1-q) = \sigma, E(|Y_i - q|^3) = q^3(1-q) + (1-q)^3 q = \rho$), we can apply the Berry-Esseen theorem to $\sqrt{n}\frac{\bar{Y}_n - \mu}{\sigma}$. By a corollary of that theorem, for $\frac{\alpha}{2} > 0$ there exists an N_1 such that

$$1 - \Phi(z) - \frac{\alpha}{2} < P(\sqrt{n}\frac{\bar{Y}_n - \mu}{\sigma} > z) < 1 - \Phi(z) + \frac{\alpha}{2},$$

and

$$\Phi(z) - \frac{\alpha}{2} < P(\sqrt{n}\frac{\bar{Y}_n - \mu}{\sigma} < -z) < \Phi(z) + \frac{\alpha}{2},$$

for all $z \in \mathbb{R}$ and $n > N_1$. Now for the given integer k pick N_2 such that

$$\frac{1}{2} - \frac{\alpha}{2} < \Phi(\frac{k}{\sigma\sqrt{N_2}}) < \frac{1}{2} + \frac{\alpha}{2}.$$

This is possible because Φ is continuous and $\Phi(0) = 1/2$. Now let

$$\phi(k) = \max\{N_1, N_2\},\ z = \frac{k}{\sigma\sqrt{\phi(k)}}.$$

Then since $\phi(k) \geq N_1$

$$P(\sqrt{\phi(k)}\frac{\bar{Y}_{\phi(k)} - \mu}{\sigma} > z) > 1 - \Phi(z) - \frac{\alpha}{2} > 1/2 - \alpha,$$

and

$$P(\sqrt{\phi(k)}\frac{\bar{Y}_{\phi(k)} - \mu}{\sigma} < -z) > \Phi(z) - \frac{\alpha}{2} > 1/2 - \alpha.$$

These two inequalities are equivalent to

$$P((S_{\phi(k)} - \phi(k)q) < -k) > 1/2 - \alpha,$$

and

$$P((S_{\phi(k)} - \phi(k)q) > k) > 1/2 - \alpha.$$

If we put $\alpha = 1/4$, we get

$$P((S_{\phi(k)} - \phi(k)q) < -k) > 1/4,$$

and

$$P((S_{\phi(k)} - \phi(k))q > k) > 1/4.$$

■

We are now ready to prove Part b) of Theorem 5.14.7.

Proof [Theorem 5.14.7, Part b)]

For the process $\{Y_i\}$ as defined above, let $n_1 = 1, m_k = n_k + \phi(n_k)$ and $n_{k+1} = m_k + \phi(m_k)$. Then define

$$D_k = (Y_{n_k+1} + \cdots + Y_{m_k} - (m_k - n_k)q < -n_k),$$
$$E_k = (Y_{m_k+1} + \cdots + Y_{n_{k+1}} - (n_{k+1} - m_k)q > m_k),$$
$$C_K = D_k \cap E_k.$$

Since $\{C_k\}$ involve non–overlapping subsequences of Y_s, they are independent events. Also D_k and E_k are independent. Now note that

$$
\begin{aligned}
Y_{n_k+1} + \cdots + Y_{m_k} - (m_k - n_k)q < -n_k \Rightarrow \\
Y_1 + \cdots + Y_{m_k} < -n_k + (m_k - n_k)q + n_k \Rightarrow \\
\bar{Y}_{m_k} < \frac{m_k - n_k}{m_k} q < q \Rightarrow \\
l q_{F_{n,Y}}(p) = r q_{F_{n,Y}} = 0 \Rightarrow \\
\{C_k, \ \ i.o.\} \subset \{l q_{F_{n,Y}}(p) = r q_{F_{n,Y}} = 0, \ \ i.o.\}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
Y_{m_k+1} + \cdots + Y_{n_{k+1}} - (n_{k+1} - m_k)q > m_k \\
\Rightarrow Y_1 + \cdots + Y_{n_{k+1}} > (n_{k+1} - m_k)q + m_k \\
\Rightarrow \bar{Y}_{n_{k+1}} > \frac{m_k + (n_{k+1} - m_k)q}{n_{k+1}} > q = 1 - p \\
\Rightarrow l q_{F_{n,Y}}(p) = r q_{F_{n,Y}}(p) = 1 \\
\Rightarrow \{C_k, \ \ i.o.\} \subset \{l q_{F_{n,Y}}(p) = r q_{F_{n,Y}}(p) = 1, \ \ i.o.\}.
\end{aligned}
$$

Let us compute the probability of C_k:

$$
\begin{aligned}
P(C_k) = \\
P(Y_{n_k+1} + \cdots + Y_{m_k} - (m_k - n_k)q < -n_k) \times \\
P(Y_{m_k+1} + \cdots + Y_{n_{k+1}} - (n_{k+1} - m_k)q > m_k) = \\
P(Y_1 + \cdots + Y_{\phi(n_k)} - \phi(n_k)q < -n_k) \times \\
P(Y_1 + \cdots + Y_{\phi(m_k)} - \phi(m_k)q > m_k) > 1/4.1/4 = 1/16.
\end{aligned}
$$

We conclude that

$$
\sum_{k=1}^{\infty} P(C_k) = \infty.
$$

By the Borel–Cantelli Lemma, $P(C_k, \ \ i.o.) = 1$. We conclude that

$$
P(l q_{F_{n,Y}}(p) = r q_{F_{n,Y}}(p) = 0, \ \ i.o.) = 1,
$$

and

$$
P(l q_{F_{n,Y}}(p) = r q_{F_{n,Y}}(p) = 1, \ \ i.o.) = 1.
$$

Hence,

$$P(\{lq_{F_{n,Y}}(p) = rq_{F_{n,Y}}(p) = 0, \ \ i.o.\} \cap \{lq_{F_{n,Y}}(p) = rq_{F_{n,Y}}(p) = 1, \ \ i.o.\}) = 1.$$

■

Proof (Theorem 5.14.7, part (c))
Suppose that $rq_F(p) = x_1 \neq lq_F(p) = x_2$ and a is an arbitrary real number. Let $h = x_2 - x_1$. We define a new chain Y as follows:

$$Y_i = \begin{cases} X_i & X_i \leq lq_{F_X}(p) \\ X_i - h & X_i \geq rq_{F_X}(p). \end{cases}$$

(See Figure 5.7.) Then $Y_1, Y_2, \cdots$ is an *i.i.d* sample. We drop the index i from Y_i and X_i in the following for simplicity and since the Y_i (as well as the X_i) are identically distributed. We claim

$$lq_{F_Y}Y(p) = rq_{F_Y}(p) = lq_{F_X}(p).$$

To prove $lq_{F_Y}(p) = lq_{F_X}(p)$, note that

$$F_Y(lq_{F_X}(p)) = P(Y \leq lq_{F_X}(p)) \geq P(X \leq lq_{F_X}(p)) \geq p \Rightarrow lq_{F_Y}(p) \leq lq_{F_X}(p).$$

(The first inequality is because $Y \leq X$.) Moreover for any $y < lq_{F_X}(p)$, $F_Y(y) = F_X(y) < p$. (Since $X, Y < lq_{F_X}(p) \Rightarrow X = Y$.) Hence $lq_{F_Y}(p) \geq lq_{F_X}(p)$ and we are done. To show $rq_{F_Y}(p) = lq_{F_X}(p)$, note that $rq_{F_Y}(p) \geq lq_{F_Y}(p) = lq_{F_X}(p)$. It only remains to show that $rq_{F_Y}(p) \leq lq_{F_X}(p)$. Suppose $y > lq_{F_X}(p)$ and let $\delta = y - lq_{F_X}(p) > 0$. First note that

$$\begin{aligned} P(\{Y \leq lq_{F_X}(p) + \delta\}) &= \\ P(\{Y \leq lq_{F_X}(p) + \delta \text{ and } X \geq rq_{F_X}(p)\} &\cup \\ \{Y \leq lq_{F_X}(p) + \delta \text{ and } X \leq lq_{F_X}(p)\}) &= \\ P(\{X - h \leq lq_{F_X}(p) + \delta \text{ and } X \geq rq_{F_X}(p)\} &\cup \\ \{X \leq lq_{F_X}(p) + \delta \text{ and } X \leq lq_{F_X}(p)\}) &= \\ P(\{rq_{F_X}(p) \leq X \leq rq_{F_X}(p) + \delta\} \cup \{X \leq lq_{F_X}(p)\}) &= \\ P(\{X \leq rq_{F_X}(p) + \delta\}). & \end{aligned}$$

Hence,

$$F_Y(y) = P(Y \leq lq_{F_X}(p) + \delta) = P(X \leq rq_{F_X}(p) + \delta) > p \Rightarrow$$

$$rq_{F_Y}(p) \leq y,\ \forall y > lq_{F_X}(p).$$

We conclude that $rq_{F_Y}(p) \leq lq_{F_Y}(p)$.
To complete the proof of part (c) observe that for every $\epsilon > 0$, we may suppose that $lq_{F_{n,Y}}(p) \in (q_{F_Y}(p) - \epsilon, q_{F_Y}(p) + \epsilon)$. Then

$$lq_{F_n,X}(p), rq_{F_{n,X}}(p) \in (lq_{F_X}(p) - \epsilon, rq_{F_X}(p) + \epsilon). \tag{5.6}$$

This is because from $lq_{F_{n,Y}}(p) \in (q_{F_Y}(p) - \epsilon, q_{F_Y}(p) + \epsilon)$, we may conclude that

$$F_{n,Y}(q_{F_Y}(p) + \epsilon) > p \Rightarrow F_{n,X}(rq_{F_X}(p) + \epsilon) > p \Rightarrow$$
$$lq_{F_{n,X}}(p), rq_{F_{n,X}}(p) < rq_{F_X}(p) + \epsilon,$$

and

$$F_{n,Y}(q_{F_Y}(p) - \epsilon) < p \Rightarrow F_{n_X}(lq_{F_X}(p) - \epsilon) < p \Rightarrow$$
$$lq_{F_{n,X}}(p), rq_{F_{n,X}}(p) > lq_{F_X}(p) - \epsilon.$$

But by part (a) of Theorem 5.14.7, $lq_{F_{n,Y}}(p) \to q_{F_Y}(p)$ and $rq_{F_{n,Y}}(p) \to q_{F_Y}(p)$. Hence for given $\epsilon > 0$ there exists an integer N such that for any $n > N$, $lq_{F_{n,Y}}(p) \in (q_{F_Y}(p) - \epsilon, q_{F,Y}(p) + \epsilon)$. By (5.6), we have shown that for every $\epsilon > 0$ there exists N such that for every $n > N$

$$q_{F_{n,X}}(p), rq_{F_{n,X}}(p) \in (lq_{F_X}(p) - \epsilon, rq_{F_X}(p) + \epsilon),$$

since

$$P(X_i \in (lq_{F_X}(p), rq_{F_X}(p)) \text{ for some } i \in \mathbb{N}) = 0.$$

We can conclude that

$$P(lq_{F_{n,X}}(p) \in (lq_{F_X}(p), rq_{F_X}(p)) \text{ for some } i \in \mathbb{N}) = 0$$

and

$$P(rq_{F_{n,X}}(p) \in (lq_{F_X}(p), rq_{F_X}(p)) \text{ for some } i \in \mathbb{N}) = 0.$$

Hence with probability 1

$$q_{F_{n,X}}(p), rq_{F_{n,X}}(p) \in (lq_{F_X}(p) - \epsilon, lq_{F_X}(p)] \cup [rq_{F_X}(p), rq_{F_X}(p) + \epsilon).$$

■

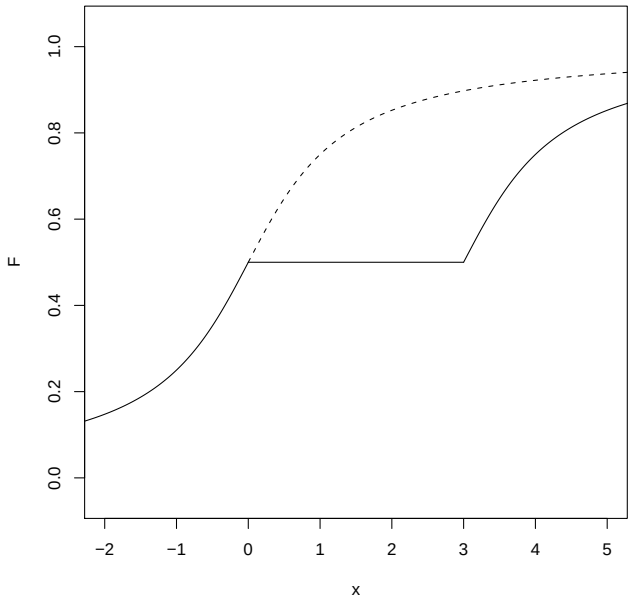

Figure 5.7: The solid line is the distribution function of $\{X_i\}$. Note that for the distribution of the X_i and $p = 0.5$, $lq_{F_X}(p) = 0, rq_{F_X}(p) = 3$. Let $h = rq(p) - lq(p) = 3$. The dotted line is the distribution function of the $\{Y_i\}$ which coincides with that of $\{X_i\}$ to the left of $lq_{F_X}(p)$ and is a backward shift of 3 units for values greater than $rq_{F_X}(p)$. Note that for the $\{Y_i\}$, $lq_{F_Y}(p) = rq_{F_Y}(p) = 1$.

5.15 Summary and discussion

This section highlights the results obtained for a two state-definition for quantiles and discuss why these results show such a consideration is useful.

Justifications and consequences of using left and right quantile functions

1. The equivariance property (under non-decreasing continuous transformations) of lq_X and rq_X makes them equivariant under the change of scale. This is a nice theoretical property. Also from a practical view it means that if we compute the quantile in one scale it can be easily calculated in another scale.

2. Considering lq_X, rq_X allowed us to find a symmetry relation on quantiles:

$$lq_X(p) = -rq_{-X}(1-p).$$

3. We found a nice formula for continuous non-increasing transformations:

$$lq_{\phi(X)}(p) = \phi(rq_X(1-p)).$$

4. We showed that $lq_{F_n}(p)$ the traditional sample quantile function and $rq_{F_n}(p)$ tend to the distribution version if and only if $lq_F(p) = rq_F(p)$. Hence finding a sufficient and necessary condition that is easy to formulate in terms of lq_F and rq_F.

5. If we start with only the traditional quantile function lq_F, then $rq_F(p)$ would arise in the limit

$$\limsup_{n\to\infty} lq_{F_n}(p) = rq_F(p).$$

6. It is widely claimed that the "median" minimizes the absolute error $E|X-a|$. In next chapters, we show that

$$argmin_a E|X-a| = [lq_X(1/2), rq_X(1/2)].$$

   We observe both $lq_X(p)$ and $rq_X(p)$ would arise if we intend to use this as a way defining quantiles. A generalization from $1/2$ to arbitrary p is left for future research.

7. We offered a physical motivation using a uniform bar to define quantiles for data vectors which resulted in a definition that coincide with lq_X, rq_X.

8. If we only use the traditional quantile function, for $p = 0$, we get $lq_X(0) = \infty$ in general. However $rq_X(0) < \infty$ is a useful value in the sense that it is the maximum a satisfying $P(X \geq a) = 1$. Also $rq_X(1) = -\infty$ in general. However $lq_X(1) > -\infty$ in general and is a useful value since it is the minimum a satisfying $P(X \leq a) = 1$.

9. Middle values of $lq_X(p), rq_X(p)$ (for example a specific weighted combination of the two) or the whole interval $[lq_X(p), rq_X(p)]$ are not preferable as a definition. This is because we showed that the range of lq_X and rq_X is exactly the set of heavy points. Points where the probability of being in any positive radius of them is positive.

10. From a practical point of view giving a value that has already occurred as quantile we can expect the same value or a close value happen again in the future. More formally, suppose a random sample $X_1, \cdots, X_n$ is given and we want to compute the sample qunatile. Then $lq_{F_n}(p)$ and $rq_{F_n}(p)$ are one of X_is by definition. If we denote X_F a future value meaning that X_F is identically distributed and independent from $X_1, \cdots, X_n$

$$P(X_F \in (X_i - \epsilon, X_i + \epsilon)) > 0.$$

A middle value might not satisfy such a property.

11. We found out a clean nice way to show in what sense exactly lq_X and rq_X are close. We showed

$$P(lq_X(p) < X < rq_X(p)) = 0.$$

For data vectors this means the two values are side by side in the sorted vector.

12. We showed that $lq_X(p)$ and $rq_X(p)$ coincide except for at most a countable subset of the reals.

13. We showed that even though $lq_X(p) \leq rq_X(p)$ in general, they are not too far apart since for a very small positive value ϵ

$$lq_X(p) \leq rq_X(p) \leq lq_X(p + \epsilon).$$

14. Given one of lq_F or rq_F, the other one can be obtained by taking the limits
$$lq_F(p_0) = \lim_{p \uparrow p_0} rq_F(p),$$
and
$$rq_F(p_0) = \lim_{p \downarrow p_0} lq_F(p).$$

15. In order to invert F, lq_F, rq_F gives us nice expressions for sets such as $x|F(x) > p$ which is equal to $(rq_X(p), \infty)$ if F is continuous at $rq_X(p)$.

16. For a continuous distribution function, we have a nice formula for the inverse based on lq_F and rq_F
$$F^{-1}(p) = [lq_X(p), rq_X(p)].$$

17. The left (right) quantile function at given probability p can be simply put as the minimal value that the distribution function reaches (passes) p.

In some practices fixing one lq or rq might be sufficient. This is because lq and rq are close in terms of the probability of the underlying random variable. For example in data vectors lq, rq will be at most one element off in terms of their position in the data vector.

In most elementary statistics text books and statistical softwares quantiles are given as a one-state solution generally a weighted combination of the left and right quantiles. In order to teach the right and left quantile functions, we suggest using a simple example $x = (1, 2, 3, 4)$ to show that there are no values in the middle and the left (2) and right median (3) are natural to consider. Then one can point out this can be generalized from $p = 1/2$ to any p without getting into details. It can also be pointed out that the left (right) quantile function at given probability p can be simply put as the minimal value that the distribution function reaches (passes) p. In a more advanced courses perhaps for mathematics, statistics or science students the teacher might like to show how the quantiles can be defined using the bar of length 1. Finally the mathematical formulas can be given to students with appropriate mathematical background (i.e. Familiar with the definition of sup and inf and their existence property for the real numbers).

In case an interpolation procedure is to be used, we suggest the interpolation procedure to be between $lq_X(p)$ and $rq_X(p)$. Surprisingly this is not the case. For example for $x = (0, 0, 0, 0, 0, 1, 1, 1, 1, 1)$ in the R package as

the quantile for $p = 0.48$, we get 0.32. But in the vector we notice that 0s have covered 50 percent of the data and since 0.48 is strictly less than 0.48, we expect 1 to be the quantile. 0.32 in our notation is both greater than $lq_x(0.50)$ and $rq_x(0.50)$.

Chapter 6

Probability loss function

6.1 Introduction

This chapter develops a "loss function" to assess the goodness of an approximation or an estimator of quantiles of a distribution (or a data vector). Suppose a quantile of a very large data vector, q is approximated by $\hat{q}$. Several classic losses can be considered. For example: absolute error $L(q, \hat{q}) = |q - \hat{q}|$ or squared error $L(q, \hat{q}) = (q - \hat{q})^2$ which was proposed by Gauss. Quoting from [30]: "Gauss proposed the square of the error as a measure of loss or inaccuracy. Should someone object to this specification as arbitrary, he writes, he is in complete agreement. He defends his choice by an appeal to mathematical simplicity and convenience." An obvious problem with this loss is its lack of invariance under re-scaling of of data. We propose a loss function that is invariant under strictly monotonic transformations. We also show that the sample version of this loss function tends uniformly to the distributional version. This loss function can be used also to find optimal ways to summarize a data vector and to define a measure of distance among random variables as shown in the next chapters.

We define the loss of estimating/approximating q by $\hat{q}$ to be the probability that the random variable falls in between the two values. A limited version of this concept only for data vectors can be found in computer science literature, where ϵ-approximations are used to approximate quantiles of large datasets. (See for example [32].) However, this concept has not been introduced as a measure of loss and the definition is limited to data vectors rather than arbitrary distributions.

6.2 Degree of separation between data vectors

Our purpose is to find good approximations to the median and other quantiles. It is not clear how such approximations should be assessed. We contend that such a method should not depend on the scale of the data. In other words it should be invariant under monotonic transformations. We

define a function δ that measures a natural "degree of separation" between data points of a data vector x. For the sake of illustration, consider the example $sort(x) = (1, 2, 3, 3, 4, 4, 4, 5, 6, 6, 7)$. Now suppose, we want to define the degree of separation of 3,4 and 7 in this example. Since 4 comes right after 3, we consider their degree of separation to be zero. There are 3 elements between 4 and 7 so it is appealing to measure their degree of separation as 3 but since the degree of separation should be relative, we cab also divide by $n = 11$, the length of the vector, and get: $\delta(4, 7) = 3/11$. We can generalize this idea to get a definition for all pairs in $\mathbb{R}$. With the same example, suppose we want to compute the degree of separation between 2.5 and 4.5 that are not members of the data vector. Then since there are 5 elements of the data vector between these two values, we define their degree of separation as 5/11. More formally, we give the following definition.

Definition Suppose $z < z'$ let $\Delta_x(z, z') = \{i | z < x_i < z'\}$. Then we define

$$\delta_x(z, z') = \frac{|\Delta_x(z, z')|}{n},$$

and $\delta_x(z, z) = 0$. We call δ_x the "degree of separation" (DOS) or the "probability loss function" associated with x.

We then have the following lemma about the properties of δ.

Lemma 6.2.1 *The degree of separation δ_x has the following properties:*

a) $\delta_x \geq 0$.

b) $y < y' < y'' \Rightarrow \delta_x(y, y'') \geq \delta_x(y, y')$.

c) If $z < z'$ and z, z' are elements of x, $\delta_x(z, z') = \frac{m_x(z) - M_x(z') - 1}{n}$. [For the definition of $m(z)$ and $M(z)$ see Chapter 5.]

d) $\delta_{\phi(x)}(\phi(z), \phi(z')) = \delta_x(z, z')$ if ϕ is a strictly monotonic transformation.

e) $y = sort(x)$ and $y' = y_i < y'' = y_j \Rightarrow \delta_x(y', y'') \leq (j - i - 1)/n$.

Proof

Both a) and b) are straightforward. We obtain c) as a straightforward consequence of the definition of $m_x(y')$ and $M_x(y')$. To show (d), suppose $z < z'$ and ϕ is strictly decreasing. (The strictly increasing case is similar.) Then $\phi(z') < \phi(z)$ and hence

$$\Delta_{\phi(x)}(\phi(z), \phi(z')) = \{i | \phi(z') < \phi(x_i) < \phi(z)\} = \{i | z < x_i < z'\} = \Delta_x(z, z').$$

Finally e) is true because $|\Delta_x(y', y'')| = |\{l | y_i < x_l < y_j\}| \leq j - i - 1$. ■

All the definitions and results above can be applied to random vectors $X = (X_1, \cdots, X_n)$ as well. In that case, $lq_X(p)$ and $rq_X(p)$ and $\delta_X(z, z')$ are random. To develop our theory, we need to study the asymptotic behavior of these statistics. We do so in later sections.

6.3 "Degree of separation" for distributions: the "probability loss function"

We define a degree of separation for distributions which corresponds to the notion of "degree of separation" defined for data vectors to measure separation between data points.

Definition Suppose X has a distribution function F. Let

$$\delta_F(z', z) = \delta_F(z, z') = \lim_{u \to z^-} F(u) - F(z') = P(z' < X < z), \;\; z > z',$$

and $\delta_F(z, z) = 0, \;\; z \in \mathbb{R}$. We also denote this by δ_X whenever a random variable X with distribution F is specified. We call δ_X the "degree of separation" or the "probability loss function" associated with X.

The following lemma is a straightforward consequence of the definition.

Lemma 6.3.1 *Suppose $x = (x_1, \cdots, x_n)$ is a data vector with the empirical distribution F_n. Then*

$$\delta_{F_n}(z, z') = \delta_x(z, z'), \;\; z, z' \in \mathbb{R}.$$

This lemma implies that to prove a result about the degree of separation of data vectors, it suffices to show the result for the degree of separation of random variables.

Theorem 6.3.2 *Let X, Y be random variables and F_X, F_Y, their corresponding distribution functions.*
a) Assume $Y = \phi(X)$, for a strictly increasing or decreasing function $\phi : \mathbb{R} \to \mathbb{R}$. Then $\delta_{F_X}(z, z') = \delta_{F_Y}(\phi(z), \phi(z')), \;\; z < z' \in \mathbb{R}$.
b) $\delta_F(z, z') \leq \delta_F(z, z''), \;\; z \leq \;\; z' \leq z''$.
c) $\delta_F(z_1, z_3) \leq \delta_F(z_1, z_2) + \delta_F(z_2, z_3) + P(X = z_2)$.

d) Suppose, $p \in [0,1]$. *Then* $\delta_F(lq_F(p), rq_F(p)) = 0$.
e) Suppose, $p_1 < p_2 \in [0,1]$. *Then* $\delta_F(lq_F(p_1), rq_F(p_2)) \leq p_2 - p_1$. *This immediately implies* $\delta_F(lq_F(p_1), lq_F(p_2)) \leq p_2 - p_1$ *and* $\delta_F(rq_F(p_1), lq_F(p_2)) \leq p_2 - p_1$ *by b).*

Remark. We may restate Part (c), for data vectors: Suppose x has length n and z_2 is of multiplicity m, (which can be zero). Then the inequality in (c) is equivalent to $\delta_x(z_1, z_3) \leq \delta_x(z_1, z_2) + \delta_x(z_2, z_3) + m/n$.

Proof

a) Note that for a strictly increasing function ϕ, we have

$$P(z < X < z') = P(\phi(z) < \phi(X) < \phi(z')).$$

Now suppose ϕ is strictly decreasing. Then $z < z' \Rightarrow \phi(z') < \phi(z)$. Let $Y = \phi(X)$. Then

$$\delta_X(z, z') = P(z < X < z') = P(\phi(z') < \phi(X) < \phi(z)) = \delta_Y(\phi(z), \phi(z')).$$

b) This is trivial.

c) Consider the case $z_1 < z_2 < z_3$. (The other cases are easier to show.) Then

$$\delta_F(z_1, z_3) = P(z_1 < X < z_3) = P(z_1 < X < z_2) + P(X = z_2) + P(z_2 < X < z_3)$$

$$= \delta(z_1, z_2) + \delta(z_2, z_3) + P(X = z_2).$$

d) This result is a straightforward consequence of Lemma 5.3.1 b) and c).

e) This result follows from

$$\delta_F(lq(p_1), rq(p_2)) = P(lq(p_1) < X < rq(p_2))$$

$$= P(X < rq(p_2)) - P(X \leq lq(p_1)) \leq p_2 - p_1.$$

The last inequality being a result of Lemma 5.3.1 a) and d). ■

Remark. We call part c) of the above theorem the pseudo–triangle inequality.

Here we give two examples about using the probability loss function and its interpretation.

Example We showed above that the triangle property does not hold for the probability loss function and that might lead to the criticism that this definition is not intuitively appealing. By an example, we now show why it makes sense that the triangle property should not hold for such a situation. Suppose a few mathematicians are standing in a line

Euclid, Khawarzmi, Khayyam, Gauss, Von Neumann.

If we were to ask Khwarzmi about his distance from Euclid, he would answer: "0, since I am right beside him." If we ask Khwarazmi again about his distance to Khayyam, he will say that "my distance is 0 since I am right beside him." However if we were to ask Euclid about his distance to Khayyam he would answer: "One unit (person) since Khwarzmi is in the middle." We observe that this distance does not satisfy the triangle property as well. In this example the people sitting in the middle are the relevant factors. If we deal with a vector of sorted observations, then observations in the middle are the relevant factors.

Example A student is told that he will receive a scholarship if he ranks first in an exam in his class in either of the subjects mathematics and physics. The teacher of the courses differ and take a practice exam in each subject. They return the students back their marks out of 100. They also publish the lists of all the marks after removing the names, to give the students a feeling of how they did in the class. Table 6.1 shows the marks in mathematics and physics.

Mathematics	Physics	Physics before scaling
80	90	81.0
65	89	79.2
63	86	74.0
61	85	72.2
54	83	68.9
54	82	67.2
53	79	62.4
50	79	62.4
49	76	57.8
48	75	56.2
47	72	51.8
47	72	51.8
46	69	47.6
44	68	46.2
30	55	30.2

Table 6.1: A class marks in mathematics and physics. The third column are the raw physics marks before the physics teacher scaled them.

Reza got 63 in math and 75 in physics. He decided to focus on just one subject that gives him a better chance in order to win the scholarship. He compared his mark in math with the best student in math: 63 against 80. So he needed

$$|\text{best mark} - \text{Reza's mark}| = 80 - 63 = 17$$

more marks to be as good as the best student. Then he compared his physics mark to the best student in physics. He found he needs 90-75=15 marks to be as good as him. So he thought it's better to focus on physics. But then he realized that different teachers use different exam and scoring methods. He had heard that the physics teacher scales the marks upward by the formula

$$\text{new mark} = \sqrt{100 \times \text{old mark}}.$$

So the student calculated the untransformed values and put the result in the third column. Now he noticed that his new mark is 56.2 while the best mark is 91. The difference this time is 24.8 which is a larger difference than before. According to his "decision-making tool", the absolute difference, he should focus on math since the absolute difference for math was only 17. But what if the mathematics teacher had used another transformation to re-scale the marks without him knowing it? This made him see a disadvantage to using the absolute value difference. Instead he realized, he can use the number

of the students between himself and the best student as a measure of the difficulty of getting the best mark. He noticed his decision in this case will be independent of how the teachers re-scaled the marks. In the math case there is only one and for physics there are 8 students between him and the best student. Hence he decided that he should focus on math.

This example was under the assumption that other students do not change their study habits or do not have access to the marks. If the other students had access to their marks or were ready to change their study focus, we need to take into account other possible actions of the other students and the problem will become game-theoretical in nature, a very interesting problem on its own right. The solution for that problem we conjecture to be the same.

6.4 Limit theory for the probability loss function

Theorem 6.4.1 *Suppose $X_1, X_2, \cdots$, is a sequence of i.i.d random variables with distribution function F. Then as $n \to \infty$,*

$$\delta_{F_n}(z, z') \to \delta_F(z, z'), \ \ a.s.,$$

uniformly in $z, z' \in \mathbb{R}$. In other words

$$\sup_{z > z' \in \mathbb{R}} |\delta_{F_n}(z, z') - \delta_F(z, z')| \to 0, \ \ a.s..$$

Proof If $z = z'$, the result is trivial. Suppose $z > z'$. We need to show that

$$\lim_{u \to z^-} F_n(u) - F_n(z') \underset{a.s.}{\to} \lim_{u \to z^-} F(u) - F(z'), \tag{6.1}$$

as $n \to \infty$, uniformly in $z > z' \in \mathbb{R}$. Suppose $\epsilon > 0$ is given. By Glivenko-Cantelli Theorem there exist $N \in \mathbb{N}$ such that for every $n > N$:

$$|F_n(u) - F(u)| < \frac{\epsilon}{2}, \ \ a.s., \ \ \forall u \in \mathbb{R}.$$

Now for $n > N$,

$$|(\lim_{u \to z^-} F_n(u) - F_n(z')) - (\lim_{u \to z^-} F(u) - F(z'))| \leq$$

$$|\lim_{u \to z^-} (F_n(u) - F(u))| + |F_n(z') - F(z')| = \lim_{u \to z^-} |F_n(u) - F(u)| + |F_n(z') - F(z')|.$$

But since $|F_n(u) - F(u)| < \frac{\epsilon}{2}$, $\lim_{u \to z^-} |F_n(u) - F(u)| \leq \frac{\epsilon}{2}$. Also $|F_n(z') - F(z')| < \frac{\epsilon}{2}$. Hence

$$|(\lim_{u \to z^-} F_n(u) - F_n(z')) - (\lim_{u \to z^-} F(u) - F(z'))| < \epsilon.$$

■

6.5 The probability loss function for the continuous case

This section studies the probability loss function when the distribution function is continuous. The results are given in the following lemmas, which show some of its desirable properties in the continuous case.

Lemma 6.5.1 *(Probability loss for continuous distributions) Suppose X is a random variable with distribution function F_X. Then $\delta_X(lq_X(p_1), rq_X(p_2)) = p_2 - p_1,\ p_2 > p_1,\ \forall p_1, p_2 \in [0,1]$ iff F_X is continuous.*

Proof

If F_X is continuous then for $p_1 < p_2$ and by Lemma 5.5.2,

$$\delta(lq_X(p_1), rq_X(p_2)) = P(lq_X(p_1) < X < rq_X(p_2)) =$$
$$P(X < rq_X(p_2)) - P(X \leq lq_X(p)) = F(rq_X(p_2)) - F(lq_X(p_2)) = p_2 - p_1.$$

If F is not continuous then there exists an x_0 such that $a = P_X(X = x_0) > 0$. Let $p_1 = P(X < x_0) + a/3$ and $p_2 = P(X < x_0) + a/2$. Clearly $lq_X(p_1) = x_0$ and $rq_X(p_2) = x_0$. Hence

$$\delta(lq_X(p_1), rq_X(p_2)) = 0 \neq p_2 - p_1.$$

■

Lemma 6.5.2 *Suppose $\delta(lq_X(p_1), rq_X(p_2)) = \delta(rq_X(p_1), lq_X(p_2)) = a,\ p_1 < p_2$.*
Then also

$$\begin{aligned} a &= \delta(lq_X(p_1), lq_X(p_2)) \\ &= \delta(rq_X(p_1), lq_X(p_2)) \\ &= \delta(rq_X(p_1), rq_X(p_2)). \end{aligned}$$

Moreover, if X is continuous, all the above are equal to $p_2 - p_1$.

Proof The result follows immediately from the fact that all the three quantities are greater than or equal to $\delta(rq_X(p_1), lq_X(p_2)) = a$ and smaller than or equal to $\delta(lq_X(p_1), rq_X(p_2)) = a$. The second part is straightforward using the previous lemma. ■

6.6 The supremum of δ_X

This section investigates how large the probability loss can become under various scenarios. The results are given in the following lemmas.

Lemma 6.6.1 *Let Dist be the set of all distribution functions. Then*

$$\sup_{F \in Dist} \delta_F(lq_F(p_1), lq_F(p_2)) = p_2 - p_1, \;\; p_2 > p_1, \; p_1, p_2 \in (0,1).$$

Proof This follows from the fact that $\delta_F(lq_F(p_1), lq_F(p_2)) \leq p_2 - p_1$ in general, as shown in Lemma 6.3.2 and $\delta_F(lq_F(p_1), lq_F(p_2)) = p_2 - p_1$ for continuous variables. ■

The same is true for data vectors as shown in the following lemma.

Lemma 6.6.2 *Suppose the supremum in the following is taken over all data vectors, then*

$$\sup_x \delta_x(lq_x(p_1), lq_x(p_2)) = p_2 - p_1, \;\; p_2 > p_1, \; p_1, p_2 \in (0,1).$$

Proof We know that $\delta_x(lq_x(p_1), lq_x(p_2)) \leq p_2 - p_1$. To show that the supremum attains the upper bound, let $x^n = (1, \cdots, n)$. Then $lq_{x^n}(p_1) = [np_1]$ or $[np_1] + 1$. Also $lq_{x^n}(p_2) = [np_2]$ or $[np_2] + 1$. Then Δ, the number of elements of x between $lq_{x^n}(p_1)$ and $lq_{x^n}(p_2)$ satisfies:

$$[np_2] - [np_1] - 1 \leq \Delta \leq [np_2] - [np_1] + 1 \Rightarrow$$

$$np_2 - 1 - np_1 - 1 - 1 \leq \Delta \leq np_2 - np_1 + 1 \Rightarrow$$

$$-3/n \leq \delta_{x^n}(p_1, p_2) - (p_2 - p_1) \leq 1/n.$$

This shows that $\delta_{x^n}(p_1, p_2)$ tends to $p_2 - p_1$ uniformly for all $p_1 < p_2 \in [0,1]$. ■

Lemma 6.6.3 *Suppose* $p_1, p_2, \cdots, p_m \in [0,1]$ *and* $m = 2k$. *Then*

$$\sup_x \max\{\delta_x(lq_x(p_1), lq_x(p_2)), \delta_x(lq_x(p_3), lq_x(p_4)), \cdots, \delta_x(lq_x(p_{m-1}), lq_x(p_m))\}$$

$$= \max\{|p_2 - p_1|, \cdots, |p_m - p_{m-1}|\}.$$

Proof The supremum is less than or equal to the left hand side by Lemma 5.3.1. Let $x^n = (1, 2, \cdots, n)$. Without loss of generality suppose $p_1 < p_2, p_3 < p_4, \cdots, p_{2k-1} < p_{2k}$. By the properties of quantiles of data vectors: $lq_{x^n}(p_i) = x_{[np_i]} = [np_i]$ or $lq_{x^n}(p_i) = x_{[np_i]+1} = [np_i] + 1$.
Also, $lq_{x^n}(p_{i+1}) = x_{[np_{i+1}]} = [np_{i+1}]$ or $lq_{x^n}(p_{i+1}) = x_{[np_{i+1}]+1} = [np_{i+1}] + 1$.
Then, $\delta_{x^n}(lq_{x^n}(p_i), lq_{x^n}(p_{i+1})) \geq \frac{1}{n}([np_{i+1}] - [np_i] - 1) \geq \frac{1}{n}(np_{i+1} - np_i - 2) = (p_{i+1} - p_i) - \frac{2}{n}$. Hence

$$\delta_{x^n}(lq_{x^n}(p_i), lq_{x^n}(p_{i+1})) > |p_{i+1} - p_i| - \frac{2}{n}, \quad i = 1, \cdots, m-1.$$

The inequality shows the supremum is greater than

$$= \max\{|p_2 - p_1| - \frac{2}{n}, \cdots, |p_m - p_{m-1}| - \frac{2}{n}\},$$

for all $n \in \mathbb{N}$. Now let $n \to +\infty$ to get the conclusion.

■

Lemma 6.6.4 *Suppose* $p_1, p_2, \cdots, p_m \in [0,1]$ *and* $a_1, a_1, \cdots, a_{2m} \in [0,1]$. *Then*

$$\sup_x [\int_{a_1}^{a_2} \delta_x(lq_x(p_1), lq_x(p))dp + \int_{a_3}^{a_4} \delta_x(lq_x(p_2), lq_x(p))dp +$$

$$\cdots + \int_{a_{2m-1}}^{a_2 m} \delta_x(lq_x(p_m), lq_x(p))dp]$$

$$= \int_{a_1}^{a_2} |p - p_1|dp + \int_{a_3}^{a_4} |p - p_2|dp + \cdots + \int_{a_{2m-1}}^{a_2 m} |p - p_m|dp.$$

Proof The proof is similar to the previous lemmas and we skip the details.

■

6.6.1 "c-probability loss" functions

This section introduces a family of loss functions that are very similar to the probability loss function but might be more useful in some contexts, particularly when the distribution function is not continuous. A defect of the probability loss function is: it can be equal to zero even if $a \neq b, a, b \in \mathbb{R}$. Also we noted that even though it resembles a metric it is not one. For example the triangle inequality does not hold. We introduce the "c-probability loss function" to solve these problems.

Definition Suppose X is a random variable, δ_X its associated probability loss function and $c \geq 0$. Then let

$$\delta_X^c(a,b) = \delta_X(a,b) + c(1 - 1_{\{0\}}(a-b)),$$

where $1_{\{0\}}$ is the indicator function at zero.

Note that the c-probability loss is the sum of two losses. The first, $\delta_X(a,b)$, is the probability of being between the two values (a and b), the second, $c(1 - 1_{\{0\}}(a-b))$, is the penalty for a and b not being equal. One question is what value of c should be chosen as the "penalty" of not being equal to the true value. It turns out that the value of c is not very important for many purposes as shown in the following lemma.

Lemma 6.6.5 *(Properties of the c-probability loss functions)*
a) $\delta_X^c(a,b) = c \Leftrightarrow a \neq b$ *and* $\delta_X(a,b) = 0$.
b) $\delta_X^c(a,b) = 0$ *or* $\delta_X^c(a,b) \geq c$.
c) δ_X^c *is invariant under strictly monotonic transformations.*
d) Let $d = \sup_{x_0 \in \mathbb{R}} P(X = x_0)$. *Then if* $c \geq d$, δ^c *satisfies the triangle inequality.*
e) $\delta_X^c(lq_X(p), rq_X(p)) \leq c$. *(It is either zero or c.)*
f) Suppose δ_X^c *is given for any* $c > 0$. *Then we can obtain any other* δ_X^d *for* $d \geq 0$.

Proof a) and b) are trivial.
c) Both δ_X and $c(1 - 1_{\{0\}}(a-b))$ are invariant under monotonic transformations.
d) We use the pseudo–triangle inequality for the probability loss function. Take $z_1, z_2, z_3 \in \mathbb{R}$. We need to show $\delta_X^c(z_1,z_3) \leq \delta_X^c(z_1,z_2) + \delta_X^c(z_2,z_3)$. If $z_1 = z_3$, the result is trivial. Otherwise $c(1 - 1_{\{0\}}(z_1 - z_3)) = c$ and

$$\delta_X^c(z_1,z_3) = \delta_X(z_1,z_3) + c \leq \delta_X(z_1,z_2) + \delta_X(z_2,z_3) + P(X = z_2) + c$$

$$\leq \delta_X(z_1, z_2) + \delta_X(z_2, z_3) + c(1 - 1_{\{0\}}(z_1 - z_2)) + c(1 - 1_{\{0\}}(z_2 - z_3)) =$$
$$\delta_X^c(z_1, z_2) + \delta_X^c(z_2, z_3).$$

e) Trivial by properties of lq, rq and δ_X as shown in Lemma 5.3.1.
f) Suppose δ_X^c is given. If $\delta_X^c(a, b) = 0$ then $a = b$ and hence $\delta_X^d(a, b) = 0$. If $a \neq b$ then $\delta_X^c(a, b) = \delta_X(a, b) + c$. From this we can obtain $\delta_X(a, b) = \delta_X^c(a, b) - c$ and hence $\delta_X^d(a, b) = \delta_X^c(a, b) - c + d$. ■

$\delta_X(X_1, X_2)$ (or $\delta_X^c(X_1, X_2)$), if $X_1, X_2 \overset{i.i.d}{\sim} X$ can be considered as a measure of disparity of the common distribution. The following lemma shows that the expectation of this quantity is constant for all continuous random variables!

Lemma 6.6.6 *Suppose X is a continuous random variable, then*

$$E(\delta_X(X_1, X_2)) = 2/3,$$

where $X_1, X_2 \overset{i.i.d}{\sim} X$. Also

$$E(\delta_X^c(X_1, X_2)) = 2/3 + c.$$

Proof We know that $F_X(X_1)$ and $F_X(X_2)$ are both uniformly distributed on (0,1) and independent. Hence

$$\begin{aligned} E(\delta_X(X_1, X_2)) &= E(|F(X_1) - F(X_2)|) = \\ \int_0^1 \int_0^1 |p_1 - p_2| dp_1 dp_2 &= 2 \int_0^1 \int_{p_2}^1 (p_1 - p_2) dp_1 dp_2 = \\ 2 \int_0^1 (1 - 2p_2 + p_2^2) dp_2 &= 2/3. \end{aligned}$$

$E(\delta_X^c(X_1, X_2)) = 2/3 + c$ is obtained by noting that $P(X_1 = X_2) = 0$ for continuous random variables. ■

Chapter 7

Approximating quantiles in large datasets

7.1 Introduction

This chapter develops an algorithm for approximating the quantiles in petascale (petabyte= one million gigabytes) datasets and uses the "probability loss function" to assess the quality of the approximation. The need for such an approximation does not arise for the sample average, another common data summary. That is because if we break down the data to equal partitions and calculate the mean for every partition, the mean of the obtained means is equal to the total mean. It is also easy to recover the total mean from the means of unequal partitions if their length is known.

However computer memories, several gigabytes (GBs) in size, cannot handle large datasets that can be petabytes (PBs) in size. For example, a laptop with 2 GBs of memory, using the well–known R package, could find the median of a data file of about 150 megabytes (MBs) in size. However, it crashed for files larger than this. Since large datasets are commonly assembled in blocks, say by day or by district, that need not be a serious limitation except insofar as the quantiles computed in that way cannot be used to find the overall quantile. Nor would it help to sub–sample these blocks, unless these (possibly dependent) sub–samples could be combined into a grand sub-sample whose quantile could be computed. That will not usually be possible in practice. The algorithm proposed here is a "worst–case" algorithm in the sense that no matter how the data are arranged, we will reach the desired precision. This is of course not true if we sample from the data because there is a (perhaps small) probability that the approximation could be poor.

We also address the following question:

> **Question**: *If we partition the data–file into a number of sub–files and compute the medians of these, is the median of the medians a good approximation to the median of the data–file?*

We first show that the median of the medians does not approximate the exact median well in general, even after imposing conditions on the number of partitions or their length. However for our proposed algorithm, we show how the partitioning idea can be employed differently to get good approximations. "Coarsening" is introduced to summarize data vector with the purpose of inferring about the quantiles of the original vector using the summaries. Then the "d-coarsening" quantile algorithm which works by partitioning the data (or use previously defined partitions) to possibly non-equal partitions, summarizing them using coarsening and inferring about the quantiles of the original data vector using the summaries. Then we show the deterministic accuracy of the algorithm in Theorem 7.4.1. The accuracy is measured in terms of the probability loss function of the original data vector. This is an extension of the work of Albasti et al. in [3] to non-equal size partition case. Theorem 7.4.1 still requires the partition sizes to be divisible by d the coarsening factor. In order to extend the results further to the case where the partitions are nit divisible by d, we investigate how quantiles of a data vector with missing data or contaminated data relate to the quantiles of the original data in Lemma 7.4.3 and Lemma 7.4.4. Also in Lemma 7.5.1, we show if the quantiles of a coarsened vector are used in place of the quantiles of the original data vector how much accuracy will be lost. Finally we investigate the performance of the algorithm using both simulations and real climate datasets.

7.2 Previous work

Finding quantiles and using them to summarize data is of great importance in many fields. One example is the climate studies where we have very large datasets. For example the datasets created by computer climate models are larger than PBs in size. In NCAR (National Center for Atmospheric sciences at Boulder, Colorado), the climate data (outputs of compute models) are saved on several disks. To access different parts of these data a robot needs to change disks form a very large storage space. Another case where we confront large datasets is in dealing with data streams which arise in many different applications such as finance and high–speed networking. For many applications, approximate answers suffice. In computer science, quantiles are important to both data base implementers and data base users. They can also be used by business intelligence applications to drive summary information from huge datasets.

As pointed out by Gurmeet et al. in [32], a good quantile approximation

algorithm should

1. not require prior knowledge of the arrival or value distribution of its inputs.
2. provide explicit and tunable approximation guarantees.
3. compute results in a single pass.
4. produce multiple quantiles at no extra cost.
5. use as little memory as possible.
6. be simple to code and understand.

Finding quantiles of data vectors and sorting them are parallel problems since once we sort a vector finding any given quantile can be done instantly. A good account of early work in sorting algorithms can be found in [28]. Munero et al. in [36] showed for P–pass algorithms (algorithms that scan the data P times) $\Theta(N/P)$ storage locations are necessary and sufficient, where N is the length of the dataset. (See Appendix C for the definitions of complexity functions such as Θ.) It is well–known that the worst-case complexity of sorting is $n \log_2 n + O(1)$ as shown in [33]. In [39], Paterson discusses the progress made in the so–called "selection" problem. He lets $V_k(n)$ be the worst–case minimum number of pairwise comparisons required to find the k–th largest out of n "distinct elements". In particular $M(n) = V_k(n)$ for $k = \lceil n/2 \rceil$. In [8], it is shown that the lower bound for $V_k(n)$ is $n + \min\{k-1, n-k\} - 1$, an achieved upper bound by Blum is $5.43n$. Better upper bounds have been achieved through the years. The best upper bound so far is $2.9423N$ and the lower bound is $(2+\alpha)N$ where α is of order 2^{-40}.

Yao in [49], showed that finding approximate median needs $\Omega(N)$ comparisons in deterministic algorithms. Using sampling this can be reduced to $O(\frac{1}{\epsilon^2} \log(\delta^{-1}))$ independent of N, where ϵ is the accuracy of the approximation in terms of the "probability loss" in our notation. In [36], Munero et al. showed that $O(N^{1/p})$ is necessary and sufficient to find an exact ϕ–quantile in p passes.

Often an exact quantile is not needed. A related problem is finding space–efficient one–pass algorithms to find approximate quantiles. A summary of the work done in this subject and a new method is given in [1]. Two approximate quantile algorithms using only a constant amount of memory were given by Jain [26] Agrawal et al. in [1]. No guarantee for the error was given. Alsabti et al. in [3], provide an algorithm and guaranteed error

in one pass. This algorithm works by partitioning the data into subsets, summarizing each partition and then finding the final quantiles using the summarized partitions. The algorithm in this chapter is an extension of this algorithm to the case of partitions of unequal length.

7.3 The median of the medians

A proposed algorithm to approximate the median of a very large data vector partitions the data into subsets of equal length, computes the median for each partition and then computes the median of the medians. For example, suppose $n = lm$ and break the data to m vectors of size l. One might conjecture that by picking l or m sufficiently large the median of the medians would ensure close proximity to the exact median. We show by an example that taking l and m very large will not help to get close to the exact median. Let $l = 2b + 1$ and $m = 2a + 1$.

partition number	Partition	Median of the partition
1	$(1, 2, \cdots, b, b+1, 10^b, \cdots, 10^b)$	$b+1$
2	$(1, 2, \cdots, b, b+1, 10^b, \cdots, 10^b)$	$b+1$
.	.	.
.	.	.
.	.	.
a	$(1, 2, \cdots, b, b+1, 10^b, \cdots, 10^b)$	$b+1$
a+1	$(1, 2, \cdots, b, b+1, 10^b, \cdots, 10^b)$	10^b
a+2	$(10^b, 10^b, \cdots, 10^b)$	10^b
.	.	.
.	.	.
.	.	.
2a+1	$(10^b, 10^b, \cdots, 10^b)$	10^b

Table 7.1: The table of data

Example

Table 7.1 shows the dataset partitioned into $m = 2a + 1$ vectors of equal length. Every vector is of length $l = 2b + 1$. The first $a + 1$ vectors are identical and 10^b is repeated b times in them. The last a vectors are also identical with all components equal to 10^b. The median of the medians turns out to be $b + 1$. However, the median of the dataset is 10^b. We show that $b + 1$ is in fact "almost" the first quantile. This is because $(b + 1)$ is smaller

than all 10^b's. There are $(a+1)b+a(2b+1)$ data points equal to 10^b. Hence $b+1$ is smaller than this fraction of the data points:

$$\frac{(a+1)b+a(2b+1)}{(2a+1)(2b+1)} = \frac{2a+2}{2a+1}\frac{b}{4b+2} + \frac{a}{2a+1} \approx 1 \times \frac{1}{4} + \frac{1}{2} \approx \frac{3}{4}.$$

With a similar argument, we can show that $b+1$ is greater than almost a quarter of the data points (the ones equal to $1, 2, \cdots, b$). Hence $b+1$ is "almost" the first quantile.

One can prove a rigorous version of the the following statement.

The median of the medians is "almost" between the first and the third quartile.

We only give a heuristic argument for simplicity. To that end, let $n = lm$ and $m = 2a+1$ and $l = 2b+1$. Let M be the exact median and M' be the median of the medians. Order the obtained medians of each partition and denote them by $M_1, \cdots, M_m$. By definition $M' \geq M_j,\ j \leq a$ and $M' \leq M_j,\ j \geq a+1$. Each $M_j,\ \ j \leq a$ is less than or equal to b data points in its partition. Hence, we conclude that M' is less than or equal to ab data points. Similarly M' is greater than or equal to ab data points (which are disjoint for the data points used before). But $\frac{ab}{n} = \frac{ab}{(2a+1)(2b+1)} \approx \frac{1}{4}$. Hence, M' is greater than or equal to $1/4$ data points and less than or equal to $1/4$ data points.

7.4 Data coarsening and quantile approximation algorithm

This section introduces an algorithm to approximate quantiles in very large data vectors. As we demonstrated in the previous section the median of medians algorithm is not necessarily a good approximation to the exact median of a data vector even if we have a large number of partitions and large length of the partitions. The algorithm is based on the idea of "data coarsening" which we will discuss shortly. The proposed algorithm can give us approximations to the exact quantile of known precisions in terms of degree of separation. After stating the algorithm, we prove some theorems that give us the precision of the algorithm. The results hold for partitions of non–equal length.

Definition Suppose a data vector x of length $n = n_1 n_2$ is given, $n_1, n_2 > 1 \in \mathbb{N}$. Also let $sort(x) = y = (y_1, \cdots, y_n)$. Then the n_2–coarsening of x, $C_{n_2}(x)$ is defined to be $(y_{n_2}, y_{2n_2}, \cdots, y_{(n_1-1)n_2})$. Note that $C_{n_2}(x)$ has length $n_1 - 1$. Let $p_i = i/n_1, i = 1, 2, \cdots, (n_1 - 1)$. Then $C_{n_2}(x) = (lq_x(p_1), \cdots, lq_x(p_{n_1-1}))$.

We can immediately generalize the coarsening operator. Suppose

$$sort(x) = (y_1, \cdots, y_n),$$

and $n_2 < n$ is given. Then by The Quotient–Remainder Theorem from elementary number theory, there exist $n_1 \in \mathbb{N} \cup \{0\}$ and $r < n_2$ such that $n = n_1 n_2 + r$. Define $C_{n_2}(x) = (y_{n_2}, \cdots, y_{n_2(n_1-1)})$. The expression is similar to before. However, there are $n_2 + r$ elements after $y_{n_2(n_1-1)}$ in the sorted vector y. In this sense this coarsening is not fully symmetric. We show that if n_2 is small compared to n this lack of symmetry has a small effect on the approximation of quantiles.

Suppose x is a data vector of length $n = \sum_{i=1}^{m} l_i$. We introduce the coarsening algorithm to find approximations to the large data vectors.

d–Coarsening quantiles algorithm:

1. Partition x into vectors of length $l_1, \cdots, l_m$. (Or use pre–existing partitions, e.g. partitions of data saved in various files on the hard disk of a computer.)

$$x^1 = (x_1, \cdots, x_{l_1}), x^2 = (x_{l_1+1}, \cdots, x_{l_1+l_2}), \cdots, x^m = (x_{\sum_{j=1}^{m-1} l_j+1}, \cdots, x_n)$$

2. Sort each x^l, $l = 1, 2, \cdots, m$ and let $y^l = sort(x^l)$, $l = 1, \cdots, m$:

$$y^1 = (y_1^1, \cdots, y_{l_1}^1), \cdots, y^m = (y_1^m, \cdots, y_{l_m}^m).$$

3. d–Coarsen every vector:

$$(y_d^1, \cdots, y_{(c_1-1)d}^1), \cdots, (y_d^m, \cdots, y_{(c_m-1)d}^m),$$

and for simplicity drop d and use the notation $w_i^j = y_{id}^j$.

$$w^1 = (w_1^1, \cdots, w_{(c_1-1)}^1), \cdots, w^m = (w_1^m, \cdots, w_{(c_m-1)}^m).$$

4. Stack all the above vectors into a single vector and call it w. Find $rq_w(p)$ (or $lq_w(p)$) and call it μ. Then μ is our approximation to $rq_x(p)$ (or $lq_x(p)$).

Theorem 7.4.1 *Suppose x is of length $n = \sum_{i=1}^m l_i$, $m \geq 2$ and $l_i = c_i d$. Let $C = \sum_{i=1}^m c_i$. Apply the coarsening algorithm to x and find μ to approximate $rq_x(p)$ (or $lq_x(p)$). Then μ is a (left and right) quantile in the interval*

$$[p - \epsilon, p + \epsilon],$$

where $\epsilon = \frac{m+1}{C-m}$. In other words $\delta_x(\mu, rq_x(p)) \leq \epsilon$ and $\delta_x(\mu, lq_x(p)) \leq \epsilon$. When $l_i = cd$, $i = 1, \cdots, m$, $\epsilon = \frac{m+1}{m-1}\frac{1}{c-1} \leq \frac{3}{c-1}$.

We need an elementary lemma in the proof of this theorem.

Lemma 7.4.2 *(Two interval distance lemma)*
Suppose two intervals $I = [a, b]$ and $J = [c, d]$ subsets of $\mathbb{R}$ are given. Then

$$\sup\{|p - q|, p \in I, q \in J\} = \max\{|a - d|, |b - c|\}.$$

Proof $\sup\{|p - q|, p \in I, q \in J\} \geq \max\{|a - d|, |b - c|\}$ is trivial because $a, b \in I$ and $c, d \in J$.
To show the converse note that $|p - q| = p - q$ or $q - p$, $p \in I, q \in J$. But

$$p - q \leq b - c,$$

and

$$q - p \leq d - a.$$

Hence

$$|p - q| \leq \max\{b - c, d - a\} \leq \max\{|b - c|, |a - d|\}.$$

This completes the proof. ■

Proof of Theorem 7.4.1.

Let $n' = \sum_{i=1}^m (c_i - 1) = \sum_{i=1}^m c_i - m = C - m$ and $M_C = \{(i, j)|i = 1, 2 \cdots, m, j = 1, \cdots, c_i - 1\}$, the index set of w. Also let $c = \max\{c_1, \cdots, c_m\}$.

Suppose, $\frac{h-1}{n'} \leq p < \frac{h}{n'}$, $h = 1, \cdots, n'$. Then since $\mu = rq_w(p)$, there are disjoint subsets of M_C, K and K' such that $|K| = h$, $|K'| = n' - h$, $\mu \geq w_j^i$, $(i, j) \in K$ and $\mu \leq w_j^i$, $(i, j) \in K'$. (This is because if we let $v = sort(w)$, $rq_w(p) = v_h$ since $[n'p] = h - 1$.)

K, K' are not necessarily unique because of possible repetitions among the w_t^i. Hence we impose another condition on K and K'. If $(i,t) \in K$ then $(i,u) \notin K',\ u < t$. It is always possible to arrange for this condition. For suppose, $(i,t) \in K$ and $(i,u) \in K',\ u < t$. Then $\mu \geq w_i^t$ and $\mu \leq w_u^i$, hence $w_t^i \leq w_i^u$. But since $u < t$ we have $w_t^i \leq w_i^u$ by the definition of w^i. We conclude that $w_t^i = w_i^u$. Now we can simply exchange (i,t) and (i,u) between K and K'. If we continue this procedure after finite number of steps we will get K and K' with the desired property.

Now define

- $$K_1 = \{(i,1)|(i,1) \in K\},$$
  with $|K_1| = k_1$ and
  $$I_1 = \{(i,j)|j \leq d, (i,1) \in K\},$$
  Then $|I_1| = k_1 d$. Also note that if $(i,j) \in I_1$, $\mu \geq w_1^i \geq y_j^i$.
- Let
  $$K_2 = \{(i,2)|, (i,2) \in K\},$$
  with $|K_2| = k_2$ and
  $$I_2 = \{(i,j)|d < j \leq 2d, (i,2) \in K\}.$$
  Then $|I_2| = k_2 d$. Also note that if $(i,j) \in I_2$, $\mu \geq w_2^i \geq y_j^i$.
- Let
  $$K_t = \{(i,t)|(i,t) \in K\},$$
  with $|K_t| = k_t$ and
  $$I_t = \{(i,j)|(t-1)d < j \leq td, (i,t) \in K\}.$$
  Then $|I_t| = k_t d$. Also note that if $(i,j) \in I_t$, $\mu \geq w_t^i \geq y_j^i$.
- Let
  $$K_{c-1} = \{(i,(c-1))|(i,c-1) \in K\},$$
  with $|K_{c-1}| = k_{c-1}$ and
  $$I_{(c-1)} = \{(i,j)|(c-2)d < j \leq (c-1)d, (i,c-1) \in K\}.$$
  Then $|I_{c-1}| = k_{c-1} d$. Also note that if $(i,j) \in I_{(c-1)}$, $\mu \geq w_{(c-1)}^i \geq y_j^i$.

Note that $K = \cup_{t=1}^{c-1} K_t$, $|K| = k_1, + \cdots + k_{c-1}$. Since the K_t are disjoint the I_t are also disjoint. Let $I = \cup_{t=1}^{c-1} I_t$ then $|I| = d(k_1 + \cdots + k_{c-1}) = d|K|$. Also note that $(i,j) \in I \Rightarrow \mu \geq y_j^i$.

Similarly define,

- $$K_1' = \{(i,1)|(i,1) \in K'\}, |K_1'| = k_1',$$

  and

  $$I_1' = \{(i,j)|d < j \leq 2d, (i,1) \in K'\}.$$

  Then $|I_1'| = k_1' d$. Also note that if $(i,j) \in I_1'$, $\mu \leq w_1^i \leq y_j^i$.

- Let

  $$K_2' = \{(i,2)|(i,2) \in K'\}, |K_2'| = k_2',$$

  and

  $$I_2' = \{(i,j)|2d < j \leq 3d, (i,2) \in K'\}.$$

  Then $|I_2'| = k_2' d$. Also note that if $(i,j) \in I_2'$, $\mu \leq w_2^i \leq y_j^i$.

- Let

  $$K_t' = \{(i,t)|(i,t) \in K'\}, |K_t'| = k't,$$

  and

  $$I_t' = \{(i,j)|td < j \leq (t+1)d, (i,t) \in K'\}.$$

  Then $|I_t'| = k_t' d$. Also note that if $(i,j) \in I_t'$ then $\mu \leq w_t^i \leq y_j^i$.

- $$K_{c-1}' = \{(i,(c-1))|(i,c-1) \in K'\}, |K_{c-1}'| = k_{c-1}',$$

  and

  $$I_{c-1}' = \{(i,j)|j > (c-1)d, (i,c-1) \in K'\}.$$

  Then $|I_{c-1}'| = k_{c-1}' d$. Also note that if $(i,j) \in I_{c-1}' \Rightarrow \mu \leq w_{(c-1)}^i \leq y_j^i$.

Then $|I| = |K|d$ and $|I'| = |K'|d$. We claim that $I \cap I' = \emptyset$. To see this note that because of how the second components in I_t and I_t' are defined, it is only possible that $I_{t+1} = \{(i,j)|td < j \leq (t+1)d, (i,t+1) \in K\}$ and $I_t' = \{(i,j)|td < j \leq (t+1)d, (i,t) \in K'\}$ intersect for some $t = 1, \cdots, c-2$. But if they intersect then there exist i,t such that $(i,t+1) \in K$ and $(i,t) \in K'$ which is against our assumption regarding K and K'. Hence by Lemma 5.2.4, μ is a quantile between

$$[\frac{|K|d}{n}, \frac{n-|K'|d}{n}] = [\frac{hd}{\sum_{i=1}^{m} c_i d}, \frac{n-(n'-h)d}{\sum_{i=1}^{m} c_i d}] = [\frac{h}{C}, \frac{m+h}{C}].$$

But we know that

$$p \in [\frac{h-1}{C-m}, \frac{h}{C-m}).$$

We are dealing with two interval in one of them μ is a quantile and the other contains p.

We showed in Lemma 7.4.2 if two intervals $[a, b]$ and $[c, d]$ are given, the sup distance between two elements of the two intervals is

$$\max\{|a-d|, |b-c|\}.$$

Applying this to the above two intervals we get,

$$\max\{|\frac{m+h}{C} - \frac{h-1}{C-m}|, |\frac{h-1}{C-m} - \frac{h}{C}|\},$$

which is equal to,

$$\max\{|\frac{mC-m^2-hm+C}{C(C-m)}|, |\frac{C-hm}{C(C-m)}|\}.$$

But $m^2 + hm \leq m^2 + (C-m)m = mC$. Hence

$$|\frac{mC-m^2-hm+C}{C(C-m)}| = \frac{mC-m^2-hm+C}{C(C-m)} \leq \frac{mC+C}{C(C-m)} = \frac{m+1}{C-m}.$$

Also

$$|\frac{C-hm}{C(C-m)}| \leq \frac{C+mC}{C(C-m)} \leq \frac{m+1}{C-m}.$$

Hence the max is smaller than $\epsilon = \frac{m+1}{C-m}$ and we conclude that μ is a quantile for p' which is at most as far as ϵ to p.

The case $l_i = cd$ is easily obtained by replacing $C = mc$ and noting that $\frac{m+1}{m-1} \leq 3$ $m \geq 2$. ■

In most applications, usually the data partitions are not divisible by d. For example the data might be stored in files of different length with common factors. Another situation involves a very large file that is needed to be read in successive stages because of memory limitations. Suppose that

we need a precision ϵ (in terms of degree of separation) and based on that we find an appropriate c and m. Note that n might not be divisible by mc.

First we prove two lemmas. These lemmas show what happens to the quantiles if we throw away a small portion of the data vector or add some more data to it. The first lemma is for a situation that we have thrown away or ignored a small part of the data. The second lemma is for a situation that a small part of the data are contaminated or includes outliers. In both cases, we show how the quantiles computed in the "imperfect" vectors correspond to the quantiles of the original vector. In both case x stands for the imperfect vector and w is the complete/clean data.

Lemma 7.4.3 *(Missing data quantile summary lemma)*
Suppose $x = (x_1, \cdots, x_n)$, $sort(x) = (y_1, \cdots, y_n)$ *and* $y' = lq_x(p), p \in [0,1]$. *Consider a vector* $x^\star$ *of length* $n^\star$ *and let* $w = stack(x, x^\star)$. *Then* $y' = lq_w(p')$, *where* $p' \in [p-\epsilon, p+\epsilon]$ *and* $\epsilon = \frac{n^\star}{n+n^\star}$.

Similarly if $y' = rq_x(p)$ *and* $p \in [0,1]$, $y' = rq_w(p')$, *where* $p' \in [p-\epsilon, p+\epsilon]$ *and* $\epsilon = \frac{n^\star}{n+n^\star}$.

Proof We prove the result for lq_x only and a similar argument works for rq_x.
Let $z = sort(w)$ then $lq_z = lq_w$. For $p = 1$ the result is easy to see. Otherwise, $\frac{i}{n} \le p < \frac{i+1}{n}$ for some $i = 0, \cdots, n-1$. But then $y' = lq_x(p) = y_i$. In the new vector z since we have added $n^\star$ elements $y' = z_j$ for some j, $i \le j < i + n^\star$. Hence $y' = lq_z(\frac{j}{n+n^\star})$. From $np - 1 < i \le np$, we conclude

$$\frac{np-1}{n+n^\star} < \frac{i}{n+n^\star} \le \frac{j}{n+n^\star} < \frac{i+n^\star}{n+n^\star} \le \frac{np+n^\star}{n+n^\star}.$$

Hence,

$$\frac{n^\star(1-p)-1}{n+n^\star} < \frac{j}{n+n^\star} - p < \frac{n^\star(1-p)}{n+n^\star} \Rightarrow$$

$$|\frac{j}{n+n^\star} - p| < \max\{|\frac{n^\star(1-p)-1}{n+n^\star}|, |\frac{n^\star(1-p)}{n+n^\star}|\}.$$

But $|\frac{n^\star(1-p)}{n+n^\star}| \le \frac{n^\star}{n+n^\star}$ and $|\frac{n^\star(1-p)-1}{n+n^\star}| \le \max\{\frac{n^\star-1}{n+n^\star}, \frac{1}{n+n^\star}\}$ since p ranges in $[0,1]$. We conclude that that

$$|\frac{j}{n+n^\star} - p| < \frac{n^\star}{n+n^\star}.$$

■

Lemma 7.4.4 *(Contaminated data quantile summary lemma)*
Suppose $x = (x_1, \cdots, x_n)$, $sort(x) = (y_1, \cdots, y_n)$ *and* $y' = lq_x(p), p \in [0,1]$. *Consider the vector* $w = (x_1, x_2, \cdots, x_{n-n^\star})$ *then* $y' = lq_w(p')$, *where* $p' \in [p-\epsilon, p+\epsilon]$ *and* $\epsilon = \frac{n^\star}{n-n^\star}$.

Similarly if $y' = rq_x(p)$ *and* $p \in [0,1]$, $y' = rq_w(p')$, *where* $p' \in [p-\epsilon, p+\epsilon]$ *and* $\epsilon = \frac{n^\star}{n-n^\star}$.

Proof We only show the case for lq_x and a similar argument works for rq_x. Let $z = sort(w)$. Then $lq_z = lq_w$. If $p = 1$ the result is easy to see. Otherwise, $\frac{i}{n} \leq p < \frac{i+1}{n}$ for some $i = 0, \cdots, n-1$. But then $y' = lq_x(p) = y_i$. In the new vector z since we have removed $n^\star$ elements $y' = z_j$ for some j, $i - n^\star \leq j \leq i$. Hence $y' = lq_z(\frac{j}{n-n^\star})$. From $np - 1 < i \leq np$, we conclude $np - 1 - n^\star < j \leq np \Rightarrow np - n^\star \leq j \leq np$. Hence

$$\frac{-n^\star + n^\star p}{n-n^\star} \leq \frac{j}{n-n^\star} - p \leq \frac{n^\star p}{n-n^\star} \Rightarrow$$

$$|\frac{j}{n-n^\star} - p| \leq \frac{n^\star}{n-n^\star}.$$

■

In the case that the partitions are not divisible by d, we can use the same algorithm with generalized coarsening. The error will increase obviously and the next two lemmas say by how much.

Lemma 7.4.5 *Suppose* x *has length* $n = lm + r$, $0 \leq r < l$ *and* $m = cd$. *To find* $lq_x(p)$, *apply the algorithm in the previous theorems to a sub–vector of* x *of length* lm. *Then the obtained quantile is a quantile for a number in* $[p-\epsilon, p+\epsilon]$, *where* $\epsilon = \frac{m+1}{m-1}\frac{1}{c-1} + \frac{r}{lm+r}$.

Proof The result is a straightforward consequence of the Theorem 7.4.1 and the Lemma 7.4.3. ■

Lemma 7.4.6 *Suppose* x *has length* $n = \sum_{i=1}^m l_i$ *and* $l_i = c_i d + r_i$, $r_i < d$. *Let* $R = \sum_{i=1}^m r_i$. *Then apply the algorithm above to* x *to find* $lq_x(p)$, *using the generalized coarsening. The obtained quantile is a quantile for a number in* $[p-\epsilon, p+\epsilon]$ *where* $\epsilon = \frac{m+1}{C-m} + \frac{R}{R+Cd}$.

Proof Let $l_i' = c_i d$. Consider x' a sub–vector of x consisting of

$$(y_1^1, \cdots, y_{l_1'}^1), (y_1^2, \cdots, y_{l_2'}^2), \cdots, (y_1^m, \cdots, y_{l_m'}^m).$$

Then x' has length $\sum_{i=1}^{m} l'_i$. By Lemma 7.4.3 p-th quantile found by the algorithm is a quantile in $[p-\epsilon_1, p+\epsilon_1]$, $\epsilon_1 = \frac{m+1}{C-m}$ for x'. x has $R = \sum_{i=1}^{m} r_i$ elements more than x'. Hence the obtained quantile is a quantile for x for a number in $[p-\epsilon, p+\epsilon]$, $\epsilon = \epsilon_1 + \frac{R}{R+Cd}$. ■

7.5 The algorithm and computations

Suppose a data vector x has length n. To find the quantiles of this vector, we only need to sort x. Since then for any $p \in (0,1)$, we can find the first h such that $p \geq h/n$. Note that

$$sort(x) = (lq_x(1/n), lq_x(2/n), \cdots, lq_x(1)) = (rq_x(0), rq_x(1/n), \cdots, rq_x(\frac{n-1}{n})).$$

We only focus on left quantiles here. Similar arguments hold for the right quantile.

Obviously, the longer the vector x, the finer the resulting quantiles are. Now imagine that we are given a very long data vector which cannot even be loaded on the computer memory. Firstly, sorting this data is a challenge and secondly, reporting the whole sorted vector is not feasible. Assume that we are given the sorted data vector so that we do not need to sort it. What would be an appropriate summary to report as the quantiles? As we noted also the sorted vector itself although appropriate, maybe of such length as to make further computation and file transfer impossible. The natural alternative would be to coarsen the data vector and report the resulting coarsened vector. To be more precise, suppose, $length(x) = n = n_1 n_2$ and $y = sort(x) = (y_1, \cdots, y_n)$. Then we can report

$$y' = C_{n_2}(y) = (y_{n_2}, \cdots, y_{(n_1-1)n_2}).$$

This corresponds to

$$(lq_{y'}(1/n_2), \cdots, lq_{y'}(1)).$$

How much will be lost by this coarsening? Suppose, we require the left quantile corresponding to $(h-1)/n < p \leq h/n$, $h = 1, \cdots, n$. Then x would give us y_h. But since $(h-1)/n < p \leq h/n$

$$np < h \leq np + 1.$$

Also suppose for some $h' = 1, \cdots, n_1$,

$$(h'-1)/(n_1-1) < p \leq (h')/(n_1-1) \Rightarrow (h'-1) < p(n_1-1) \leq h'$$

$$\Rightarrow (n_1-1)p \leq h' < p(n_1-1)+1.$$

Then

$$(h-1)(n_1-1)/n < h' < h(n_1-1)/n + 1,$$

and

$$(h-1)(n_1-1)n_2/n < h'n_2 < h(n_1-1)n_2/n + n_2. \tag{7.1}$$

Using the coarsened vector, we would report $y_{h'(n_2)}$ as the approximated quantile for p. The degree of separation between this element and the exact quantile using Equation 7.1 is less than or equal to

$$\max\{\frac{|h-(h-1)(n_1-1)n_2/n|}{n}, \frac{|h(n_1-1)n_2/n + n_2 - h|}{n}\}.$$

This equals

$$\max\{|\frac{-hn_2 - n_1n_2 + n_2}{n^2}|, |\frac{-hn_2 + nn_2}{n^2}|\}.$$

But

$$|\frac{-hn_2 - n_1n_2 + n_2}{n^2}| = \frac{n_2(n_1+n-1)}{n^2} < \frac{n_2(n_1+n)}{n^2} = \frac{1}{n} + \frac{n_2}{n},$$

and

$$|\frac{-hn_2 + nn_2}{n^2}| < \frac{n_2}{n}.$$

Hence the degree of separation is less than $1/n + 1/n_1$. We have proved the following lemma.

Lemma 7.5.1 *Suppose x is a data vector of the length $n = n_1n_2$ and $y = sort(x)$, $y' = C_{n_2}(y)$. Then if we use the quantiles of y' in place of x, the accuracy lost in terms of the probability loss of x (δ_x) is less than $1/n + 1/n_1$.*

The algorithm proposes that instead of sorting the whole vector and then coarsening it, coarsen partitions of the data. The accuracy of the quantiles obtained in this way is given in the theorems of the previous section. This allows us to load the data into the memory in stages and avoid program failure due to the length of the data vector. We are also interested in the

performance of the method in terms of speed, and do a simulation study using the "R" package (a well–known software for statistical analysis) to assess this. In order to see theoretical results regarding the complexity of the special case of the algorithm for equal partitions see [3]. For the simulation study, we create a vector, x, of length $n = 10^7$. We apply the algorithm for $m = 1000, c = 20, d = 500$. We create this vector in a loop of length 1000. During each iteration of the loop, we generate a random mean for a normal distribution by first sampling from $N(0, 100)$. Then we sample 10,000 points from a normal distribution with this mean and standard deviation 1. We compare two scenarios:

1. Start by a NULL vector x and in each iteration add the full generated vector of length 10000 to x. After the loop has completed its run, sort the data vector which now has length 10^7 by the command sort in R and use this to find the quantiles.

2. Start with a NULL vector w. During each iteration after generating the random vector, d–coarsen the data by $d = 500$. (Hence $m = 1000$, $c = 20$.) In order to do that computing, first apply the sort command to the data and then simply d–coarsen the resulting sorted vector. During each iteration, add the coarsened vector to w. After all the iterations, sort w and use it to approximate quantiles.

Remark. The first part corresponds to the straightforward quantiles' calculation and the second corresponds to our algorithm. Note that in the real examples instead of the loop, we could have a list of 1000 data files and still this example serves as a way of comparing the straightforward method and our algorithm.

Remark. Note that if we wanted to create an even longer vector say of length 10^{10} then the first method would not even complete because the computer would run out of memory in saving the whole vector x.

Remark. The final stage of the algorithm can use the fact that w is built of ordered vectors to make the algorithm even faster. We will leave that a problem to be investigated in the future.

We have repeated the same procedure for $n = 2 \times 10^7, m = 1000, d = 500$ and $n = 10^8, m = 1000, d = 500$. The results of the simulation are given in Table 7.2, in which "DOS" stands for the degree of separation between the exact median and the approximated median. The "DOS bound" bounds the degree of separation obtained by the theorems in the previous section. For $n = 10^7, n = 2 \times 10^7$ significant time accrue by using the algorithm. For a vector of length 10^8, R crashed when we tried to sort the original vector

and only the algorithm could provide results. For all cases the exact and approximated quantiles are close. In fact the dos is significantly smaller than the dos bound. This is because this is a "worst–case" bound. The exact and approximated quantiles for $n = 10^7$ are plotted in Figure 7.1.

Length	$n = 10^7$	$n = 2 \times 10^7$	$n = 10^8$
Exact median value	1.847120	1.857168	NA
Algorithm median value	1.866882	1.846463	1.846027
DOS	0.00012	-6.475×10^{-5}	NA
DOS bound	0.05268421	0.02566667	0.005030151
Time for exact median	186 sec	461 s	NA
Time for the algorithm	6 sec	18 s	98 s

Table 7.2: Comparing the exact method with the proposed algorithm in R run on a laptop with 512 MB memory and a processor 1500 MHZ, $m = 1000, d = 500$. "DOS" stands for degree of separation in the original vector. "DOS bound" is the theoretical degree of separation obtained by Theorem 7.4.1.

Next, we apply the algorithm on a real dataset. The dataset includes the daily maximum temperature for 25 stations over Alberta during the period 1940–2004. We focus on the 95th percentile. The results are given in Table 7.3. The algorithm finds the percentile more quickly but the time difference is not as large as the simulation. This is because most of the time of the algorithm and the exact computation is spent on reading the files from the hard drive. The dos bound is about 0.01 (on the 0–1 probability scale). The true degree of separation is about 0.001. The estimated quantiles and the exact quantiles are plotted in Figure 7.2. Notice that the exact and approximated values match except at the very beginning (very close to zero) and end (when it is close to 1), where we see that the circles (corresponding to exact quantiles) and the +s (corresponding to the approximated quantiles) do not completely match. This difference is at most 0.01 in terms of dos in any case.

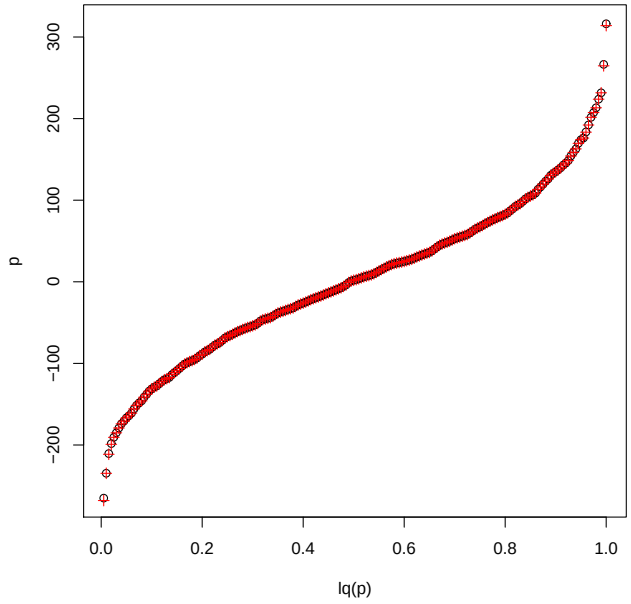

Figure 7.1: Comparing the approximated quantiles to the exact quantiles $N = 10^7$. The circles are the exact quantiles and the + are the corresponding approximated quantiles.

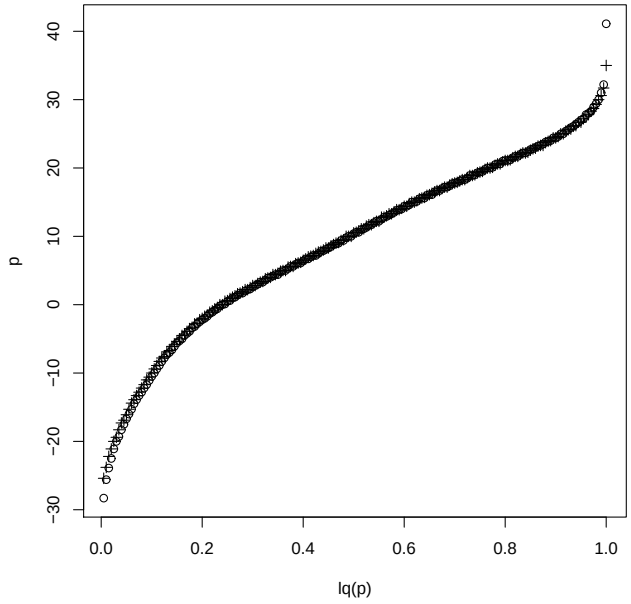

Figure 7.2: Comparing the approximated quantiles to the exact quantiles for MT (daily maximum temperature) over 25 stations in Alberta 1940–2004. The circles are the exact quantiles and the + the approximated quantiles.

Exact 95th percentile	27 C
Algorithm 95th percentile	26.7 C
DOS	0.001278726
DOS bound	0.01052189
time for exact median	8 min 6 sec
time for the algorithm	7 min 29 sec

Table 7.3: Comparing the exact method with the proposed algorithm in R (run on a laptop with 512 MB memory and processor 1500 MHZ) to compute the quantiles of MT (daily maximum temperature) over 25 stations with data from 1940 to 2004.

Chapter 8

Quantile data summaries

8.1 Introduction

This chapter introduces techniques to summarize data (using quantiles), manipulate and combine such summaries. "Weighted data vectors", which are an extension of data vectors are introduced. The operators *sort* and *stack* are extended to weighted data vectors and the operator *comp* (compress) is introduced to compress a data vector as much as possible with no loss of information. In the quantile definition chapter, we expressed a few appealing properties that quantiles should satisfy. We established the equivariance and symmetry properties and left the following to later:

1. The "amount" of data between $q_x(p_1)$ and $q_x(p_2)$ should be a $p_2 - p_1$, $p_1 < p_2$ fraction of the "data amount" of the whole data.

2. If we cut a sorted data vector up until the p_1-th quantile and compute the p_2-th quantile for the new vector, we should get the p_1p_2-th quantile of the original vector. For example the median of a sorted vector upto its median should be the first quartile of the original vector.

A natural definition for the "amount of data" between a, b would be the number of data points between a, b divided by the length of the whole vector. However, by this definition there is no hope of establishing property (1) knowing that $p_2 - p_1$ can be irrational. Also for the second property one might conjecture that if we define the cut operator to be the sorted vector from left to $lq_x(p_1)$ (or $rq_x(p_1)$) then this property holds. However, consider $x = (1, 2)$ and a cut of length 0.6. Then we get the same vector $x' = (1, 2)$ after the cut using this definition since $lq_x(0.6) = 2$. Now the 0.7th left (or right) quantile of the cut vector x' is

$$lq_{x'}(0.7) = 2.$$

However,

$$lq_x(0.6 \times 0.7) = lq_x(0.42) = 1.$$

In the following, we define the cut operator for $p \in (0,1)$ in a way that it ends with $lq_x(p)$ but satisfies property (2). The idea can be explained in the example by considering the vector $x = (1,2)$ as a weighted vector with weights $(1/2, 1/2)$ and give 2 less "weight" than 1 after the cut. In summary, this chapter provides a framework to establish these properties, using the "partition" operator and the "cut" operator.

When dealing with summarized data the following general question is a fundamental one:

Question: Suppose x is a data vector which consists of m subvectors

$$x^1, \cdots, x^m.$$

In other words $x = stack(x^1, \cdots, x^m)$. Assume we do not have access to the x^i but to the w^i, their summaries (possibly a result of coarsening of the x^i). Then how can we approximate the quantiles of the original data vector x and assess how good this approximation is?

We have already encountered such a problem in Chapter 7, where we answered the question in some specific cases. We do not answer the question in general in this chapter but provide a framework to formalize and answer these type of questions.

In computer science quantiles are sometimes used to summarize large datasets. A good summary of the work for creating quantile summaries of datasets in a single pass is given in [19].

In order to make a summary (of length k) of a data vector using the quantiles, one has various choices to pick certain probability indices

$$p_1 \leq p_2 \leq \cdots \leq p_k,$$

and save the corresponding quantiles. Using the probability loss function, we find an optimal way of doing this. Then we consider the problem of finding

$$\underset{a}{argmin} E(L(X,a)),$$

for various L (loss) functions. It is widely claimed that if L is the absolute value function, the $argmin$ is the median of X. We show that the $argmin$ is in fact $[lq_X(1/2), rq_X(1/2)]$. We also find the

$$\underset{a}{argmin} E(\delta_X(X,a)).$$

Finally, we find optimal "probability index vectors" to assign quantiles to a random sample $X_1, \cdots, X_n$, which can be used to make a quantile–quantile plot. Some previous techniques to make a q–q plot are discussed in [24].

8.2 Generalization to weighted vectors

This section extends the definitions and ideas developed before (quantiles, probability loss function, sorting, stacking etc.) from ordinary data vectors to weighted vectors. A weighted vector has two extra components compared to an ordinary vector: a weight allocation and a data amount. This allows us to summarize information in some cases. For example, consider the vector $(1,1,1,1,1,1,1,1,1,2)$. We observe that 1 is repeated 9 times and 2 only one time. We can summarize this by giving the elements $(1,2)$ a weight allocation $(0.9, 0.1)$ and a data amount 10 which is the length of the vector in this case. Weighted vectors also enable us to define the "cut" operator to cut data vectors.

Definition We call a triple $\chi = (x, w^\chi, n^\chi)$ a weighted vector if $length(x) = length(w^\chi) = l_x$, $x = (x_1, \cdots, x_l)$, $w^\chi = (w_1^\chi, \cdots, w_l^\chi)$, $\sum_{i=1}^{l} w_i^\chi = 1$ and n^χ a positive real number. Note that n^χ is not necessarily equal to the length of x. We call w^χ the "weight vector" of χ and n^χ the "data amount" of χ.

Remark. Note that in order to specify a weight vector w, we do not need to specify the last component since the weights must sum up to one.

Examples:

1. $\chi = ((1,2,3),(1/3,1/3,1/3),3)$. This is equivalent to an ordinary vector of length 3 in a sense we make clear soon.

2. $\chi = ((1,2,3),(1/3,1/3,1/3),6)$. Notice this weighted vector has the same elements as before with a data amount of 6 which is two times the previous vector. This vector is equivalent to the ordinary vector $x = (1,1,2,2,3,3)$.

3. $\chi = ((1,1,2,3),(1/6,1/6,1/3,1/3),3)$. This is equivalent to vector given in 1. Note that one is repeated two times here. However, the sum of the weights for 1 is 1/6+1/6=1/3 which is the same as the vector defined in 1.

4. $\chi = ((1),(1),1/2)$. Here we only have 1/2 data amount. i.e. we have less than one observation! (1/2 of an observation to be precise.)

5. $\chi = ((1,2),(1/2,1/2),\sqrt{3})$.

The first vector, x, in the definition $\chi = (x, w^\chi, n^\chi)$, is the vector of possible values, the second one, w^χ, is the corresponding weights for elements of x and the third component, n^χ, is a measure of how fine the vector is.

A vector is called an ordinary vector if the length of x, l_x, is equal to n^χ and $w_i^\chi = w_j^\chi,\ i, j \in 1, \cdots, l_x$. The ordinary vector corresponds to the usual data vectors. Denote the space of all weighted vectors by Υ. We define some operations and an equivalence relation on Υ.

Definition Suppose $\chi = (x, w^\chi, n^\chi)$ then $comp(\chi) = \xi = (y, w^\xi, n^\xi)$, where $y = (y_1, \cdots, y_r)$ is a non-decreasing vector of all disjoint elements of x, $w_i^\xi = \sum_{x_j = y_i} w_j^\chi$ and $n^\xi = n^\chi$.

It is clear that *comp* (compress operator) is an operator from Υ to Υ. Then we define an equivalence relation on Υ.

Definition $\chi \sim \xi$ in Υ iff $comp(\chi) = comp(\xi)$.

Clearly, $\sim$ is an equivalence relation. Let us define a transformation of a weighted vector.

Definition Suppose $\chi = (x, w^\chi, n^\chi)$ is a weighted vector and ϕ a transformation of $\mathbb{R}$ (not necessarily increasing). Then $\phi(\chi) = \zeta = (z, w^\zeta = w^\chi, n^\zeta = n^\chi)$, where $z_i = \phi(x_i),\ i = 1, 2, \cdots, l_x$.

For ordinary vectors x, y, $comp(x) = comp(y)$ iff $sort(x) = sort(y)$. Also *comp* leaves the last component of a weighted vector (the data amount) unchanged.

Since x and w^χ have the same length, we can show an element of Υ by pair consisting of a matrix of dimension $2 \times l_x$ and a number n^χ:

$$\chi = (\begin{pmatrix} x_1 & \cdots & x_{l_x} \\ w_1^\chi & \cdots & w_{l_x}^\chi \end{pmatrix}, n^\chi)$$

Given a weighted vector $\chi = (x, w^\chi, n^\chi)$, we can naturally define a distribution function as follows.

Definition Suppose $\chi = (x, w^\chi, n^\chi)$ is a weighted vector. The the empirical distribution of χ is defined as

$$F_\chi(a) = \sum_{i,\ x_i \le a} w_i^\chi.$$

Remark. If χ is an ordinary vector then F_χ is the usual empirical function.

Then we extend the definition of the stack operator to weighted vectors.

Definition Suppose $\chi = (x, w^\chi, n^\chi)$ and $\xi = (y, w^\xi, n^\xi)$ are given then

$$stack : \Upsilon \times \Upsilon \to \Upsilon,$$

$$(\chi, \xi) \mapsto \zeta = (z, w^\zeta, n^\chi + n^\xi),$$

where (z, w^ζ) in the matrix notation is given by

$$\begin{pmatrix} x_1 & \cdots & x_{l_x} & y_1 & \cdots & y_{l_y} \\ w_1^\chi \frac{n^\chi}{n^\chi+n^\xi} & \cdots & w_{l_x}^\chi \frac{n^\chi}{n^\chi+n^\xi} & w_1^\xi \frac{n^\xi}{n^\chi+n^\xi} & \cdots & w_{l_y}^\xi \frac{n^\xi}{n^\chi+n^\xi} \end{pmatrix}.$$

Remark. In the definition, notice how the data amounts are used to adjust the weights.

Remark. For ordinary vectors x, y the *stack* operator coincide to concatenating x and y.

Lemma 8.2.1 *(Stack operator properties)*
a) The stack operator preserves the equivalence relation defined above, i.e.

$$\chi_1 \sim \xi_1, \;\; \chi_2 \sim \xi_2, \; \textit{then } stack(\chi_1, \chi_2) \sim stack(\xi_1, \xi_2)$$

b)

$$stack(\chi_1, stack(\chi_2, \chi_3)) \sim stack(stack(\chi_1, \chi_2), \chi_3)$$

Proof a) Suppose $\chi_i = (x^i, w^{\chi_i}, n^{\chi_i}), \xi_i = (y^i, w^{\xi_i}, n^{\xi_i})$ and $\chi_i \sim \xi_i$ for $i = 1, 2$. Let

$$\chi = comp(stack(\chi_1, \chi_2)), \; comp(stack(\xi_1, \xi_2)) = \xi.$$

We need to show $\chi = \xi$. Let $\chi = (x, w^\chi, n^\chi)$ and $\xi = (y, w^\xi, n^\xi)$. From $\chi_i = \xi_i$ for $i = 1, 2$, we conclude $n^{\chi_i} = n^{\xi_i}, i = 1, 2$, which in turn gives

$$n^\chi = n^{\chi_1} + n^{\chi_2} = n^{\xi_1} + n^{\xi_2} = n^\xi.$$

Also $x = y$ since both x and y are increasingly sorted and every element in x is an element of x^1 or x^2 which have the same elements as y^1 or y^2. Now to show $w_i^\chi = w_i^\xi, i = 1, 2, \cdots, l_x$, suppose $x_i = y_i$ be the corresponding element in $x = y$. Assume that the corresponding weight for x_i in χ_1 is w and in χ_2 is w'. Then the corresponding weight in ξ_1 and ξ_2 must be w and

w' respectively by the assumed equivalence relations. Hence w_i^{χ} and w_i^{ξ} are equal to

$$w.\frac{n^{\chi_1}}{n^{\chi_1}+n^{\chi_2}}+w'.\frac{n^{\chi_2}}{n^{\chi_1}+n^{\chi_2}},$$

and

$$w.\frac{n^{\xi_1}}{n^{\xi_1}+n^{\xi_2}}+w'.\frac{n^{\xi_2}}{n^{\xi_1}+n^{\xi_2}},$$

which are equal.
b) Let

$$\chi=(x,w^{\chi},n^{\chi})=comp[stack(\chi_1,stack(\chi_2,\chi_3))]$$

and

$$\chi'=(x',w^{\chi'},n^{\chi'})=comp[stack(stack(\chi_1,\chi_2),\chi_3))].$$

We show $\chi=\chi'$. Firstly, note that

$$n^{\chi}=n^{\chi_1}+(n^{\chi_2}+n^{\chi_3})=(n^{\chi_1}+n^{\chi_2})+n^{\chi_3}=n^{\chi'}.$$

$x=x'$ is trivial. Fix $x_i=x_i'$ in $x=x'$. Suppose its corresponding weight in χ_j is equal to $w_j, j=1,2,3$. To show that the corresponding weights w_i^{χ} and $w_i^{\chi'}$ are equal, note that the corresponding weight of x_i in χ is a combination of its weights in χ_1 and $stack(\chi_2,\chi_3)$:

$$w_i^{\chi}=w_1\frac{n^{\chi_1}}{n^{\chi_1}+(n^{\chi_2}+n^{\chi_3})}+[w_2\frac{n^{\chi_2}}{n^{\chi_2}+n^{\chi^3}}+w_3\frac{n^{\chi_3}}{n^{\chi_2}+n^{\chi^3}}]\frac{n^{\chi_2}+n^{\chi^3}}{n^{\chi_1}+(n^{\chi_2}+n^{\chi_3})}$$

and the corresponding weight of x_i in χ' is a combination of its weights in $stack(\chi_1,\chi_2)$ and χ_3:

$$w_i^{\chi'}=[w_1\frac{n^{\chi_1}}{n^{\chi_1}+n^{\chi_2}}+w_2\frac{n^{\chi_2}}{n^{\chi_1}+n^{\chi_2}}]\frac{n^{\chi_1}+n^{\chi_2}}{(n^{\chi_1}+n^{\chi_2})+n^{\chi_3}}+w_3\frac{n^{\chi_3}}{(n^{\chi_1}+n^{\chi_2})+n^{\chi_3}}.$$

But the previous two expressions are equal and the proof is complete.

■

This lemma implies that we can use the notation $stack(\chi_1,\cdots,\chi_m)$.

Definition of quantiles and DOS for weighted vectors

Now let us get to the definition of quantiles. We can proceed exactly in the same way as we did before by having in mind a bar of length one. Or alternatively, we can apply the quantile function definition for usual distributions to the empirical distribution of a weighted vector F_{χ}. This time, we proceed in a slightly different fashion which is equivalent to these

methods. Suppose $\chi = (x, w^\chi, n^\chi)$ is given and $\zeta = comp(\chi) = (z, w^\zeta, n^\chi)$. We assume z has length l_z. First, we define

$$lqind_\chi : (0,1] \to \{1, 2, \cdots, l_z\},$$

and

$$rqind_\chi : [0,1) \to \{1, 2, \cdots, l_z\},$$

the "left quantile index" and "right quantile index" functions and then define the left and right quantile functions using the index functions. If $\zeta = comp(\chi) = (z, w^\zeta, n^x)$ then we define

$$lq_\chi(p) = z_{lqind_\chi(p)},\ p \in (0,1],\ lq_\chi(p) = -\infty,\ p = 0,$$

and

$$rq_\chi(p) = z_{rqind_\chi(p)},\ p \in [0,1),\ rq_\chi(p) = \infty,\ p = 1.$$

Let $\zeta = comp(\chi)$. $lqind_\chi$ and $rqind_\chi$ are defined as follows:

- $p = 0$ then $lqind_\chi(p)$ not defined and $rqind_\chi(p) = 1$.
- $0 < p < w_1^\zeta$ then $lqind_\chi(p) = rqind_\chi(p) = 1$.
- $p = w_1^\zeta$ then $lqind_\chi(p) = 1$ and $rqind_\chi(p) = 2$.
  $\vdots$
- $w_1^\zeta + \cdots + w_{i-1}^\zeta < p < w_1^\zeta + \cdots + w_i^\zeta$ then $lqind_\chi(p) = rqind_\chi(p) = i$.
- $p = w_1^\zeta + \cdots + w_i^\zeta$ then $lqind_\chi(p) = i, rqind_\chi(p) = i + 1$.
  $\vdots$
- $p = 1$ then $lqind_\chi(p) = l_z$ and $rqind_\chi$ is not defined.

Remark. It is easy to see that $\chi \sim \xi$ then $lq_\chi = lq_\xi, rq_\chi = rq_\xi$.

Remark. For ordinary vectors, this is equivalent to the definition given in the previous sections.

Remark. Consider the natural distribution function F_χ corresponding to a weighted vector χ then $lq_\chi = lq_{F_\chi}$ and $rq_\chi = rq_{F_\chi}$. Hence, lq_χ, rq_χ satisfy all the properties proved for left and right quantile functions of a distribution function.

Definition We generalize the degree of separation (probability loss function) δ_χ on the set of weighted vectors as follows:

$$\delta_\chi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+ \cup \{0\},$$

$$\delta_\chi(z', z) = \delta_\chi(z, z') = \sum_{z < x_j < z'} w_j^\chi,\ z < z',$$

and $\delta_\chi(z, z) = 0$.

Lemma 8.2.2 *(Properties of the probability loss function for weighted vectors)*
a) $\delta_\chi = \delta_{F_\chi}$.
b) δ_χ *only depends on comp*(χ).
c) δ_χ *satisfies the pseudo–triangle property.*

Proof a) and b) are trivial and c) follows from a) and pseudo–triangle property for the probability loss functions for distributions.

■

8.2.1 Partition operator

This section introduces the partition operator to partition data into arbitrarily sized partitions. This allows us to address the two remaining properties for quantiles we pointed out in the introduction (in Lemma 8.2.5). The idea behind the definition of the partition operator can be explained as follows. Suppose a weighted vector $\chi = (x, w^\chi, n^\chi)$ is given and we want to partition it to smaller vectors with weights $(p_1, \cdots, p_m), \sum_{i=1}^m p_i = 1$. Consider a bar of length 1 and then color it from left to right using colors corresponding to the x_i with length w_i^χ. Then cut the bar from left to right using the given weights $(p_1, \cdots, p_m)$. Now each one of the small bars is the partitions we needed. More formally, we have the following definition:

Definition Suppose $\mathcal{P} = (p_1, p_2, \cdots, p_m)$ is given, such that $\sum_{i=1}^m p_i = 1$. Then a $\mathcal{P}$-partition of a weighted data vector $\chi = (x, w^\chi, n^\chi)$ is denoted

by $part(\mathcal{P},\chi) = (\chi^1,\cdots,\chi^m)$ and is a collection of m weighted vectors $\chi^1 = (x^1, w^{\chi^1}, n^{\chi^1} = n^\chi.p_1),\cdots,\chi^m = (x^m, w^{\chi^m}, n^{\chi^m} = n^\chi.p_m)$ defined as follows:

1. $x^1 = (x_{s_1},\cdots,x_{t_1}),\ s_1 = 1, v_1 = \sum_{1\le j\le t_1} w_j \ge p_1,\ \sum_{1\le j<t_1} w_j < p_1$

2. $x^2 = (x_{s_2},\cdots,x_{t_2}),\ v_2 = \sum_{1\le j\le t_2} w_j - p_1 \ge p_2,\ \sum_{1\le j<t_2} w_j - p_1 < p_2,\ s_2 = \begin{cases} t_1+1 & v_1 = p_1 \\ t_1 & v_1 > p_1 \end{cases}$

$\vdots$

k. $x^k = (x_{s_k},\cdots,x_{t_k}),\ v_k = \sum_{1\le j\le t_k} w_j - \sum_{j=1}^{k-1} p_j \ge p_k,\ \sum_{1\le j<t_2} w_j - \sum_{j=1}^{k-1} p_j < p_k,\ s_k = \begin{cases} t_{k-1}+1 & v_{k-1} = p_{k-1} \\ t_{k-1} & v_{k-1} > p_{k-1} \end{cases}$.

$\vdots$

The corresponding weight vectors and data amounts are defined as:

1. $w^{\chi^1} = \frac{1}{p_1}(w^\chi_{s_1}, w^\chi_{s_2},\cdots,w^\chi_{t_1} - (v_1 - p_1))$,

$\vdots$

k. $w^{\chi^k} = \begin{cases} \frac{1}{p_k}(w^\chi_{s_k}, w^\chi_{s_k+1},\cdots,w^\chi_{t_k} - (v_k - p_k)) & v_{k-1} = p_{k-1} \\ \frac{1}{p_k}(v_{k-1} - p_{k-1}, w^\chi_{s_k+1},\cdots,w^\chi_{t_k} - (v_k - p_k)) & v_{k-1} > p_{k-1} \end{cases}$.

$\vdots$

Lemma 8.2.3 *If $\chi = (x, w^\chi, n^\chi)$ is an ordinary vector and $l_x = n^\chi = n_1 + \cdots + n_m$. Let $\mathcal{P} = (\frac{n_1}{n^\chi},\cdots,\frac{n_m}{n^\chi})$ then the $\mathcal{P}$-partition of χ is simply obtained by starting from the left and partitioning x to vectors of length $n_1, n_2,\cdots,n_m$.*

Proof This is a straightforward conclusion of the definition. ■

Lemma 8.2.4 *Suppose $\chi = (x, w^\chi, n^\chi)$ is partitioned by some $\mathcal{P} = (p_1,\cdots,p_m)$ to $\chi^1,\cdots,\chi^m$ then*

$$stack(\chi^1,\cdots,\chi^m) \sim \chi.$$

Proof Let $\chi' = stack(\chi^1,\cdots,\chi^m)$ and suppose $\chi' = (x', w^{\chi'}, n^{\chi'})$. Then clearly x' and x have the same distinct elements. (Although it might be the

case that $x' \neq x$ since some elements of x are repeated more than once in x.) Also

$$n^{\chi'} = \sum_{i=1}^{m} p_i n^{\chi} = n^{\chi}.$$

In order to show that for z an element of the vector x, its corresponding weight is equal in χ and χ', suppose z is equal to $x_{i_1}, \cdots, x_{i_r}$ in x with corresponding weights $w^{\chi}_{i_1}, \cdots, w^{\chi}_{i_r}$. Then the weight corresponding to z in χ is equal to $\sum_{k=1}^{r} w^{\chi}_{i_k}$. Now note that any of $x_{i_k}, k = 1, \cdots, r$, corresponds to one or two elements in $stack(\chi^1, \cdots, \chi^m)$ by the definition of the partitions operator. It can be the case that x_{i_k} only appears in χ^s or in χ^s, χ^{s+1} if x_{i_k} is at the end of the partition χ^s and at the beginning of the next. In the first case when x_{i_k} only appears in χ^s, its weight in χ^s will be $\frac{1}{p_s} w^{\chi}_{i_k}$ and hence its weight contribution in $stack(\chi^1, \cdots, \chi^m)$ will be $\frac{n^{\chi}.p_s}{n^{\chi}} \frac{1}{p_s} w^{\chi}_{i_k} = w^{\chi}_{i_k}$. In the second case its weight in χ^s will be $\frac{1}{p_s}(w^{\chi}_{i_k} - (v_s - p_s))$ and in χ^{s+1} will be $\frac{1}{p_{s+1}}(v_s - p_s)$. Hence its weight contribution in $stack(\chi^1, \cdots, \chi^m)$ coming from χ^s, χ^{s+1} is $\frac{n^{\chi} p_s}{n^{\chi}} \frac{1}{p_s}(w^{\chi}_{i_k} - (v_s - p_s)) + \frac{n^{\chi} p_{s+1}}{n^{\chi}} \frac{1}{p_{s+1}}(v_s - p_s) = w^{\chi}_{i_k}$. Summing up all the weights in $stack(\chi^1, \cdots, \chi^m)$, we get the same value of $\sum_{k=1}^{r} w^{\chi}_{i_k}$.

■

Using the partition operator, we can easily define the cut operator as follows.

Definition Let $D = \{(a, b) |\ a, b \in (0, 1),\ a < b\}$. Then $cut : \Upsilon \times D \to \Upsilon$ is defined to be

$$cut(\chi, p_1, p_2) = \chi^2,$$

where χ^2 is the second component of $part(\mathcal{P}, comp(\chi)) = (\chi^1, \chi^2, \chi^3)$, the result of applying a partition operator with weights $\mathcal{P} = (p_1, p_2 - p_1, 1 - p_2)$ to $comp(\chi)$. We also define left cut and right cuts,

$$lcut, rcut : (0, 1) \to \mathbb{R},$$

$$lcut(\chi, p) = \chi^1, rcut(\chi, 1 - p) = \chi^2,$$

where χ^1 and χ^2 are the first and second component of the partition of χ by $\mathcal{P} = (p, 1 - p)$.

Lemma 8.2.5 *Suppose $\chi = (x, w^{\chi}, n^{\chi})$ is a weighted vector and (p_1, p_2) in D. Then*

a) The amount of data in $cut(\chi, p_1, p_2)$ *is* $n^\chi(p_2 - p_1)$.
b) $cut(\chi, p_1, p_2)$ *starts with* $rq_\chi(p_1)$ *and ends with* $lq_\chi(p_2)$.
c) The vector of $lcut(\chi, p)$ *ends with* $lq_\chi(p)$.
d) The vector of $rcut(\chi, p)$ *starts with* $rq_\chi(1-p)$.
e) Suppose $p_1, p_2 \in (0,1)$ *then* $lcut(lcut(\chi, p_1), p_2) = lcut(\chi, p_1 p_2)$.
f) Suppose $p_1, p_2 \in (0,1)$ *then* $rcut(rcut(\chi, p_1), p_2) = rcut(\chi, p_1 p_2)$.

Proof a) is trivial. To prove b), consider the definition of the partition operator as given in Definition 8.2.1 for arbitrary $\mathcal{P} = (p'_1, \cdots, p'_m)$. For the first partition, $x_{s_1} = x_1 = lq_\chi(p'_1)$ and for x_{t_1}, we have

$$\sum_{1 \le j \le t_1} w_j \ge p'_1, \text{ and } \sum_{1 \le j < t_1} w_j < p'_1,$$

which concludes $lq_\chi(p'_1) = x_{t_1}$. For the k-th partition,

$$s_k = \begin{cases} t_{k-1} + 1 & v_{k-1} = p'_{k-1} \\ t_{k-1} & v_{k-1} > p'_{k-1} \end{cases}.$$

If $v_{k-1} = p'_{k-1}$, then $\sum_{1 \le j \le t_{k-1}} w_j = \sum_{i=1}^{k-1} p'_i$. Hence $rq_\chi(\sum_{i=1}^{k-1} p'_i) = x_{t_{k-1}+1} = x_{s_k}$. For t_k, we have $\sum_{1 \le j < t_k} w_j < \sum_{i=1}^{k} p'_i$ and $\sum_{1 \le j \le t_k} w_j \le \sum_{i=1}^{k} p'_k$. Hence $lq_\chi(\sum_{i=1}^{k} p'_k) = x_{t_k}$. To finish the proof, let $m = 3$ and $p'_1 = p_1, p'_2 = p_2 - p_1, p'_3 = 1 - p_2$ and note that $cut(\chi, p_1, p_2)$ corresponds to the second component of the partition operator of $\mathcal{P} = (p'_1, p'_2, p'_3)$ on χ.
The proof of c) is similar to b). d) can be either done by a similar direct proof or by using the Quantile Symmetry Theorem.
To prove e) let

$$\chi^1 = lcut(\chi, p_1) = ((x_1, \cdots, x_{t_1}), (w_1^1, \cdots, w_{t_1}^1), n^\chi.p_1)$$

$$\chi^{1,2} = lcut(lcut(\chi, p_1), p_2) = ((x_1, \cdots, x_{t_{1,2}}), (w_1^{1,2}, \cdots, w_{t_1}^{1,2}), n^\chi.p_1.p_2)$$

$$\chi^{12} = lcut(\chi, p_1 p_2) = ((x_1, \cdots, x_{t_{12}}), (w_1^{12}, \cdots, w_{t_1}^{12}), n^\chi.p_1.p_2)$$

We want to show $\chi^{1,2} = \chi^{12}$. It is clear that their data amount is equal. By applying the definition of *lcut* to the above three equations, we conclude the following:

$$\sum_{1 \le j < t} w_j < p_1, \quad \sum_{1 \le j \le t} w_j \ge p_1, \tag{8.1}$$

$$\sum_{1 \le j < t_{1,2}} w_j^1 < p_2, \quad \sum_{1 \le j \le t_{1,2}} w_j^1 \ge p_2, \tag{8.2}$$

$$\sum_{1\leq j<t_{12}} w_j < p_1p_2, \sum_{1\leq j\leq t_{12}} w_j \geq p_1p_2. \tag{8.3}$$

If $j < t_1$ then $w_j^1 = \frac{1}{p_1}w_j$. Hence from the first equation in 8.2, we conclude

$$\frac{1}{p_1}\sum_{1\leq j<t_{1,2}} w_j < p_2 \Rightarrow \sum_{1\leq j<t_{1,2}} w_j < p_1p_2.$$

Now consider two cases:
Case I: $t_{1,2} < t_1$. In this case, similarly, from the second equation in 8.2, we conclude

$$\frac{1}{p_1}\sum_{1\leq j\leq t_{1,2}} w_j \geq p_2 \Rightarrow \sum_{1\leq j\leq t_{1,2}} w_j \geq p_1p_2.$$

Case II: $t_{1,2} = t_1$. In this case note that for $j < t_{1,2} = t_1$, we still have $w_j^1 = \frac{1}{p_1}w_j$ and for $j = t_{1,2} = t$, we have $w_j^1 \leq \frac{1}{p_1}w_j$. But

$$\sum_{1\leq j\leq t_{1,2}=t} w_j^1 = 1 \Rightarrow \sum_{1\leq j\leq t_{1,2}=t} w_j^1 \geq p_1 \geq p_1p_2.$$

In both cases, we showed that $\sum_{1\leq j\leq t_{1,2}} w_j \geq p_1p_2$ and $\sum_{1\leq j\leq t_{1,2}=t_1} w_j^1 \geq p_1 \geq p_1p_2$. We conclude that $t_{1,2} = t_{12}$. In order to show that the weight vectors of $\chi^{1,2}$ and χ^{12} are the same, note that they have the same length. We only need to show that they match on all the components except for the last one because the equality of the last one will follow. But if $j < t_{1,2} = t_{12}$ then $w_j^{1,2} = \frac{1}{p_2}(\frac{1}{p_1}w_j)$ and $w_j^{12} = \frac{1}{p_1p_2}w_j$.
f) can be done either by a similar argument as e) or using the Quantile Symmetry Theorem.

■

Remark. Part a) and e) address the two remaining properties we were seeking in the introduction.

8.2.2 Quantile data summaries

Here, we formally define quantile data summaries. They arise when a large data vector is summarized by a smaller vector and possibly some other information about the original vector and how the summary is been created. A large vector might have been partitioned into smaller vectors and the smaller vectors might have been summarized. First we define a probability index vector which is needed to define quantile data summaries.

Definition A vector $\mathcal{P} = (p_1, \cdots, p_k)$ is called a probability index vector if $0 \leq p_1 < \cdots < p_k \leq 1$.

Definition Suppose $\chi = (x, w^\chi, n^\chi)$, a weighted vector and a probability index vector $\mathcal{P} = (p_1, \cdots, p_m)$ is given such that $0 \leq p_1 < p_2 < \cdots < p_m \leq 1$. Then a $\mathcal{P}$-quantile summary of χ is defined to be

$$qs(\mathcal{P}, \chi) = (lq_\chi(p_1), \cdots, lq_\chi(p_m)).$$

Definition A summary triple is defined to be a triple $(qs(\mathcal{P}, \chi), \mathcal{P}, n^\chi)$, where qs is the summarized vector as defined above, $\mathcal{P}$ is the summary probability index vector and n^χ is the data amount of the original vector.

We also define an ϵ-summary for $\epsilon < 1/2$.

Definition Let $h = [1/\epsilon]$. Then the ϵ-summary for χ is defined to be the triple $(qs(\epsilon, \chi), \epsilon, n^\chi)$:

$$qs(\epsilon, \chi) = (lq_\chi(\epsilon), lq_\chi(2\epsilon), \cdots, lq_\chi((h-1)\epsilon)).$$

Note that $[0, \epsilon), [\epsilon, 2\epsilon), \cdots, [(h-1)\epsilon, 1]$ is a partition of [0,1] to intervals of the same length ϵ other than the last one, which can be greater than ϵ. However it is less than 2ϵ. If $\epsilon = 1/s$ for a natural number s, then the $1/s$ summary is going to be

$$qs(1/s, \chi) = (lq_\chi(1/s), lq_\chi(2/s), \cdots, lq_\chi((s-1)/s)).$$

Remark. For an ordinary vector $x = (x_1, \cdots, x_n)$, suppose $n = n_1 n_2$. Then we defined the n_2–coarsening operator to be

$$C_{n_2}(x) = (lq_x(p_1), \cdots, lq_x(p_{n_1 - 1}),$$

where $p_i = i/n,\ i = 1, \cdots, n_1 - 1$. This is the same as

$$qs(\epsilon, x),$$

for $\epsilon = 1/n_1$. Hence the coarsening operator is a special case of creating an ϵ-summary.

We also define summary lists.

Definition Suppose $\chi = stack(\chi^1, \cdots, \chi^m)$ and m probability index vectors $\mathcal{P}_1, \cdots, \mathcal{P}_m$ are given. Then let $\xi^i = qs(\chi^i, \mathcal{P}_i)$. Then the list

$$\xi = \begin{pmatrix} \xi_1 & \mathcal{P}_1 & n^{\chi^1} \\ \vdots & \vdots & \vdots \\ \xi_m & \mathcal{P}_m & n^{\chi^m} \end{pmatrix}$$

is called a quantile summary list of χ. Note that ξ is not a matrix in general since the length of the summary indices might differ.

Quantile summary vectors or quantile summary lists are to be used to infer the original vector χ. They can be used as "inputs" to procedures for approximating lq_χ. The formal definition of a data summary procedure is defined below.

Definition Suppose χ is a weighted vector and *input* is a quantile summary list. Then a quantile summary procedure is defined to be a left quantile function:

$$proc(input, \chi) : [0, 1] \to \mathbb{R}.$$

"proc" tries to approximate the quantiles of the original vector χ using the input. It is desirable to find procedures that have good accuracy.

Example The d-coarsening algorithm can be viewed as an example of the above framework. There the vector χ is simply an ordinary vector of length n which is a concatenation of $x^1, \cdots, x^l$. The summary list consists of d-coarsening of partitions $x^1, \cdots, x^l$. In other words $x^1, \cdots, x^m$ which are of length $l_i = c_i d$ are summarized by $\mathcal{P}_i = (1/c_i, \cdots, (c_i - 1)/c_i,\ i = 1, \cdots, m)$ to $w^1, \cdots, w^m$. Finally the "proc" is simply the left quantile function of the concatenation of $w^1, \cdots, w^m$. The accuracy in terms of the probability loss was bounded by $\epsilon = \frac{m+1}{C-m}$, $C = \sum_{i=1}^m c_i$. In other words

$$\sup_{p \in (0,1)} \delta_x(proc(input, x)(p), lq_x(p)) \leq \epsilon.$$

8.3 Optimal probability indices for vector data summaries

Suppose a data vector x or a distribution X is given. The data vector x might be too long to carry around or save in the memory. Similarly the distribution X might be too complicated or unknown. To make inferences about a data vector x or the distribution of X, we might use a summary or

some other procedure. For example, we might save a vector data summary instead of the vector x of length n, where n is very large:

$$qs(\mathcal{P}, x) = (lq_x(p_1), \cdots, lq_x(p_m)),\ p_1 < \cdots < p_m.$$

The following question motivates our ensuing development:
Question: How should $\mathcal{P} = \{p_1, \cdots, p_m\}$ be chosen to provide good approximation/prediction to x (or X)?

A natural way to approximate x or X is to estimate all the quantiles. (This is equivalent to approximating or estimating the whole data vector x or the distribution function of X.) We are given an input. In the case of a data vector it is usually a quantile data summary and in the case of the random variable X it might be a random sample. Then a "procedure" can be employed to approximate/estimate the quantiles of x or X. For any given p the left quantile $lq_x(p)$ or $lq_X(p)$ is approximated/estimated by the procedure using the input. We denote this value by $proc(input, x)(p)$ or $proc(input, X)(p)$. Then a loss L can be used to assess the goodness of such a procedure:

$$L(proc(input, x)(p), lq_x(p)).$$

To assess the overall goodness of such a procedure, we can use either the sup loss or the integral loss:

$$\sup_{p\in[0,1]} L(proc(input, x)(p), lq_x(p)),$$

or

$$\int_{p\in[0,1]} L(proc(input, x)(p), lq_x(p))dp.$$

For simplicity, we restrict to data vectors from here. We use the probability loss δ_x as the most natural choice. We want to minimize this loss n order to find optimal ways to summarize data (create input) and find optimal procedures.

Definition We define the crudity of the procedure *proc* at p given the *input* to be

$$crud(proc(input, x)(p)) = \delta_x(proc(input, x)(p), lq_x(p)).$$

Also the "sup crudity" and "integral crudity" are respectively given by

$$SC(proc(input, x)) = \sup_{p\in[0,1]} \delta_x(proc(input, x)(p), lq_x(p)),$$

and

$$IC(proc(input, x)) = \int_{p\in[0,1]} \delta_x(proc(input, x)(p), lq_x(p))dp.$$

Using the above framework, we look for good procedures to summarize data vectors and later distribution functions.

A quantile data summary was defined to be

$$qs(\mathcal{P}, x) = (lq_x(p_1), \cdots, lq_x(p_m)),\ p_1 < \cdots < p_m,$$

for a probability index vector $\mathcal{P} = (p_1, \cdots, p_m)$. There is a natural procedure associated with this input that is a quantile data summary, which we define below.

Definition Suppose x is a data vector which has been summarized by $\mathcal{P} = (p_1, \cdots, p_m)$. Then we define the shortest distance quantile procedure of x associated with $\mathcal{P}$ to be

$$proc(input, x)(p) = lq_x(p_i),\ i = \underset{j}{argmin}\{|p - p_j|, j = 1, \cdots, m\}.$$

If there were more than one minimum above, take the smaller value. We denote this procedure by $shproc(x, \mathcal{P})$.

The shortest distance procedure be specified by the notation "$\mapsto$" as shown below:
1. $0 \le p \le p_1 + \frac{p_2-p_1}{2} \mapsto lq_x(p_1)$.
2. $p_1 + \frac{p_2-p_1}{2} < p \le p_2 + \frac{p_3-p_2}{2} \mapsto lq_x(p_2)$.
$\vdots$
m. $p_{m-1} + \frac{p_m-p_{m-1}}{2} < p \le 1 \mapsto lq_x(p_m)$.
The largest loss in the first part of the procedure is the maximum of the two values,

$$\delta_x(lq_x(0), lq_x(p_1)), \delta_x(lq_x(p_1), lq_x(p_1 + \frac{p_2 - p_1}{2})). \tag{8.4}$$

For the second part, it is the maximum of

$$\delta_x(lq_x(p_1 + \frac{p_2 - p_1}{2}), lq_x(p_2)), \delta_x(lq_x(p_2), lq_x(p_2 + \frac{p_3 - p_2}{2})). \tag{8.5}$$

For the m-th part it is the maximum of

$$\delta_x(lq_x(p_{m-1} + \frac{p_m - p_{m-1}}{2}), lq_x(p_m)), \delta_x(lq_x(p_m), lq_x(1)). \tag{8.6}$$

We use quantile data summaries to save space and memory for operations on very large datasets. Hence, we have a limitation on m. The interesting question is what is an optimal index set $\mathcal{P}$ of length m to summarize data vectors? In the beginning, we usually do not have any information about x so the $\mathcal{P}$ should be chosen in way that works well for all possible data vectors. Hence, we settle for either

$$\underset{\mathcal{P}}{argmin} \sup_x SC(shproc(input, x)(p), lq_x(p)) =$$

$$\underset{\mathcal{P}}{argmin} \sup_x \sup_{p\in[0,1]} \delta_x(shproc(input, x)(p), lq_x(p)),$$

or

$$\underset{\mathcal{P}}{argmin} \sup_x IC(shproc(input, x)(p), lq_x(p)) =$$

$$\underset{\mathcal{P}}{argmin} \sup_x \int_0^1 \delta_x(shproc(input, x)(p), lq_x(p))dp.$$

We sort out the sup crudity case first. By Lemma 6.6.3, taking the sup of the max over all x in Equations 8.4, 8.5 and 8.6, we get the maximum of the following quantities:

1. p_1, $\frac{p_2-p_1}{2}$.
2. $\frac{p_2-p_1}{2}$, $\frac{p_3-p_2}{2}$.
3. $\frac{p_3-p_2}{2}$, $\frac{p_4-p_3}{2}$.

$\vdots$

m. $\frac{p_m-p_{m-1}}{2}$, $1 - p_m$.

Hence,

$$\sup_x \sup_{p\in[0,1]} \delta_x(shproc(input, x)(p), lq_x(p)) =$$

$$\max_{p\in[0,1]}\{p_1, \frac{p_2 - p_1}{2}, \frac{p_2 - p_1}{2}, \frac{p_3 - p_2}{2}, \frac{p_3 - p_2}{2}, \frac{p_4 - p_3}{2}, \cdots, \frac{p_m - p_{m-1}}{2}, 1 - p_m\}.$$

After omitting the repetitions, we need to minimize:

$$\max\{p_1, \frac{p_2 - p_1}{2}, \frac{p_3 - p_2}{2}, \frac{p_4 - p_3}{2}, \cdots, \frac{p_m - p_{m-1}}{2}, 1 - p_m\},$$

over all $p_1 < p_2 < \cdots < p_m \in [0,1]$. We claim that

$$p_1 = \frac{1}{2m}, p_2 - p_1 = 1/m, p_3 - p_2 = 1/m, \cdots, p_{m-1} = 1/m, p_m = 1 - \frac{1}{2m},$$

is the solution. Note that in this case the max is equal to $1/2m$. We show that we cannot do better. Let

$$\begin{aligned}
\alpha_1 &= p_1, \\
\alpha_2 &= \frac{p_2 - p_1}{2}, \\
\alpha_3 &= \frac{p_3 - p_2}{2}, \\
&\vdots \\
\alpha_m &= \frac{p_m - p_{m-1}}{2}, \\
\alpha_{m+1} &= p_m.
\end{aligned}$$

We have $\alpha_1 + 2\alpha_2 + \cdots + 2\alpha_m + \alpha_{m+1} = 1$. The α_i are non-negative, there are $1 + 2(m-2) + 1$ of them (counting the ones with multiple 2 two times) and they sum up to 1. If all of them are less than $\frac{1}{2m}$ the sum will be less than 1. Hence we conclude the maximum is obtained when they are all equal to $1/2m$.

Now let us do the integral crudity case. We claim the solution is the same. We compute the integral in the following, using 6.6.4 in the second equality:

$$\begin{aligned}
\sup_x \int_0^1 \delta_x(lq_x(p), shproc(input, x)(p))dp &= \\
\sup_x [\int_0^{p_1 + \frac{p_2 - p_1}{2}} \delta_x(lq_x(p_1), lq_x(p))dp &+ \int_{p_1 + \frac{p_2 - p_1}{2}}^{p_2 + \frac{p_3 - p_2}{2}} \delta_x(lq_x(p_2), lq_x(p))dp \\
&+ \cdots + \int_{p_{m-1} + \frac{p_m - p_{m-1}}{2}}^1 \delta_x(lq_x(p_m), lq_x(p))dp]
\end{aligned}$$

$$
\begin{aligned}
&= \int_0^{p_1+\frac{p_2-p_1}{2}} |p-p_1|dp + \int_{p_1+\frac{p_2-p_1}{2}}^{p_2+\frac{p_3-p_2}{2}} |p-p_2|dp + \cdots + \int_{p_{m-1}+\frac{p_m-p_{m-1}}{2}}^{1} |p-p_m|dp \\
&= \int_0^{p_1} (p_1-p)dp + \int_{p_1}^{p_1+\frac{p-p_1}{2}} (p-p_1)dp + \int_{p_1+\frac{p_2-p_1}{2}}^{p_2} (p_2-p)dp \\
&+ \int_{p_2}^{p_2+\frac{p_3-p_2}{2}} (p-p_2)dp + \cdots + \int_{p_{m-1}+\frac{p_m-p_{m-1}}{2}}^{p_m} (p_m-p)dp + \int_{p_m}^{1} (p-p_m)dp \\
&= \int_0^{p_1} pdp + \int_0^{\frac{p_2-p_1}{2}} pdp + \int_0^{\frac{p_2-p_1}{2}} pdp \\
&+ \int_0^{\frac{p_3-p_2}{2}} pdp + \cdots + \int_0^{\frac{p_m-p_{m-1}}{2}} pdp + \int_0^{1-p_m} pdp \\
&= (1/2)\alpha_1^2 + \alpha^2 + \cdots + \alpha_m^2 + (1/2)\alpha_{m+1}^2 = (1/2)(\alpha_1^2 + 2\alpha^2 + \cdots + 2\alpha_m^2 + \alpha_{m+1}^2),
\end{aligned}
$$

where $\alpha_1 = p_1, \alpha_2 = \frac{p_2-p_1}{2}, \cdots, \alpha_m = \frac{p_m-p_{m-1}}{2}, \alpha_{m+1} = 1-p_m$.
We have the restriction $\alpha_1 + 2\alpha_2 + \cdots + 2\alpha_m + \alpha_{m+1} - 1 = 0$ and $\alpha_i \geq 0$.
In order to minimize

$$\alpha_1^2 + 2\alpha_2^2 + \cdots + 2\alpha_m^2 + \alpha_{m+1}^2,$$

we use Lagrange Multiplier's Method. Let

$$f(x_1, \cdots, x_{m+1}) = x_1^2 + 2x_2^2 \cdots + 2x_m^2 + x_{m+1}^2 - \lambda(x_1 + 2x_2 + \cdots + 2x_m + x_{m+1} - 1).$$

Taking the partial derivatives and putting them equal to zero, we get:

$$
\begin{aligned}
\frac{\partial g}{\partial x_1} &= 2x_1 - \lambda = 0, \\
\frac{\partial g}{\partial x_2} &= 4x_2 - 2\lambda = 0, \\
&\vdots \\
\frac{\partial g}{\partial x_m} &= 4x_m - 2\lambda = 0, \\
\frac{\partial g}{\partial x_{m+1}} &= 2x_{m+1} - \lambda = 0.
\end{aligned}
$$

By summing up the equations we get:

$$2(x_1 + 2x_2 + \cdots + 2x_m + x_{m+1}) - 2\lambda(m-1) - 2\lambda = 2 - 2\lambda(m-1) - 2\lambda = 0.$$

Hence $\lambda = \frac{1}{m}$. This gives $x_i = \frac{1}{2m}$. Hence $p_1 = p_m = \frac{1}{2m}$ and $p_2 - p_1 = \cdots = p_m - p_{m-1} = \frac{1}{m}$.

8.4 Other loss functions

It is well-known that

$$\underset{a}{argmin} E_X(X-a)^2,$$

is the mean, when it exists. This fact is used in classical statistics for estimation of parameters and regression. It is also widely claimed that

$$\underset{a}{argmin} E_X|X-a|,$$

is "the median". In particular for data vectors $x = (x_1, \cdots, x_n)$, this will take the form

$$\underset{a}{argmin} \frac{1}{n}\sum_{i=1}^{n}|x_i - a|.$$

It is not clear what is meant by "the median"? For data vectors does it mean that the classic median (the middle value when there is odd number of elements and the average of the two middle values otherwise) is the unique solution? In general, is the answer unique? What is the connection of the solution to the left and right quantiles? We provide answers to some of these questions in the following theorem.

Theorem 8.4.1 *Suppose X is a random variable and $E|X-a|$ is finite for some $a \in \mathbb{R}$ then*

$$\underset{a}{argmin} E|X-a| = [lq_X(\frac{1}{2}), rq_X(\frac{1}{2})].$$

Proof

$$E|X-a| = \int_{\mathbb{R}} |X-a|dP = \int_{X>a}(X-a)dP + \int_{X<a}(a-X)dP.$$

We prove the theorem in three steps:
1. If $a < lq_X(1/2)$ then $E|X-a| > E|X-lq_X(1/2)|$.
2. If $a > rq_X(1/2)$ then $E|X-a| > E|X-rq_X(1/2)|$.
3. If $lq_X(1/2) \leq a, b \leq rq_X(1/2)$ then $E|X-a| = E|X-b|$.

Step **1**. Let $b = lq_X(1/2)$ and $\epsilon = b - a > 0$. Then

$$\begin{aligned} E|X-b| = & \int_{X\geq b}(X-b)dP + \int_{X<b}(b-X)dP \\ = & \int_{X\geq b}(X-a-\epsilon)dP + \int_{X<b}(a+\epsilon-X)dP \\ \leq & \int_{X\geq b}|X-a|-\epsilon dP + \int_{X<b}(|X-a|+\epsilon)dP \\ = & E|X-a| - \epsilon(P(X\geq b) - P(X<b)). \end{aligned}$$

But $P(X \geq b) - P(X < b)$ is non-negative since $P(X < lq_X(1/2)) \leq 1/2$. Hence $E|X-b| \leq E|X-a|$. To show that the equality cannot happen take $a < a' < b$ and let $\epsilon' = a' - a$ then

$$\begin{aligned} E|X-a'| = & \int_{X\geq a'}(X-a')dP + \int_{X<a'}(a'-X)dP \\ = & \int_{X\geq a'}(X-a-\epsilon')dP + \int_{X<a'}(a+\epsilon'-X)dP \\ \leq & \int_{X\geq a'}|X-a|-\epsilon' dP + \int_{X<a}(|X-a|+\epsilon')dP \\ = & E|X-a| - \epsilon'(P(X\geq a') - P(X<a')). \end{aligned}$$

But $P(X \geq a') - P(X < a')$ is positive since $P(X < a') < 1/2$ and $a' < lq_X(p) \Rightarrow P(X < a') < 1/2$. Hence $E|X-a'| < E|X-a|$. But also since $a' < b$, we have

$$E|X-b| \leq E|X-a'| < E|X-a|.$$

Step **2.** For $a > rq_X(1/2) = c$ one can either repeat a similar argument to that in Step 1 or use the Quantile Symmetry Theorem as we do here. Consider the random variable $-X$. Then

$$a > rq_X(1/2) \Rightarrow -a < -rq_X(1/2) = lq_{-X}(1/2)$$

Now since $-a < -c = lq_{-X}(1/2)$ by applying Step 1 to $-X$, we get

$$E|-X-(-c)| < E|-X-(-a)| \Rightarrow E|X-a| < E|X-c|.$$

Step **3.** If $lq_X(1/2) = rq_X(1/2)$ the result is trivial. Otherwise let $b = lq_X(1/2) < rq_X(1/2) = c$ and $a < a' \in [b, c]$. By Lemma 5.3.1 if $lq_X(p) < rq_X(p)$. So $P(X \leq lq_X(p)) = p$ and $P(X \geq rq_X(p)) = 1 - p$. Hence $P(X \leq b) = P(X \geq c) = 1/2$. Let $\epsilon = a' - a$. Then

$$\begin{aligned} E|X-a| &= \int_{b<X<c} |X-a|dP + \int_{X\geq c}(X-a)dP + \int_{X\leq b} +\epsilon/2 - \epsilon/2 \\ &= \int_{X\geq c}(X-a-\epsilon)dP + \int_{X\leq b}(a-X+\epsilon)dP \\ &= \int_{X\geq c}(X-a')dP + \int_{X\leq b}(a'-X)dP \\ &= E|X-a'|. \end{aligned}$$

■

Corollary 8.4.2 *Suppose F_X is continuous and $\exists a \in \mathbb{R}$, $E|X-a| < \infty$. Then*

$$\underset{a}{argmin} E|X-a| = \{a|F(a) = 1/2\}.$$

Proof Note that if F is continuous $F(a) = p \Leftrightarrow a \in [lq_X(p), rq_X(p)]$, by Lemma 5.5.2. ■

Now let us find

$$\underset{a}{argmin} E(\delta_F(X,a)).$$

We solve the problem for continuous variables only here and leave the general case as an interesting open problem. Our conjecture is that the same result holds in general.

Lemma 8.4.3 *Suppose X be a random variable with continuous distribution function F. Then*

$$\underset{a}{argmin} E(\delta_F(X,a)) = [lq_X(1/2), rq_X(1/2)].$$

Proof If F is continuous then $F(X) \sim U(0,1)$. Also $\delta_F(X,a) = |F(X) - F(a)|$.

$$\underset{a}{argmin} E(\delta_F(X,a)) = \underset{a}{argmin} \int_\Omega |F(X) - F(a)|dP.$$

The last expression is minimized if $F(a)$ equals the median of the uniform. We conclude $F(a) = 1/2$ and the proof is complete. ■

8.4.1 Optimal index vectors for assigning quantiles to a random sample

Given a sample $X_1, \cdots, X_n$, $i.i.d \sim X$, we can find the sample order statistics $X_{(i)}, i = 1, \cdots, n$. Suppose we want to assign these order statistics to quantiles, $lq_X(p_i), i = 1, \cdots, n$, of the true distribution of X. In other words, what is the optimal index vector $\mathcal{P} = (p_1, \cdots, p_n)$ to assign $lq_X(p_i)$ to $X_{(i)}$. This can be used to make a qq–plot. We define the optimal vector to be the index vector that minimizes the expected probability loss

$$E[\frac{1}{n}\sum_{i=1}^{n}\delta_X(X_{(i)}, lq_X(p_i))].$$

We only solve the problem for continuous variables and leave the general case as an open problem. Under the continuity assumption, we have

$$E[\frac{1}{n}\sum_{i=1}^{n}\delta_X(X_{(i)}, lq_X(p_i))] = \frac{1}{n}\sum_{i=1}^{n}E(|F_X(X_{(i)}) - p_i|),$$

which is minimized if and only if the individual terms $E(|F_X(X_{(i)}) - p_i|)$ are minimized. Since F_X is a continuous random variable, $F_X(X_{(i)})$ is also continuous. Hence the minimum is obtained by solving $P(F_X(X_{(i)}) \leq x) = 1/2$ by the corollary of Theorem 8.4.1. By Lemma 5.5.1, this is equivalent to $P(X_{(i)} \leq rq_F(x)) = 1/2$. The distribution of the order statistics, $X_{(i)}$ is given by

$$P(X_{(i)} \leq y) = \sum_{j=i}^{n}\binom{n}{j}F(y)^j(1 - F(y))^{n-j},$$

as discussed by Casella and Berger in [11]. Hence, the minimum is obtained by solving

$$\sum_{j=i}^{n}\binom{n}{j}F(rq_X(x))^j(1 - F(rq_X(x)))^{n-j} = \sum_{j=i}^{n}\binom{n}{j}x^j(1-x)^{n-j} = 1/2,$$

which does not have a closed form solution in general. Also note that the solution does not depend on F. However, the solution always exists and is unique since $\sum_{j=i}^{n}\binom{n}{j}x^j(1-x)^{n-j}$ is increasing, continuous on (0,1) and ranges between 0 and 1. We also prove that the resulting index vector is symmetric in the sense that $p_{n-i+1} = 1 - p_i$, $i = 1, 2, \cdots, n$. For the proof, consider the random sample $(Y_1, \cdots, Y_n) = (-X_1, \cdots, -X_n)$. Then the sorted vector is $(Y_{(1)}, \cdots, Y_{(n)}) = (-X_{(n)}, \cdots, -X_{(1)})$. Hence $Y_{(i)} =$

$-X_{(n-i+1)}$. Suppose $p_1, \cdots, p_n$ is an optimal summary index vector. Then p_i is the solution of the first equation below

$$\underset{a}{argmin} E|F_Y(Y_{(i)}) - a| = \underset{a}{argmin} E|1 - F_X(Y_{(i)}) - a| =$$

$$\underset{a}{argmin} E|F_X(X_{(n-i+1)}) - (1-a)|.$$

But if we let $b = 1-a$ the solution to the last equation is $b = 1-a = p_{n-i+1}$. We conclude that $p_i = 1 - p_{n-i+1}$.

As examples, we solve the equation for $n = 1, 2$, where closed form solutions exist.
$n = 1$. Then $X_{(1)} = X_1$. It is easy to see that the solution is $p = 1/2$.
$n = 2$. Then we want to solve two equations

$$\sum_{j=1}^{2} \binom{2}{j} x^j (1-x)^{2-j} = 1/2,$$

and

$$\sum_{j=2}^{2} \binom{2}{j} x^j (1-x)^{2-j} = 1/2,$$

which are equivalent to

$$2x(1-x) + x^2 = 1/2,$$

and

$$x^2 = 1/2,$$

We get $p_1 = \frac{1}{\sqrt{2}}$ and $p_2 = 1 - \frac{1}{\sqrt{2}}$.

Note that in general for n, the last equation is $x^n = 1/2$. Hence $p_n = 1/\sqrt[n]{2}$ and $p_1 = 1 - 1/\sqrt[n]{2}$.

Chapter 9

Quantile distribution distance and estimation

9.1 Introduction

This chapter uses the probability loss function as a basis for estimating unknown parameters of a distribution and defining a distance among distribution functions. The "probability loss" and "c-probability loss" functions were introduced to measure the distance between quantiles. This is not the same as any other specific loss functions that have been proposed in statistical decision theory [30], where the loss function, L, is the loss of the statistician in estimating the true parameter vector $\theta = (\theta_1, \cdots, \theta_k) \in \Theta$, by an estimator $\hat{\theta}(X_1, \cdots, X_n)$ which is a function of the data (a random sample $X_1, \cdots, X_n$ drawn from the distribution parameterized by θ). The estimator is then chosen in such a way that $L(\hat{\theta}(X_1, \cdots, X_n), \theta)$ becomes small in some sense. However, it is not possible to use the probability loss function in the same manner for parameter estimation. We defined

$$\delta_X(z', z) = \delta_X(z, z') = P(z' < X < z),\ z' \leq z,\ z, z' \in \mathbb{R}.$$

Now it is clear that $\delta_X(\theta, a)$ cannot even be evaluated since θ is a k-dimensional vector and k is possibly greater than 1. This chapter presents two methods to estimate the parameters of distributions. More theoretical and applied development is necessary to justify such estimation procedures which we leave for future research. The first method derives from considering families of distributions that are identified by their values on certain quantiles and the second method from defining a distance among distributions and then trying to minimize that distance.

These methods are designed to give estimates that are equivariant under continuous strictly monotonic transformations. The distances associated with probability measures in this section are based on the distances between the quantiles using the probability loss function and they are invariant under monotonic transformations. This property does not hold in classical methods. For example the sample mean $\bar{x}$, an estimator of the location parameter

for normal distribution is equivariant under linear transformations but not all continuous strictly monotonic transformations.

Quantile distance allows us to measure closeness of distributions to each other. We also define a quantile distance for the tails of the distributions. We show that even though two distributions are very close in terms of "overall quantile distance", they might not be very close in terms of "tail quantile distance". This shows that to study extremes (for example extremely hot temperature) if we use a good overall fit, our results might not be reliable. We use this observation in the next chapter in choosing our method of studying extreme temperature events.

9.2 Quantile–specified parameter families

This section considers families of distributions that are identified by their values on certain quantiles. In this case the parameters in the vector $\theta = (\theta_1, \cdots, \theta_k)$ are certain quantiles. Then we use the "probability loss" or the "c-probability loss" to characterize the loss and thus yield optimal parameter estimators.

Definition A family of random variables $\{X_\theta\}_{\theta=(\theta_1,\cdots,\theta_k)\in\Theta}$, and a probability index vector $\mathcal{P} = (p_1, \cdots, p_k), 0 \leq p_1 < p_2 < \cdots < p_k \leq 1$ are called a left–quantile–specified family if

$$(\theta_1, \cdots, \theta_k) = (lq_{X_\theta}(p_1), \cdots, lq_{X_\theta}(p_k)),$$

and the distribution of X_θ is know given θ. Note that this implies that $\theta \in \Theta$ then $\theta_1 \leq \theta_2 \leq \cdots \theta_k$.

We can similarly define:

Definition A family of random variables $\{X_\theta\}_{\theta=(\theta_1,\cdots,\theta_k)\in\Theta}$, and a probability index vector $\mathcal{P} = (p_1, \cdots, p_k), 0 \leq p_1 < p_2 < \cdots < p_k \leq 1$ are called a right–quantile–specified family

$$(\theta_1, \cdots, \theta_k) = (rq_{X_\theta}(p_1), \cdots, rq_{X_\theta}(p_k)),$$

and the distribution of X_θ is know given θ. Note that this implies that $\theta \in \Theta$ then $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_k$.

Example Consider the family $\{U(0, 2a)\}_{a\in\mathbb{R}^+}$, of uniformly distributed random variables on $(0, 2a), a > 0$. Then, we can express this family as the quantile–specified family $\{X_\theta\}_{\theta\in\mathbb{R}^+}$ with $\mathcal{P} = (1/2)$. The reason is if $X_\theta \sim U(0, 2a)$ then $\theta = lq_{X_\theta}(1/2) = a$.

Example Consider the family $\mathcal{N} = \{N(\mu, \sigma^2) | -\infty < \mu < +\infty, \sigma^2 > 0\}$. Then we claim this is a quantile–specified family. To verify that claim let $\mathcal{P} = (1/2, p_2)$ where $p_2 = P(Z \leq 1)$ and Z has the standard normal distribution. Let

$$\mu = lq_X(1/2) = \theta_1,$$

and

$$\mu + \sigma^2 = lq_X(p_2) = \theta_2.$$

Then we can equivalently represent $\mathcal{N}$ by $\{X_\theta\}_{\theta=(\theta_1,\theta_2)\in\Theta}$, where

$$\Theta = \{(\theta_1, \theta_2) | \theta_1 < \theta_2\}.$$

Because (μ, σ^2) is in 1:1 correspondence with $\theta = (\theta_1, \theta_2)$ as defined above, where

$$P(X \leq \mu + \sigma^2) = P(Z \leq 1) = p_2.$$

Note that this representation is not unique. For example, we can take $\mathcal{P} = (1/2, p_2)$ with $p_2 = P(Z \leq 2)$. Then the alternate re-parametrization in terms of variables is

$$\mu = lq_X(1/2) = \theta_1,$$

and

$$\mu + 2\sigma^2 = lq_X(p_2) = \theta_2.$$

It should be clear that if the goal is to infer the parameters of the original family, i.e. a in $U(0, 2a)$ and (μ, σ^2) then it is desirable that the θ_i are simple functions of the original parameters and the original parameters be easily obtainable from the θ_i. Linear combinations seem to be the easiest to handle.

We suggest the following framework to estimate the parameters:

- Express the original parameterized family X_β as a quantile specified family X_θ with $\mathcal{P} = (p_1, \cdots, p_k)$.
- Use

  $$argmin_{D_i \in \mathcal{F}} E[L(\theta_i, D_i(input)],\ i = 1, \cdots, k$$

  where *input* is the information available to us, usually a random sample,

  $$(X_1, \cdots, X_n),$$

  D_i is an estimator of $\theta_i = lq_X(p_i)$ (a function of the random sample), L is a loss function and $\mathcal{F}$ is the class of the estimators. The loss functions of our interest are $L = \delta_{X_\theta}$ and $L = \delta^c_{X_\theta}, c > 0$.

- Using the estimated parameters solve for the original parameters, the β_i.

Note that $\delta^c_{X_\theta}, c > 0$ depends on the unknown distribution function X_θ. Many issues in the above framework need to be addressed including: the existence and uniqueness of the *argmin*, properties of the estimators and so on which we leave for future research. In next subsections we show the Equivariance property of the method and apply it to a particular class of estimators using simulations.

9.2.1 Equivariance of quantile–specified families estimation

Here, we show the equivariance property of estimation using quantile–specified families in the following lemmas.

Lemma 9.2.1 *Suppose $\{X_\theta\}_{\theta\in\Theta}$ is left–quantile–specified with*

$$\mathcal{P} = (p_1, \cdots, p_k),$$

and ϕ is a continuous strictly increasing transformation which induces a map on $\mathbb{R}^k$:

$$\Phi : \mathbb{R}^k \to \mathbb{R}^k,$$

$$(\theta_1, \cdots, \theta_k) \mapsto (\phi(\theta_1), \cdots, \phi(\theta_k)).$$

Let $\Theta' = \Phi(\Theta), \theta' = \Phi(\theta)$ for $\theta \in \Theta$ and consider the family of distributions $Y_{\theta'} = \phi(X_\theta)$. Then $\{Y_{\theta'}\}_{\theta'\in\Theta'}$ is also a left–quantile–specified family with the same index vector $\mathcal{P} = (p_1, \cdots, p_k)$.

Proof Suppose the distribution of X_θ is specified by F_θ. Then

$$P(Y_{\theta'} \le a) = P(\phi(X_\theta) \le a)$$

$$= F_\theta(\phi^{-1}(a)) = F_{\Phi^{-1}(\theta')}(\phi^{-1}(a)).$$

Hence the distribution of $Y_{\theta'}$ is known given θ'. It remains to show that for $\theta' \in \Theta'$,

$$(\theta'_1, \cdots, \theta'_k) = (lq_{Y_{\theta'}}(p_1), \cdots, lq_{Y_{\theta'}}(p_k)).$$

But

$$\begin{aligned}
(lq_{Y_{\theta'}}(p_1), \cdots, lq_{Y_{\theta'}}(p_k)) &= \\
(lq_{\phi(X_\theta)}(p_1), \cdots, lq_{\phi(X_\theta)}(p_k)) &= \\
(\phi(lq_{X_\theta}(p_1)), \cdots, \phi(lq_{X_\theta}(p_k))) &= \\
(\phi(\theta_1), \cdots, \phi(\theta_k)) &= \\
(\theta'_1, \cdots, \theta'_k) &\quad .
\end{aligned}$$

■

Lemma 9.2.2 *Suppose* $\{X_\theta\}_{\theta\in\Theta}$ *is left–quantile–specified with*

$$\mathcal{P} = (p_1, \cdots, p_k),$$

and ϕ *is a continuous strictly decreasing transformation which induces a map on* $\mathbb{R}^k$*:*

$$\Phi : \mathbb{R}^k \to \mathbb{R}^k,$$

$$(\theta_1, \cdots, \theta_k) \mapsto (\phi(\theta_k), \cdots, \phi(\theta_1)).$$

Let $\Theta' = \Phi(\Theta), \theta' = \Phi(\theta)$ *for* $\theta \in \Theta$ *and consider the family of distributions* $Y_{\theta'} = \phi(X_\theta)$*. Then* $\{Y_{\theta'}\}_{\theta'\in\Theta'}$ *is a right–quantile–specified family with the index vector* $\mathcal{P} = (1 - p_k, \cdots, 1 - p_1)$*.*

Proof Suppose the distribution of X_θ is specified by F_θ. Then since F_θ the left closed distribution of X_θ is known, the right closed distribution of X_θ, $G^c_X(X_\theta)$ is also known. Then

$$P(Y_{\theta'} \le a) = P(\phi(X_\theta) \le a) = P(X_\theta \ge \phi^{-1}(a))$$

$$= G^c_\theta(\phi^{-1}(a)) = G^c_{\Phi^{-1}(\theta')}(\phi^{-1}(a)),$$

where G^c_θ is the right closed distribution function. Hence the distribution of $Y_{\theta'}$ is known given θ'. It remains to show that for $\theta' \in \Theta'$,

$$(\theta'_1, \cdots, \theta'_k) = (rq_{Y_{\theta'}}(1 - p_k), \cdots, rq_{Y_{\theta'}}(1 - p_1)).$$

But

$$\begin{aligned}
(rq_{Y_{\theta'}}(1 - p_k), \cdots, rq_{Y_{\theta'}}(1 - p_1)) &= \\
(rq_{\phi(X_\theta)}(1 - p_k), \cdots, rq_{\phi(X_\theta)}(1 - p_1)) &= \\
(\phi(lq_{X_\theta}(p_k)), \cdots, \phi(lq_{X_\theta}(p_1))) &= \\
(\phi(\theta_k), \cdots, \phi(\theta_1)) &= \\
(\theta'_1, \cdots, \theta'_k).
\end{aligned}$$

■

For a parameter θ, we want to find

$$argmin_{D\in\mathcal{F}} E(\delta_X(lq_X(p), D))$$

where $\mathcal{F}$ is a family of estimators for θ and $D \in \mathcal{F}$ is a function

$$D : \mathbb{R}^n \to \mathbb{R},$$

where n is the size of the sample and $D(X_1, \cdots, X_n)$ is the estimator of $\theta = lq_X(p)$.

Lemma 9.2.3 *Suppose a random sample $X_1, \cdots, X_n$ is given, X_θ is a left–quantile–specified family with $\theta = lq_X(p)$, ϕ a strictly monotonic continuous transformation on $\mathbb{R}$, $\mathcal{F}$ is a family of estimators to estimate θ and the following argmin is nonempty*

$$argmin_{D\in\mathcal{F}} E(\delta_X(lq_X(\theta), D)),$$

and let $\mathcal{F}' = \phi(\mathcal{F})$. Then
a) if ϕ is strictly increasing

$$argmin_{D'\in\mathcal{F}'} E(\delta_{\phi(X)}(lq_{\phi(X)}(p), D')) = \phi(argmin_{D\in\mathcal{F}} E(\delta_X(lq_X(p), D)))$$

b) if ϕ is strictly decreasing

$$argmin_{D'\in\mathcal{F}'} E(\delta_{\phi(X)}(lq_{\phi(X)}(p), D')) = \phi(argmin_{D\in\mathcal{F}} E(\delta_X(rq_X(1-p), D)))$$

Proof We only prove a) and b) is similar.

$$\begin{aligned} &\min_{D'\in\mathcal{F}'} E(\delta_{\phi(X)}(lq_{\phi(X)}(p), D')) \\ =& \min_{D\in\mathcal{F}} E(\delta_{\phi(X)}(\phi(lq_X)(p), \phi(D))) \\ =& \min_{D\in\mathcal{F}} E(\delta_X(lq_X(p), D)) \end{aligned}$$

■

Note that for a general family of estimators, $\mathcal{F}$

$$argmin_{D\in\mathcal{F}} E(\delta_X(lq_X(p), D))$$

depends on the unknown distribution X by δ_X. We suggest two possible ways to get around this issue:

- Restrict to a family $\mathcal{F}$ that

  $$argmin_{D\in\mathcal{F}} E(\delta_X(lq_X(p), D))$$

  does not depend on the distribution.

- Use the empirical distribution to approximate the expression

$$E(\delta_X(lq_X(p), D)).$$

We will not explore the second method here and leave it for future research. Next subsection shows an important instance of the first method.

9.2.2 Continuous distributions with the order statistics family of estimators

Suppose that the desired distribution X is continuous then

$$E(\delta_X(lq_X(p), D)) = E|F_X(lq_X(p)) - F_X(D)| = E|p - F_X(D)|.$$

Now suppose a random sample $X_1, \cdots, X_n$ is given and we want to estimate $lq_X(p)$. We restrict to an important family of estimators, order statistics:

$$\mathcal{F} = \{X_{1:n}, \cdots, X_{n:n}\}.$$

Then for $i = 1, \cdots, n$:

$$E|p - F_X(X_{i:n})|,$$

does not depend on F_X. This is because the distribution of $F_X(X_{i:n})$ does not depend on F_X. It can be obtained as shown below:

$$G_i(y) = P(F_X(X_{i:n}) \leq y) = P(X_{i:n} \leq lq_X(y)) =$$
$$\sum_{j=i}^{n} \binom{n}{j} P(X_1, \cdots, X_j \leq lq_X(y) \text{ and } X_{j+1}, \cdots, X_n > lq_X(y)) =$$
$$\sum_{j=i}^{n} \binom{n}{j} P(X \leq lq_X(y))^j P(X > lq_X(y))^{n-j} = \sum_{j=i}^{n} \binom{n}{j} y^j (1-y)^{n-j}.$$

By taking the derivative of the above expression we can find the density function $g_i(p)$ and conclude:

$$E|p - F_X(X_{i:n})| = \int_0^1 |p - y| g_i(y) dy.$$

For a given p we want to find the i that minimize above which does not on F_X. We can approach this problem theoretically to find such an i. Or we could try to estimate these integral using numerical methods. However, here we use simulation for two examples and leave the general case for future research.

Example Consider a family of continuous variables, quantile–specified by $\mathcal{P} = (1/2, P(Z \leq 1))$ where Z is the standard normal. Suppose a random sample $X_1, \cdots, X_n$ is given and we want to estimate $lq_X(1/2)$ and $lq_X(P(Z \leq 1))$ using the family of estimators, order statistics:

$$\mathcal{F} = \{X_{1:n}, \cdots, X_{n:n}\}.$$

We estimate the parameters for $n = 25$ and $n = 20$. In order to minimize the loss we can approximate the loss by approximating the integral in Equation 9.2.2 or approximating

$$E|p - F_X(X_{i:n})|,$$

using an arbitrary continuous distribution such as standard normal to do the simulations. For a large number M, we create M samples of length n from normal and for every sample we find the i that minimize the loss. Then for every i, we compute the mean of such losses and find out which has the smallest mean loss. We do that for $M = 1, \cdots, 1000$. The results for $n = 25$ are given in Figure 9.1. We see that for large M the estimator for $lq_X(1/2)$ is $X_{13:25}$ and for $lq_X(P(Z < 1))$ it is $X_{22:25}$. The results for $n = 20$ are given in Figure 9.2. The estimator for $lq_X(1/2)$ has changed between $X_{10:20}$ and $X_{11:21}$ and it is $X_{18:20}$ for $lq_X(P(Z \leq 1))$. This shows that the argmin is not necessarily unique.

9.3 Probability divergence (distance) measures

In probability theory, physics and statistics several measures have been introduced as the "distance" of two probability measures (or random variables). These measures have several applications, one of which is parameter estimation. We list some of these measures in this section. The next section then introduces new measures of distance among probability measures using the c-probability loss functions ($c \geq 0$).

- The Kullback-Leibler (KL) distance: Suppose P, Q are probability measures and P is absolutely continuous with respect to Q. Then consider the Radon-Nikodym derivative of P with respect to Q, $\frac{dP}{dQ}$ [See [9]]. Then we define:

$$D_{KL}(P, Q) = \int_\Omega \log \frac{dP}{dQ} dP.$$

  If P and Q have density functions over $\mathbb{R}$, $p(x), q(x)$ then

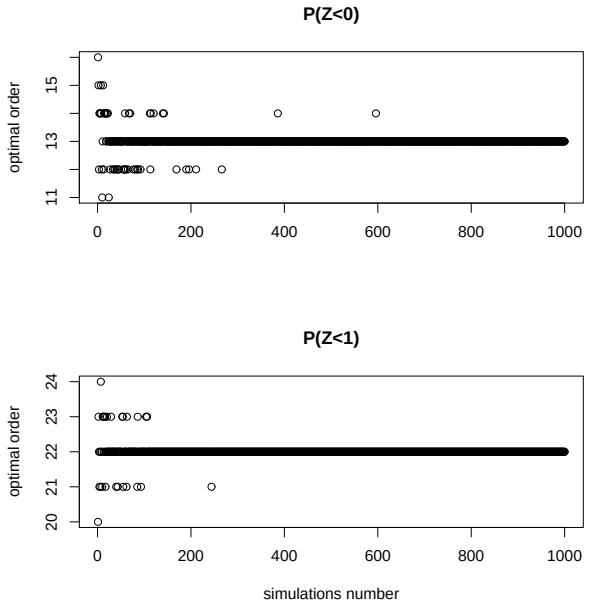

Figure 9.1: The order statistics family members that estimate $lq_X(1/2)$ and $lq_X(P(Z \leq 1))$ for a random sample of length 25 obtained by generating samples of size 1 to 1000 from a standard normal distribution

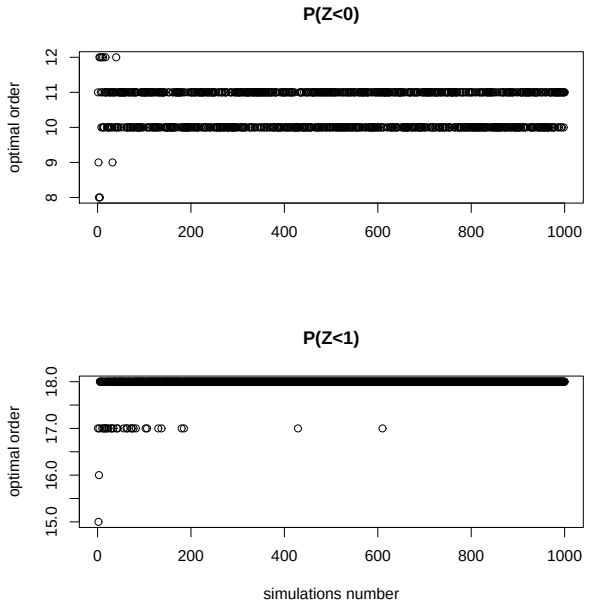

Figure 9.2: The order statistics family members that estimate $lq_X(1/2)$ and $lq_X(P(Z \leq 1))$ for a random sample of length 20 obtained by generating samples of size 1 to 1000 from a standard normal distribution

$$\int_{\mathbb{R}} p(x) log(\frac{p(x)}{q(x)}) dx.$$

The symmetric version of this distance is called Kullback-Jeffreys

$$D_{KJ}(P,Q) = D_{KL}(P,Q) + D_{KL}(Q,P).$$

We show that the Kullback-Leibler distance is invariant under bijective differentiable monotonic transformations when the density functions exists and are positive everywhere on the real line. Let g be a monotonic, bijective and differentiable (bijective and differentiable will automatically imply strictly monotonic) transformation and X, Y random variables with density functions $f_X(x)$ and $f_Y(x)$, positive on $\mathbb{R}$. Then the density functions of $g(X)$ and $g(Y)$ are respectively $(g^{-1})'(x) f_X(g^{-1}(x))$ and $(g^{-1})'(x) f_Y(g^{-1}(x))$. Hence

$$D_{KL}(\phi(X), \phi(Y)) =$$
$$\int_{-\infty}^{\infty} (g^{-1})' f_X(g^{-1}(x)) \log \frac{(g^{-1})' f_X(g^{-1}(x))}{(g^{-1})' f_Y(g^{-1}(x))} dx =$$
$$\int_{-\infty}^{\infty} (g^{-1})' f_X(g^{-1}(x)) \log \frac{(f_X(g^{-1}(x))}{f_Y(g^{-1}(x))} dx.$$

We use the change of variable $x = g(y)$. Then $dx = (g^{-1})' dy$ and the proof is complete. For the strictly decreasing case note that the density function of $g(X)$ and $g(Y)$ are respectively $-(g^{-1})'(x) f_X(g^{-1}(x))$ and $-(g^{-1})(x)' f_Y(g^{-1}(x))$ and a similar argument works. We leave the general case (where the density function does not exist or is not positive over all the real line) as an open(?) problem.

- Let P and Q be two probability distributions over a space Ω such that P is absolutely continuous with respect to Q. Then, for a convex function f such that $f(1) = 0$, the f-divergence of Q from P is

$$I_f(P,Q) = \int_{\Omega} f\left(\frac{dP}{dQ}\right) dQ.$$

Note that the same argument as the one for KL distance shows that this distance is invariant for monotonic differentiable bijective transformations when the density functions exist and are positive.

- The Kolmogorov-Smirnov distance: Suppose X, Y are random variables on $\mathbb{R}$ with distribution functions F_X and F_Y. Then

$$KS(X,Y) = \sup_{x\in\mathbb{R}} |F_X(x) - F_Y(x)|.$$

The Gilvenko-Cantelli Theorem states that if $X_1, \cdots, X_n$ is a random sample drawn from the distribution F_{θ_0} and F_n, the empirical distribution function

$$\lim_{n\to\infty} KS(F_{\theta_0}, F_n) > \epsilon = 0, \;\; a.s..$$

Note that the KS metric is invariant under monotonic transformations. Take ϕ to be strictly monotonic on $\mathbb{R}$. Then

$$\begin{aligned}
\sup_{x\in\mathbb{R}} |F_{\phi(X)}(x) - F_{\phi(Y)(x)}| &= \\
\sup_{x\in\mathbb{R}} |F_X(\phi^{-1}(x)) - F_Y(\phi^{-1}(x))| &= \\
\sup_{\phi^{-1}(x)\in\mathbb{R}} |F_X(\phi^{-1}(x)) - F_Y(\phi^{-1}(x))| &= \\
\sup_{x\in\mathbb{R}} |F_X(x) - F_Y(x)|.
\end{aligned}$$

Although the KS metric is invariant under strictly monotonic transformations, it is not intuitively very appealing as we show in the following example.

Example Consider $X \sim U(0,1)$, $Y \sim U(1/2, 3/2)$ and let Z be distributed as F_Z:

$$F_Z(z) = \begin{cases} 0 & z < 0 \\ 1/2 & 0 \le z \le 1/2 \\ z & 1/2 < z < 1 \\ 1 & z \ge 1 \end{cases}.$$

Then we have $KS(X,Y) = KS(X,Z) = 1/2$. But we observe that F_Z matches F_X on $(1/2, 1)$ while F_X and F_Y differ by $1/2$ on $(0,1)$. Another way to see the defect is the quantiles of Z and X match half of the time but the quantiles of X and Y are off as much as one half of a unit at all times.

To overcome the above problem one might (naively) suggest using an integral version

$$IKS(X,Y) = \int_{x\in\mathbb{R}} |F_X(x) - F_Y(x)|dx.$$

However, this definition is not well-defined. To see that consider $F_X(x) = 1 - 8/x, x > 8$ and $F_Y(x) = 1 - 9/x,\ x > 9$. Then $|F_X(x) - F_Y(x)| = 1/x$ on $[8,\infty]$, which does not have finite integral. It is also not invariant under strictly monotonic transformations for if ϕ is strictly monotonic and differentiable,

$$IKS(\phi(X),\phi(Y)) = \int_{x\in\mathbb{R}} |F_X(\phi^{-1}(x)) - F_Y(\phi^{-1}(x))|dx.$$

In the right hand side of the above equation the factor $(\phi^{-1})'$, that would make the distance invariant under transformations, is missing.

- Lévy distance: Suppose $(\Omega,\Sigma,P_\theta)_{\theta\in\Theta}$ be a statistical space, where the P_θ are probability measures on Ω with σ-field Σ. Then we define

$$Lev(F_{\theta_1},F_{\theta_2}) = \inf\{\epsilon > 0 | F_{\theta_1}(x-\epsilon) < F_{\theta_2}(x) < F_{\theta_1}(x+\epsilon),\ \forall x \in \mathbb{R}\}.$$

It can be shown that convergence in the Lévy metric implies weak convergence for distribution function in $\mathbb{R}$ [31]. It is shift invariant but not scale invariant as discussed in [31].

9.4 Quantile distance measures

This section introduces the quantile distance measure to measure the distance among distribution functions on $\mathbb{R}$ (or random variables). We begin with a general definition using the quantiles and then consider interesting particular cases. The intuition behind all these metrics lies in their capability to measure the separation in the quantiles of two random variables.

Definition Suppose a statistical space $(\Omega,P,\{X_\theta\}_{\theta\in\Theta})$ and a loss function L defined over the extended real numbers $\mathbb{R}\cup\{-\infty,+\infty\}$ are given. Also let E be a measurable subset of (0,1) and $d\mu_E$ is a measure on E. Then we can define the following two measures of distance between X_{θ_1} and X_{θ_2},

$$SQD_L^E(X_{\theta_1}, X_{\theta_2}) = \sup_{p\in E} L(lq_{X_{\theta_1}}(p), lq_{X_{\theta_2}}(p)),$$

and

$$IQD_L^E(X_{\theta_1}, X_{\theta_2}) = \int_{p\in E} L(lq_{X_{\theta_1}}(p), lq_{X_{\theta_2}}(p))d\mu_E,$$

which we call the sup quantile distance and integral quantile distance respectively.

Remark. Note that in general SQD_L^E and IQD_L^E are neither well-defined nor metrics on the space of random variables..
Remark. We can also take $L(rq_{X_{\theta_1}}(p), rq_{X_{\theta_2}}(p))$ in the above definitions.
Remark. The natural choice for E is $(0,1)$ and the measure $\mu = \mathcal{L}$, where $\mathcal{L}$ is the Lebègues measure on $(0,1)$. However, one might choose another E depending on the purpose. For example $E = (0.8, 1)$ might be more appropriate if the purpose is modeling the high extremes.
Remark. Interesting choices for L are $\delta_{X_{\theta_1}}$, $\delta^c_{X_{\theta_1}}$, $\delta_{X_{\theta_1}} + \delta_{X_{\theta_2}}$ and $\delta^c_{X_{\theta_1}} + \delta^c_{X_{\theta_2}}$. Note that in all these cases the quantile distance is defined since these quantities are bounded respectively by $1, 1+c, 2, 2+2c$.

The rest of this report focuses on quantile distances obtained from c-probability losses ($c \geq 0$). (Note that $c = 0$ corresponds to the usual probability loss.)

9.4.1 Quantile distance invariance under continuous strictly monotonic transformations

This subsection show the invariance of quantile distance under strictly monotonic tranformations in the following lemmas.

Lemma 9.4.1 *(Quantile distance invariance under continuous strictly increasing transformations)*
Suppose X, Y are random variables, let

$$IQD_{\delta^c_X}^E(X, Y) = \int_E L(lq_X(p), lq_Y(p))d\mu_E,$$

and

$$SQD_{\delta^c_X}^E(X, Y) = \sup_{p\in E} L(lq_X(p), lq_Y(p)),$$

where $E \subset (0,1)$, $c \geq 0$ and μ_E is a measure on E. Then

$$IQD_{\delta^c_X}^E(X, Y) = IQD_{\delta^c_{\phi(X)}}^E(\phi(X), \phi(Y)),$$

and

$$SQD^E_{\delta^c_X}(X,Y) = SQD^E_{\delta^c_{\phi(X)}}(\phi(X),\phi(Y)),$$

for all $\phi : \mathbb{R} \to \mathbb{R}$ *continuous and strictly increasing transformations.*

Proof The proof attains from noting that

$$\begin{aligned}
\delta_{\phi(X)}(lq_{\phi(X)}(p), lq_{\phi(Y)}(p)) &= \\
\delta_{\phi(X)}(lq_{\phi(X)}(p), lq_{\phi(Y)}(p)) + c(1 - 1_{\{0\}}(lq_{\phi(X)}(p) - lq_{\phi(Y)}(p))) &= \\
\delta_{\phi(X)}(\phi(lq_X(p)), \phi(lq_Y(p))) + c(1 - 1_{\{0\}}(lq_X(p) - lq_Y(p))) &= \\
\delta_X(lq_X(p), lq_Y(p)) + c(1 - 1_0(lq_X(p) - lq_Y(p))) &= \\
\delta^c_X(lq_X(p), lq_Y(p)).
\end{aligned}$$

■

Remark. The above lemma is also true for $\delta^c_X + \delta^c_Y$, which follows immediately.

Lemma 9.4.2 *If E a measurable subset of [0,1] then the two following distance measures are equal:*

$$LQD^E_{\delta_X}(X,Y) = \int_E \delta_X(lq_X(p), lq_Y(p))dp,$$

and

$$RQD^E_{\delta_X}(X,Y) = \int_E \delta_X(rq_X(p), rq_Y(p))dp.$$

The following two measures are also equal:

$$LQD^E_{\delta_X+\delta_Y}(X,Y) = \int_E (\delta_X + \delta_Y)(lq_X(p), lq_Y(p))dp,$$

and

$$RQD^E_{\delta_X+\delta_Y}(X,Y) = \int_E (\delta_X + \delta_Y)(rq_X(p), rq_Y(p))dp.$$

Proof We prove the first part part of the lemma and the second part is deduced from the first. We showed in the quantile definition section that the set $\{p|lq_X(p) \neq rq_X(p)\}$ is countable. Hence,

$$\{p|lq_X(p) \neq rq_X(p)\} \cup \{p|lq_Y(p) \neq rq_Y(p)\},$$

is also countable. In the complement of this set

$$\delta_X(lq_X(p), lq_Y(p)) = \delta_X(rq_X(p), rq_Y(p)).$$

Hence the integral values are the same. ■

Remark. Note that the above theorem also holds for any measure μ on any $E \subset (0,1)$ which is continuous with respect to the Lebègue measure. Because of this lemma we will not worry about the left or right quantile in the definitions.

The following lemma establishes a relationship between LQD_{δ_X} and $LQD_{\delta_X^c}$.

Lemma 9.4.3 *Let E be a measurable subset of [0,1] and*

$$k_E = \mathcal{L}\{p \in E | lq_X(p) \neq lq_Y(p)\},$$

where $\mathcal{L}$ is the Lebègue measure. Let

$$LQD^E_{\delta_X^c}(X,Y) = \int_E \delta_X^c(lq_X(p), lq_Y(p))dp,$$

and

$$LQD^E_{\delta_X}(X,Y) = \int_E \delta_X(lq_X(p), lq_Y(p))dp.$$

Then

$$LQD^E_{\delta_X^c}(X,Y) = LQD^E_{\delta_X}(X,Y) + ck_E.$$

Proof

$$\begin{aligned}
LQD_{\delta_X^c}(X,Y) &= \\
\int_E \delta_X^c(lq_X(p), lq_Y(p))dp &= \\
\int_{lq_X(p)=lq_Y(p), p\in E} \delta_X^c(lq_X(p), lq_Y(p))dp &+ \\
\int_{lq_X(p)\neq lq_Y(p), p\in E} \delta_X^c(lq_X(p), lq_Y(p))dp &= \\
\int_{lq_X(p)=lq_Y(p), p\in E} \delta_X(lq_X(p), lq_Y(p))dp &+ \\
\int_{lq_X(p)\neq lq_Y(p), p\in E} [\delta_X(lq_X(p), lq_Y(p)) + c(1-1_{\{0\}})(lq_X(p) - lq_Y(p))]dp &= \\
LQD^E_{\delta_X}(X,Y) + ck_E.
\end{aligned}$$

■

Remark. Note that the same is true for $RQD^E_{\delta^c_X}$ and $RQD^E_{\delta_X}$. Also

$$\mathcal{L}\{p \in E | lq_X(p) \neq lq_Y(p)\} = \mathcal{L}\{p, p \in E | rq_X(p) \neq rq_Y(p)\},$$

because lq_X, rq_X and lq_Y, rq_Y are unequal only on a measure zero set. Hence the constant k_E is the same as before and

$$RQD^E_{\delta^c_X}(X,Y) = RQD^E_{\delta_X}(X,Y) + ck_E.$$

Lemma 9.4.4 *Suppose E a measurable subset of [0,1] then the two following distance measures are equal*

$$LQD^E_{\delta^c_X}(X,Y) = \int_{p\in E} \delta^c_X(lq_X(p), lq_Y(p))dp,$$

and

$$RQD^E_{\delta^c_X}(X,Y) = \int_{p\in E} \delta^c_X(rq_X(p), rq_Y(p))dp.$$

Also these two measures are equal

$$LQD^E_{\delta^c_X+\delta^c_Y}(X,Y) = \int_{p\in E} (\delta^c_X + \delta^c_Y)(lq_X(p), lq_Y(p))dp,$$

and

$$RQD^E_{\delta^c_X+\delta^c_Y}(X,Y) = \int_{p\in E} (\delta^c_X + \delta^c_Y)(rq_X(p), rq_Y(p))dp.$$

Proof

This is a straightforward consequence of the previous two lemmas.

■

Remark. Note that the above theorem also holds for any measure μ on any $E \subset (0,1)$ which is continuous with respect to the Lebègue measure.

Lemma 9.4.5 *(Quantile distance invariance under continuous strictly monotonic transformations)*
Suppose X, Y are random variables and let

$$QD^E(X,Y) = LQD^E_{\delta_X}(X,Y), \tag{9.1}$$

$$QD^E_c(X,Y) = LQD^E_{\delta^c_X}(X,Y), \tag{9.2}$$

where, $E \subset (0,1)$ symmetric, meaning $p \in E \Leftrightarrow (1-p) \in E$, and μ is absolutely continuous with respect to the Lebègue measure and symmetric on E in the sense that if A is measurable then so is $1-A$ while $\mu(A) = \mu(1-A)$. Then 9.1 and 9.2 are invariant under continuous strictly monotonic transformations, i.e.
a) $QD^E(\phi(X),\phi(Y)) = LQD^E_{\delta_{\phi(X)}}(\phi(X),\phi(Y)) = QD^E(X,Y) = QD^E_{\delta_X}(X,Y)$,
b) $QD^E_c(\phi(X),\phi(Y)) = LQD^E_{\delta^c_{\phi(X)}}(\phi(X),\phi(Y)) = QD^E_c(X,Y) = QD^E_{\delta^c_X}(X,Y)$.

Proof For ϕ continuous and strictly increasing transformations, we have shown the result in Lemma 9.4.1. Suppose ϕ is continuous and strictly decreasing.
a) We use $lq_{\phi(X)}(p) = \phi(rq_X(1-p))$ which we proved above using quantile symmetries:

$$\begin{aligned} \delta_{\phi(X)}(lq_{\phi(X)}(p), lq_{\phi(Y)}(p)) &= \\ \delta_{\phi(X)}(\phi(rq_X(1-p)), \phi(rq_Y(1-p))) &= \\ \delta_{-\phi(X)}(-\phi(rq_X(1-p)), -\phi(rq_Y(1-p))), & \end{aligned}$$

where the last equality is because $\delta_X(a,b) = \delta_{-X}(-a,-b)$. Now since $-\phi$ is continuous and increasing, the above is equal to

$$\delta_X(rq_X(1-p), rq_Y(1-p)).$$

We use this result in the following:

$$\begin{aligned} QD^E(X,Y) &= \int_E \delta_X(lq_X(p), lq_Y(p))d\mu_E \\ &= \int_E \delta_X(rq_X(1-p), rq_Y(1-p))d\mu_E. \end{aligned}$$

Then we do a change of variable $p \to (1-p)$ and by symmetry of μ, we find that the above is equal to

$$\int_E \delta_X(rq_X(p), rq_Y(p))d\mu_E.$$

But by the previous lemmas and since μ is continuous with respect to the Lebègue measure, this is equal to

$$\int_E \delta_X(lq_X(p), lq_Y(p))d\mu_E.$$

b) We only consider continuous and strictly decreasing functions ϕ:

$$LQD^E_{\delta^c_{\phi(X)}}(\phi(X),\phi(Y)) =$$
$$\int_E c(1-1_{\{0\}}(lq_{\phi(X)}(p)-lq_{\phi(Y)}(p)))dp + LDQ_{\delta_{\phi(X)}}(X,Y) =$$
$$ck_E + LDQ^E_{\delta_{\phi(X)}}(\phi(X),\phi(Y)),$$

where,

$$\begin{aligned} k_E = \mu\{p\in E|lq_{\phi(X)}(p)\neq lq_{\phi(Y)}(p)\} &= \\ \mu\{p\in E|\phi(rq_X(1-p))\neq\phi(rq_Y(1-p))\} &= \\ \mu\{p\in E|rq_X(1-p)\neq rq_Y(1-p)\} &= \\ \mu\{p\in E|rq_X(p)\neq rq_Y(p)\} &= \\ \mu\{p\in E|lq_X(p)\neq lq_Y(p)\}. & \end{aligned}$$

We showed in a) that

$$LDQ^E_{\delta_{\phi(X)}}(\phi(X),\phi(Y)) = LDQ^E_{\delta_X}(X,Y)$$

and because we just showed that $k_E = \mu\{p\in E,|(lq_X(p))\neq(lq_Y)(p)\}$, we conclude

$$DQ^E_{\delta^c_{\phi(X)}} = ck_E + LDQ^E_{\delta_{\phi(X)}}(\phi(X),\phi(Y)) = ck_E + LDQ^E_{\delta_X}(X,Y) = LQD^E_{\delta^c_X}(X,Y).$$

■

9.4.2 Quantile distance closeness of empirical distribution and the true distribution

The next theorem shows that the quantile distance between the sample distribution and the true distribution tends to zero when the sample size becomes large.

Theorem 9.4.6 *Let $X_1, X_2, \cdots$ be an i.i.d. random sample drawn from an arbitrary distribution function F. Then*

$$(a)\ \ SQD_{\delta_X}(F,F_n) = \sup_{p\in(0,1)} \delta_F(lq_{F_n}(p), lq_F(p)) \to 0.,\ a.s.,$$

and

$$(b)\ \ IQD_{\delta_X}(F,F_n) = \int_{p\in(0,1)} \delta_F(lq_{F_n}(p), lq_F(p)) \to 0.,\ a.s..$$

Proof

We only need to prove (a) since (b) is a straightforward consequence of (a). Clearly $lq_{F_n}(p) = X_{i:n}$ for $p \in ((i-1)/n, i/n], i = 1, 2, \cdots, n$. Also $F_n^c(X_{i:n}) \geq i/n$ and $F_n^o(X_{i:n}) \leq (i-1)/n$. Pick an N large enough in the Glivenko-Cantelli Theorem such that

$$n > N \Rightarrow |F_n(x) - F(x)| < \epsilon, \text{ and } |F_n^o(x) - F^o(x)| < \epsilon,$$

uniformly in x. Consider two cases:
Case I: $X_{i:n} < lq_F(p)$. Then

$$\begin{aligned} \delta_F(lq_{F_n}(p), lq_F(p)) &= \delta_F(X_{i:n}, lq_F(p)) = \\ F^o(lq_F(p)) - F^c(X_{i:n}) &\leq F^o(lq_F(p)) - F_n^c(X_{i:n}) + \epsilon \\ &\leq p - i/n + \epsilon \leq \epsilon. \end{aligned}$$

Case II: $X_{i:n} > lq_F(p)$. Then

$$\begin{aligned} \delta_F(lq_{F_n}(p), lq_F(p)) &= \delta_F(X_{i:n}, lq_F(p)) = \\ F^o(X_{i:n}) - F^c(lq_F(p)) &\leq F_n^o(X_{i:n}) + \epsilon - p \\ &\leq (i-1)/n + \epsilon - p \leq \epsilon. \end{aligned}$$

Since this holds for $i = 1, 2, \cdots, n$ and $(0,1) = \cup_{i=1,2,\cdots,n}(\frac{i-1}{n}, \frac{i}{n}]$, the supremum is also less than ϵ. ■

9.4.3 Quantile distance and KS distance closeness

Clearly if $X \sim Y$, then $LQD_L^E(X,Y) = 0$. In the following theorem we study the inverse question for $L = \delta_X^c,\ c \geq 0$ and $E = [0,1]$. The Kolmogorov Smirnoff distance was defined to be

$$KS(X,Y) = \sup_{x\in\mathbb{R}} |F_X(x) - F_Y(x)|.$$

We also define the "open Kolmogorov Smirnoff" distance as

$$KS^o(X,Y) = \sup_{x\in\mathbb{R}} |F_X^o(x) - F_Y^o(x)|.$$

Lemma 9.4.7 *Suppose X, Y are random variables, then*

$$KS^o(X,Y) = KS(X,Y).$$

To prove the lemma, we show that

$$KS(X,Y) \leq \epsilon \Leftrightarrow KS^o(X,Y) \leq \epsilon.$$

Suppose $KS(X,Y) \leq \epsilon$. If the R.H.S does not hold then there exist $x \in \mathbb{R}$ such that

$$F_X^o(x) > F_Y^o(x) + \epsilon.$$

Since F_X^o is left continuous, we conclude there is a $y < x$ such that

$$F_X^o(y) > F_Y^o(x) + \epsilon.$$

Hence,

$$F_X^c(y) \geq F_X^o(y) > F_Y^o(x) + \epsilon \geq F_Y^c(y) + \epsilon,$$

which is a contradiction.
Inversely, suppose $KS^o(X,Y) \leq \epsilon$. If the L.H.S does not hold then there exist $x \in \mathbb{R}$ such that

$$F_X^c(x) > F_Y^c(x) + \epsilon.$$

Since F_Y^c is right continuous, we conclude there is $y > x$ such that

$$F_X^c(x) > F_Y^c(y) + \epsilon.$$

Hence,

$$F_X^o(y) \geq F_X^c(x) > F_Y^c(x) + \epsilon \geq F_Y^o(y) + \epsilon,$$

which is a contradiction.

Lemma 9.4.8 *Kolmogorov Smirnoff closeness implies Quantile distance closeness. More formally if for two random variables X, Y, $KS(X,Y) \leq \epsilon$ then*

$$SQD_{\delta_X}(X,Y) = \sup_{p \in (0,1)} \delta_X(lq_X(p), lq_Y(p)) \leq \epsilon.$$

Proof

For $p \in (0,1)$, suppose $lq_X(p) < lq_Y(p)$. Then

$$\delta(lq_X(p), lq_Y(p)) = F_X^o(lq_Y(p)) - F_X^c(lq_Y(p)) \leq$$

$$F^o(lq_Y(p)) + \epsilon - p \leq p + \epsilon - p = \epsilon.$$

The discussion for $lq_Y(p) < lq_X(p)$ is similar. ■

Remark. By symmetry also $KS(X,Y) \leq \epsilon \Rightarrow SQD_{\delta_Y}(X,Y) \leq \epsilon$.

The converse needs the continuity assumption:

Lemma 9.4.9 *Suppose X, Y are continuous random variables. Then quantile distance closeness implies Kolmogorov Smirnoff distance closeness. More formally, suppose*

$$SQD_{\delta_X}(X,Y) = \sup_{p\in(0,1)} \delta_X(lq_X(p), lq_Y(p)) \leq \epsilon$$

and

$$SQD_{\delta_Y}(X,Y) = \sup_{p\in(0,1)} \delta_Y(lq_X(p), lq_Y(p)) \leq \epsilon.$$

Then

$$KS(X,Y) \leq \epsilon.$$

Proof Suppose the result is not true and there exists x such that

$$|F_X(x) - F_Y(x)| \geq \epsilon.$$

Then let $p_1 = F_X(x)$ and $p_2 = F_Y(x)$ and without loss of generality assume $p_2 > p_1$. Since $F_Y(x) = p_2$, $lq_Y(p_2) \leq x$. But $lq_X(p_2) = y > x$. Otherwise $p_2 \leq F_X(lq_X(p_2)) = F_X(x) = p_1$ which is a contradiction.

$$\delta_X(lq_X(p_2), lq_Y(p_2)) = F_X^o(y) - F_Y(x) = F_X(y) - F_Y(x) \geq p_2 - p_1 > \epsilon,$$

which is a contradiction. Note that we have used continuity of X in the second equality. ■

Remark. This is not true in general. Consider X with $P(X = 0) = 1$ and Y with $P(Y = 1) = 1$. Then $F_X(1/2) - F_Y(1/2) = 1$ and $SQD_{\delta_X}(X,Y) + SQD_{\delta_Y}(X,Y) = 0$.

In the next theorem we show that if the quantile distance between two variables are zero and one of them is continuous then they are identically distributed.

Theorem 9.4.10 *Suppose F_1, F_2 distribution functions, F_1 continuous and their quantile distance is zero. In other words,*

$$\sup_{p\in(0,1)} \delta_{F_1}(lq_{F_1}(p), lq_{F_2}(p)) = 0.$$

Then $F_1 = F_2$.

Proof Suppose the result does not hold. Then we have two cases.

Case I: $\exists x,\ p_1 = F_1(x) < F_2(x) = p_2$.

$$F_1(x) = p_1 \Rightarrow lq_{F_1}(p_2) = y > x,$$

and

$$F_2(x) = p_2 \Rightarrow lq_{F_2}(p_2) = z \leq x.$$

Hence

$$\delta_{F_1}(lq_{F_1}(p_2), lq_{F_2}(p_2)) = F_1(y) - F_1(z) \geq F_1(y) - F_1(x) \geq p_2 - p_1.$$

Case II: $\exists x,\ p_1 = F_1(x) > F_2(x) = p_2$.
Take $p_3 \in (p_2, p_1)$. Then

$$F_1(x) = p_1 \Rightarrow lq_{F_1}(p_3) = y \leq x.$$

However if $lq_{F_1}(p_3) = x$, we conclude

$$F_1(lq_{F_1}(p_3)) = F_1(x) \Rightarrow p_3 = p_1,$$

which is a contradiction. Note that we have used the continuity of F_1 in $F_1(lq_{F_1}(p_3)) = p_3$.
Also

$$F_2(x) = p_2 \Rightarrow lq_{F_2}(p_3) = z > x.$$

Hence

$$\delta_{F_1}(lq_{F_1}(p_3), lq_{F_2}(p_3)) = \delta_{F_1}(y, z) = F_1(z) - F_1(y) \geq F_1(x) - F_1(y) \geq p_1 - p_3.$$

■

Here we prove an easy lemma regarding the continuity of δ.

Lemma 9.4.11 *Suppose F is a continuous distribution function. For any fixed $b \in \mathbb{R}$, $\delta_F(a, b)$ is a continuous function in a.*

Proof Note that $\delta_F(a, b) = |F(b) - F(a)|$ because F is a continuous function. ■

Lemma 9.4.12 *Suppose F_1, F_2 are distribution functions, F_1 is continuous and*

$$\delta_{F_1}(lq_{F_1}(p_0), lq_{F_2}(p_0)) = \Delta > 0,$$

for some $p_0 \in (0, 1)$ then there exist $0 < \epsilon < p_0$ such that

$$\delta_{F_1}(lq_{F_1}(p), lq_{F_2}(p)) > \Delta/3,\ p \in (p_0 - \epsilon, p_0).$$

Proof Since F_1 is continuous

$$\delta_{F_1}(lq_{F_1}(p), lq_{F_2}(p)) = |p - F_1(lq_{F_2}(p))|.$$

Let $lq_{F_2}(p_0) = x_1$ and $F_1(x_1) = p_1$. Then $|p_0 - p_1| = \Delta$.
By continuity of F_1 there exist $\epsilon' > 0$ such that

$$x \in (x_1 - \epsilon', x_1 + \epsilon') \Rightarrow F_1(x) \in (p_1 - \frac{\Delta}{3}, p_1 + \frac{\Delta}{3}).$$

By left continuity of lq_{F_2} for ϵ' positive, there exists an $0 < \epsilon < \min(\Delta/3, p_0)$ such that

$$p \in (p_0 - \epsilon, p_0) \Rightarrow lq_{F_2}(p) \in (x_1 - \epsilon', x_1).$$

Hence for $p \in (p_0 - \epsilon, p_0)$, we have $F_1(lq_{F_2}(p)) \in (p_1 - \Delta/3, p_1 + \Delta/3)$. Hence

$$\delta_{F_1}(lq_{F_1}(p), lq_{F_2}(p)) = |p - F_1(lq_{F_2}(p))| \geq$$

$$|p_0 - p_1| - \epsilon - \frac{\Delta}{3} \geq \Delta/3.$$

■

Lemma 9.4.13 *Suppose F_1, F_2 are distribution functions and F_1 is continuous. Also assume*

$$IDQ_{\delta_{F_1}}(F_1, F_2) = \int_0^1 \delta_{F_1}(lq_{F_1}(p), lq_{F_2}(p)) = 0.$$

Then $F_1 = F_2$.

Proof The assumption implies that $\delta_{F_1}(lq_{F_1}(p), lq_{F_2}(p)) = 0,\ \forall p \in (0, 1)$. For otherwise if $\delta_{F_1}(lq_{F_1}(p_0), lq_{F_2}(p_0)) = \Delta > 0$, for some p_0. By the previous lemma there exist $0 < \epsilon < p_0$ such that

$$\delta_{F_1}(lq_{F_1}(p), lq_{F_2}(p)) > \Delta/3,\ p \in (p_0 - \epsilon, p_0).$$

This implies that

$$\int_0^1 \delta_{F_1}(lq_{F_1}(p), lq_{F_2}(p)) \geq \epsilon\Delta,$$

which is a contradiction. Now we can use Lemma 9.4.10 to conclude $F_1 = F_2$.
■

9.4.4 Quantile distance for continuous variables

From now on we only consider continuous variables and the probability loss function with $c = 0$, δ_X. Some results can be generalized to the general distributions but we leave that for future research. We use the simpler notations:

$$QD_X(X, X_\theta) = LQD_{\delta_X}(X, X_\theta) = \int_0^1 \delta_X(lq_X(p), lq_{X_\theta}(p))dp.$$

Also

$$QD(X, X_\theta) = QD_X(X, X_\theta) + QD_{X_\theta}(X, X_\theta).$$

Quantile distance in the continuous case can be obtained by:

$$QD_X(X, X_\theta) = \int_0^1 \delta_X(lq_X(p), lq_{X_\theta}(p))dp =$$
$$\int_0^1 |F_X \circ lq_X(p) - F_X \circ lq_{X_\theta}(p)|dp = \int_0^1 |p - F_X \circ lq_{X_\theta}(p)|dp.$$

We can also consider the quantile distance closeness in the tails. Consider the tails to correspond to probabilities $E = (0, 0.025) \cup (0.0975, 1)$. Then $\mathcal{L}(E) = 0.05$ ($\mathcal{L}$ being the Lèbegue measure) and we can define

$$QD_X^{tail}(X, X_\theta) = \int_E \delta_X(lq_X(p), lq_{X_\theta}(p))dp/0.05 =$$
$$\int_E |F_X \circ lq_X(p) - F_X \circ lq_{X_\theta}(p)|dp/0.05 = \int_E |p - F_X \circ lq_{X_\theta}(p)|dp/0.05.$$

We have divided the integral by 0.05 the length of E to make this measure comparable to the overall measure over [0,1], which has length 1.

Then we compute the quantile distance of the standard normal to some known distributions. Both the overall quantile distance and the tail quantile distance are calculated (by approximating the integrals) and the results are given in Table 9.1 and 9.2. For the overall quantile distance we observe that QD_X and QD_Y have almost the same value. A theoretical result regarding this observation is desirable and we leave this for future research. This is not true in general for the tail distance.

Then we find the closest Cauchy with scale parameter in (0,4) (and location parameter=0) to the standard normal. Once using the quantile distance and once using the tail quantile distance. We find the quantile distance of

the standard normal to all Cauchy distributions with scale parameters on the grid $(0.01, 0.02, \cdots, 4.00)$ (and location parameter=0). The results are given in Figures 9.3 and 9.5 respectively. For the overall quantile distance the optimal Cauchy is the one with scale parameter 0.66 and for the tail quantile distance, the optimal Cauchy is the one with scale parameter 0.12. Figure 9.4 depicts the normal distribution functions compared with a few Cauchy distributions including the optimal and Figure 9.6 depicts the normal distribution in the upper tail with a few Cauchy distributions including the optimal in tails with scale parameter 0.12. Figure 9.7 depicts the standard normal distribution compared with the optimal Cauchy for the overall quantile distance and the optimal Cauchy for the tail quantile distance. We conclude that a fit that is optimally might not be optimal on the tails. We use this fact later in choosing our method to model extreme temperature events.

Distribution	$QD_X(X,Y)$	$QD_Y(X,Y)$	QD
$Y = N(1,1)$	0.2605080	0.2605080	0.5210159
$Y = N(0.5,1)$	0.138301	0.138301	0.276602
$Y = N(0,2)$	0.1024215	0.1024207	0.2048422
$Y = t(1)$	0.06382985	0.0637436	0.1275734
$Y = t(10)$	0.0078747	0.007872528	0.01574723
$Y = t(100)$	0.000795163	0.0007951621	0.001590325
$Y = Cauchy(scale = 1)$	0.06376941	0.06376579	0.1275352
$Y = \chi^2(1)$	0.2190132	0.2190249	0.4380381
$U(-0.5, 0.5)$	0.1522836	0.1522991	0.3045827
$U(-1,1)$	0.06562216	0.06563009	0.1312522
$U(-2,2)$	0.05612716	0.0561283	0.1122555
$U(-3,3)$	0.1171562	0.1171562	0.2343124

Table 9.1: Comparing standard normal with various distributions using quantile distance, where U denotes the uniform distribution and χ^2 the Chi-squared distribution.

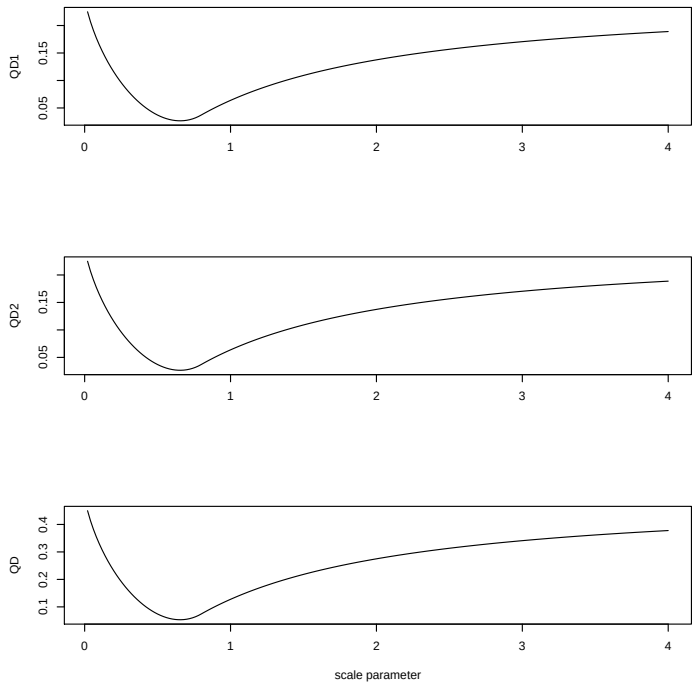

Figure 9.3: Cauchy distribution's distance with different scale parameter (and location parameter=0) to the standard normal. In the plots $QD_1 = Q_X$ and $QD_2 = QD_Y$ and $QD = QD1 + QD2$, where X is the standard normal and Y is the Cauchy.

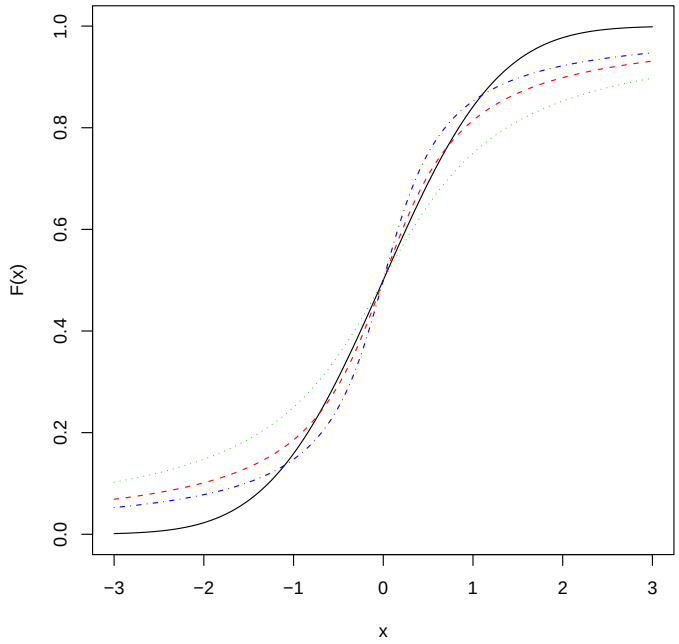

Figure 9.4: The distribution function of standard normal (solid) compared with the optimal Cauchy (and location parameter=0) picked by quantile distance minimization with scale parameter=0.66 (dashed curve), Cauchy with scale parameter=1 (dotted) and Cauchy with scale parameter=0.5 (dot dashed).

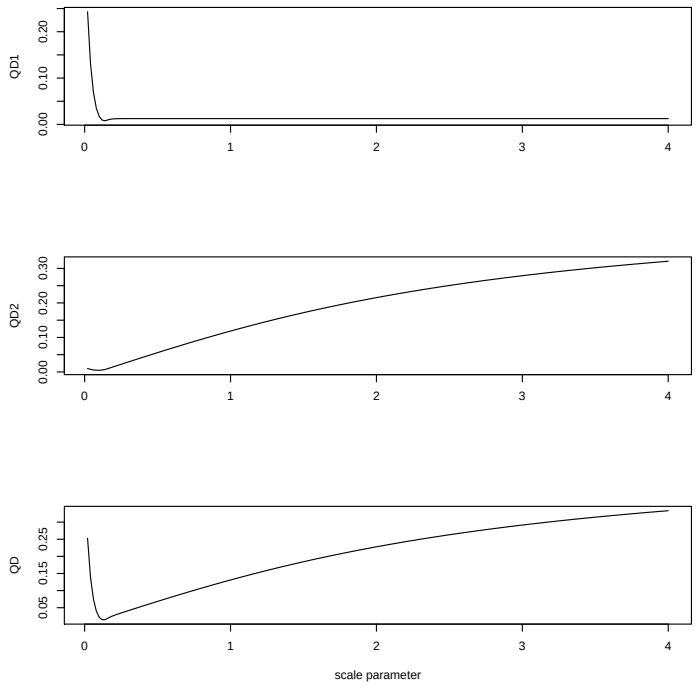

Figure 9.5: Cauchy distribution's distance with different scale parameter (and location parameter=0) to the standard normal on the tails. In the plots $QD_1 = Q_X$ and $QD_2 = QD_Y$ and $QD = QD1 + QD2$, where X is the standard normal and Y is the Cauchy.

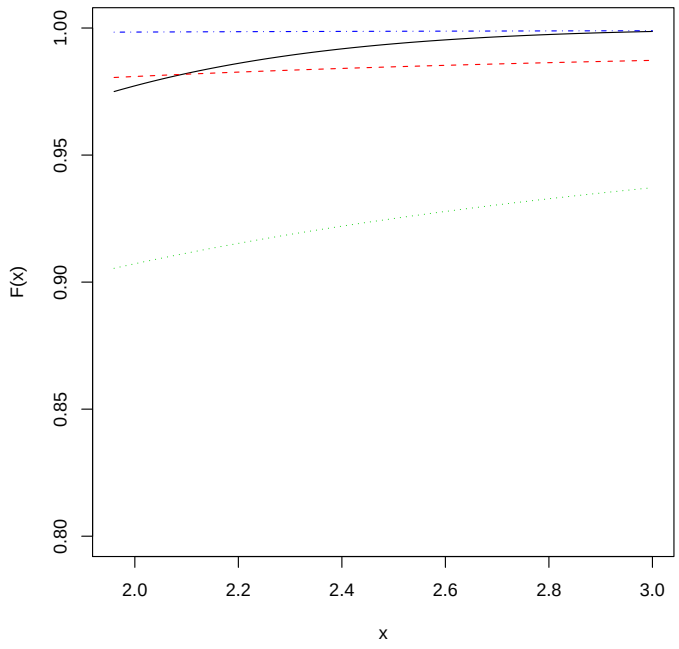

Figure 9.6: The distribution function of standard normal (solid) compared with the optimal Cauchy picked by tail quantile distance minimization with scale parameter=0.12 (dashed curve), Cauchy with scale parameter=0.65 (dotted) and Cauchy with scale parameter=0.01 (dot dashed).

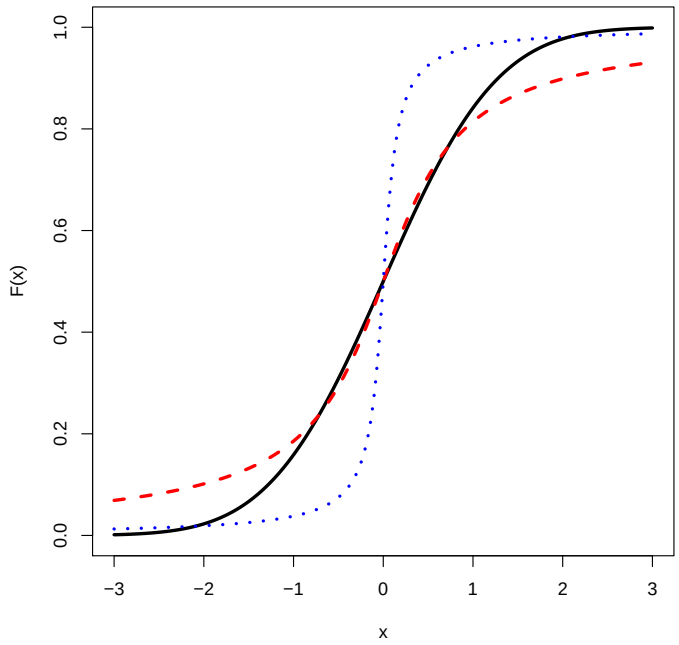

Figure 9.7: Comparing the standard normal distribution (solid) with optimal Cauchy picked by quantile distance (dashed) and the optimal Cauchy picked by tail quantile distance minimization (dotted).

Distribution	$QD_X^{tail}(X,Y)$	$QD_Y^{tail}(X,Y)$	$QD^{tail}(X,Y)$
$Y = N(1,1)$	0.05075276	0.05075276	0.10150552
$Y = N(0.5,1)$	0.01824013	0.01824013	0.03648026
$Y = N(0,2)$	0.01249034	0.11206984	0.12456018
$Y = t(1)$	0.0125000	0.1184949	0.1309949
$Y = t(10)$	0.007631262	0.011192379	0.018823642
$Y = t(100)$	0.0009740074	0.0010122519	0.0019862594
$Cauchy(scale = 1)$	0.0125000	0.1180231	0.1305231
$Y = \chi^2(1)$	0.25006521	0.06467072	0.31473593
$U(-0.5,0.5)$	0.3004565	0.0125000	0.3129565
$U(-1,1)$	0.1523052	0.0125000	0.1648052
$U(-2,2)$	0.01313629	0.01205279	0.02518908
$U(-3,3)$	0.01083494	0.10054194	0.11137688

Table 9.2: Comparing standard normal on the tails with some distributions using quantile distance, where U denotes the uniform distribution and χ^2 the Chi-squared distribution.

9.4.5 Equivariance of estimation under monotonic transformations using the quantile distance

Suppose a family of distributions $\{X_\theta\}_{\theta\in\Theta}$, $\Theta \subset \mathbb{R}^k$ is given. Also assume ϕ is a continuous and strictly monotonic transformation on $\mathbb{R}$. Consider the family of distributions $\{Y_\theta = \phi(X_\theta)\}_{\theta\in\Theta}$. Then the family $\{Y_\theta\}_{\theta\in\Theta}$ is parameterized by the same parameters since

$$P(Y_\theta < a) = P(\phi(X_\theta) < a) = P(X_\theta < \phi^{-1}(a)).$$

Then the following lemma shows the equivariance property of quantile distance estimation.

Lemma 9.4.14 *Suppose a random variable X and a family of distributions $\{X_\theta\}_{\theta\in\Theta}$ are given,*

$$A = argmin_{\theta\in\Theta} \int_0^1 \delta_X(lq_X(p), lq_{X_\theta}(p))dp,$$

is nonempty and ϕ is a continuous and strictly monotonic transformation. Let

$$B = argmin_{\theta\in\Theta} \int_0^1 \delta_{\phi(X)}(lq_{\phi(X)}(p), lq_{\phi(X_\theta)}(p))dp.$$

Then $A = B$. In other words if X_θ is an optimal estimator of X, then $\phi(X_\theta)$ is an optimal estimator of $\phi(X)$.

Proof This is trivial by invariance properties of quantile distance under continuous strictly monotonic transformations. ■

Remark. The above is also true if we use replace the integral quantile distance by the sup quantile distance.

9.4.6 Estimation using quantile distance

Here we only consider estimation using integral quantile distance. In order to estimate a distribution X using a parameterized family $\{X_\theta\}_{\theta\in\Theta}$, one can try to find

$$argmin_{\theta\in\Theta}\int_0^1 \delta_X(lq_X(p), lq_{X_\theta}(p))dp.$$

However, the above expression depends on δ_X an unknown. The available information to us is usually a random sample $X_1, \cdots, X_n$.
Remark. If we use the empirical distribution instead of the distribution of X is above, we get:

$$argmin_{\theta\in\Theta}\int_0^1 \delta_{F_n}(lq_{F_n}(p), lq_{X_\theta}(p))dp.$$

The *argmin* can be checked again to be equivariant under continuous and strictly monotonic transformations.

Tables 9.3 and 9.4 compare the maximum likelihood estimation to the quantile distance estimation method for a sample of size $N = 20$ and $N = 100$ respectively. In each case we generate 50 samples of length N and estimate the parameters using both methods. Then we assess the performance by a few measures: mean absolute error, mean square error, mean probability loss error and mean quantile distance. In both cases maximum likelihood has done slightly better in terms of all errors except the quantile distance error in which case the quantile distance estimation has done significantly better. The histogram for both estimation methods for $N = 20$ and $N = 100$ are given in Figures 9.8 and 9.9 respectively. For both maximum likelihood and quantile distance estimations for $N = 100$ the parameters have a symmetric (close to normal) distribution.

Error type	QD error	s.e. of QD error	ML error	s.e. ML error
Mean probability loss error for $\mu = lq_{N(\mu,\sigma^2)}(1/2)$	0.077	0.061	0.077	0.055
Mean probability loss for $\sigma^2 + \mu = lq_{N(\mu,\sigma^2)}(P(Z < 1))$	0.185	0.114	0.176	0.096
Mean abs. error for μ	0.198	0.160	0.196	0.143
Mean abs. error for σ	0.159	0.127	0.132	0.085
Mean square error μ	0.064	0.089	0.058	0.077
Mean square error for σ	0.041	0.065	0.025	0.028
Mean QD error	0.035	0.009	0.122	0.073

Table 9.3: Assessment of Maximum likelihood estimation and quantile distance estimation using several measures of error for a sample of size 20. In the table *s.e.* stands for the standard error.

Error type	QD error	*s.e.* of QD error	ML error	s.e. ML error
Mean probability loss for $\mu = lq_{N(\mu,\sigma^2)}(1/2)$	0.028	0.020	0.027	0.020
Mean probability loss for $\sigma^2 + \mu = lq_{N(\mu,\sigma^2)}(P(Z < 1))$	0.157	0.046	0.165	0.038
Mean abs. error for μ	0.070	0.051	0.068	0.051
Mean abs. error for σ	0.079	0.052	0.061	0.039
Mean square error μ	0.007	0.009	0.007	0.009
Mean square error for σ	0.009	0.011	0.005	0.005
Mean QD error	0.014	0.003	0.045	0.026

Table 9.4: Assessment of Maximum likelihood estimation and quantile distance estimation using several measures of error for a sample of size 100. In the table *s.e.* stands for the standard error.

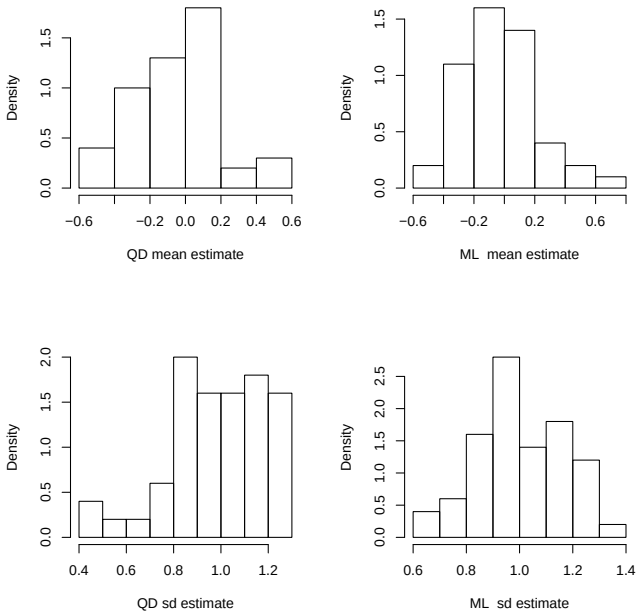

Figure 9.8: Histograms for the parameter estimates using quantile distance and maximum likelihood methods for a sample of size 20.

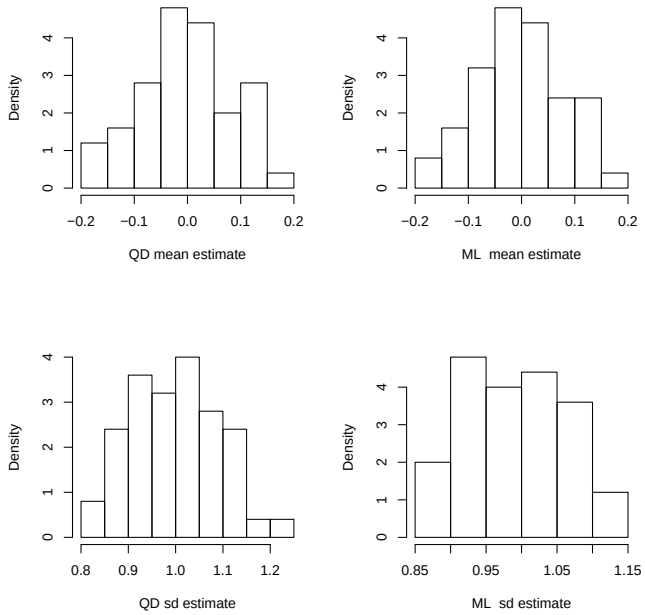

Figure 9.9: Histograms for the parameter estimates using quantile distance and maximum likelihood methods for a sample of size 100.

Chapter 10

Binary temperature processes

10.1 Introduction

This chapter uses the theory developed in previous chapters to find appropriate models for extreme temperature events. We consider both low and high temperatures. The temperature is measured in degrees centigrade. We define a day with minimum temperature (mt) less than zero as extremely cold and denote it by e:

$$e(t) = \begin{cases} 1 & mt(t) \leq 0 \text{ (deg C)} \\ 0 & mt(t) > 0 \text{ (deg C)} \end{cases}.$$

Taking 0 (deg C) to be the cut–off for low temperature seems reasonable in the absence of any other considerations, since it is the usual definition of a frost. In agriculture, where most plants contain a lot of water this can be considered as an important cut–off. No seemingly natural cut-off like that for minimum temperature exists for extremely high temperature. To define extreme events, we ask the following questions:

1. Should the definition of an extreme event depend on the purpose of our model?
2. Should it depend on the time of the year and location?
3. What should be the cut–off (threshold) to define an extreme event?
4. Should we use a certain quantile as the cut-off? In that case which quantile should be used?

We provide some answers in the following:

1. The answer to the first question is clearly affirmative. For example, a high temperature day for agriculture purposes is different from energy

providing purposes. Even for the farmer, different crops may have different tolerances to hot or cold weather.

2. The answer of the second question depends on the model's purpose. We might want to vary the definition over time and space for some purposes.

3. We do not know of any such natural cut–off for high temperatures like that for low temperature.

4. Quantiles have long been used to determine the extreme events. Choosing the level of the quantile depends on the purpose. Some extreme–value modelers pick the quantile high enough to insure the validity of the assumptions underlying their models as Embrechts et al. discuss in [16]. For example, a well–known result asserts that $P(X - u < v|X > u)$ follows a known distribution (extreme value distributions e.g. Pareto) when u is large. [See [16].] We do not favor such methods of choosing the threshold. The threshold should be picked primarily to reflect our needs in the real problem rather than satisfy the assumptions of the models. If the models do not satisfy the conditions, we should find others rather than move the threshold up.

Based on the above discussion with the statistician's knowledge alone, one cannot define the extreme events. Ralph Wright (personal communication) in AAFRD (Agriculture and Rural Development in Alberta, Canada) raises similar points. In particular he said the following about the droughts:

"Drought is really defined by the impact that the moisture deficit has on a specific use or uses. Its definition can vary both with time of year and from place–to–place. Drought can be short–term or long–term. For example, one month of hot dry weather can significantly reduce crop yields, despite the fact that normal amounts of precipitation have been received over the past year. On the other hand, crops may do fine in dry weather conditions if precipitation has been received in a timely manner and temperatures have been favorable. However under the same conditions, a dam operator in the same area may have severe shortages in the reservoir and declare drought like conditions (e.g. with low winter snow–fall and poor spring run–off). You will need to define your drought based on whom or what is being impacted by the water shortage."

Since we do not have any standard definition of an extremely hot day, we use the data. In our example, to define a binary process of (hot)/(not hot) for temperature, we pick the global spatial/temporal 95th percentile using

the data from 25 stations over Alberta that had daily maximum temperature (MT) data from 1940 to 2004. The 95th percentile was computed using the quantile algorithm developed in previous chapters and turned out to be 26.7. The exact value was also found and turned out to be $q = 27$ (deg C). Then We define the binary process of extremely hot temperature as:

$$E(t) = \begin{cases} 1 & MT(t) \geq q \\ 0 & MT(t) < q \end{cases},$$

where $q = 27$ (deg C) here.

In order to study extreme events (e.g. for MT) three approaches come to mind:

1. Model the whole daily MT process and use that to infer about the extremes. For MT, we have shown that a Gaussian distribution fits the daily values fairly well. However, in the tails, usually of paramount concern, the fit does not do well as shown in the qq–plots in Chapter 2. Another difficulty with this approach is picking a covariance function to model the covariance over time. Also in Chapter 9, we showed that even though two distributions are very close in terms of overall quantile distance, they might not be very close in terms of tail quantile distance (Figure 9.7). This shows in order to study extremes (for example extremely hot temperature) if we use a good overall fit, our results might not be reliable.

2. Use a specified threshold and model the values exceeding the threshold. This approach has several drawbacks. Firstly we cannot answer the question of how often or in what periods of the year the extremes happen. This is because we model the actual extreme values and ignore the non–extreme values. Secondly, strong assumption of independence is needed for this method. Thirdly we need to pick the threshold high enough to make the model reasonable as mentioned before. This might not be an optimal threshold from a practical point of view.

3. Based on a real problem, use a threshold to define a new binary process of (extreme)/(not extreme) values and then model that binary process. This is the method we use and it does not have the issues mentioned in 1 and 2 because the threshold is not taken to satisfy some statistical property and we make few assumptions about the binary chain.

10.2 rth–order Markov models for extreme minimum temperatures

This section looks for appropriate models for the binary process $e(t)$ of cold/not cold temperature days. This is a binary process and the Categorical Expansion Theorem (Theorem 3.5.6) gives the form of all such rth–order Markov chains. Here we also consider other covariates such as the minimum temperature of the previous day and two days ago as well as seasonal covariates (deterministic). The next subsection uses graphical tools and exploratory techniques to investigate the properties the model should have. Then we use the BIC criterion and compare several proposed models. We use partial likelihood techniques to estimate parameters as proposed by Kedem et al. in [27].

10.2.1 Exploratory analysis for binary extreme minimum temperatures

Here we perform an exploratory analysis of the binary process $e(t)$ using two stations for this purpose, Banff and Medicine Hat which have data from 1895 to 2006. The transition probabilities are computed from the historical data considering years as independent observations. The results are summarized a follows:

- Figures 10.1 and 10.2 plot the probability of a freezing day over the course of a year for the Banff and Medicine Hat stations, respectively. A regular seasonal pattern is seen. Medicine Hat seems to have a much longer frost–free period.

- Figures 10.3 and 10.4 plot the estimated transition probabilities, $\hat{p}_{01}$ and $\hat{p}_{11}$ for the Banff and Medicine Hat stations. If the chain were a 0th–order Markov chain then these two curves would overlap. This is not the case and Markov chain at least of 1st–order seems necessary. In the $\hat{p}_{01}$ curve for both Banff and Medicine Hat, high fluctuations are seen at the beginning and end of the year which corresponds to the cold season. This is not surprising because there are very few pairs in the data with a freezing day followed by a non–freezing day in a cold season in Alberta.

- In Figure 10.4, $\hat{p_{11}}$ is missing for a period over the summer. This is because no freezing day is observed over this period in the summer and hence $\hat{p}_{11}$ could not be estimated.

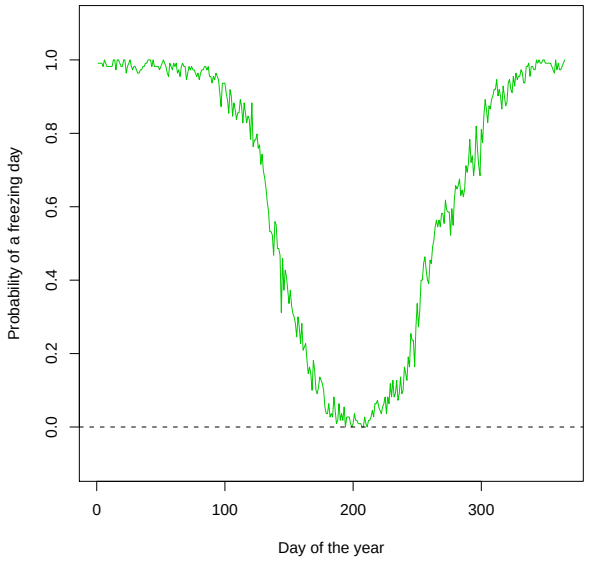

Figure 10.1: The estimated probability of a freezing day for the Banff site for different days of a year computed using the historical data.

- Figures 10.5 and 10.6 give the plots for the 2nd–order transition probabilities. They overlap substantially and hence a 2nd–order Markov chain does not seem to be necessary.

10.2.2 Model selection for extreme minimum temperature

This section finds models for the extreme minimum temperature process $e(t)$. Here Z_{t-1} denotes the covariate process. We investigate the following predictors:

- $e^k(t) \equiv e(t-k)$. Was it an extremely cold day k days ago?
- $mt^k(t) \equiv mt(t-k)$, the actual minimum temperature k days ago.
- N^k, the number of freezing days during the k previous days.
- SIN, COS, $SIN2$ and $COS2$ which are abbreviations for $\sin(\omega t)$, $\cos(\omega t)$, $\sin(2\omega t)$ and $\cos(2\omega t)$, respectively (with $\omega = \frac{2\pi}{366}$).

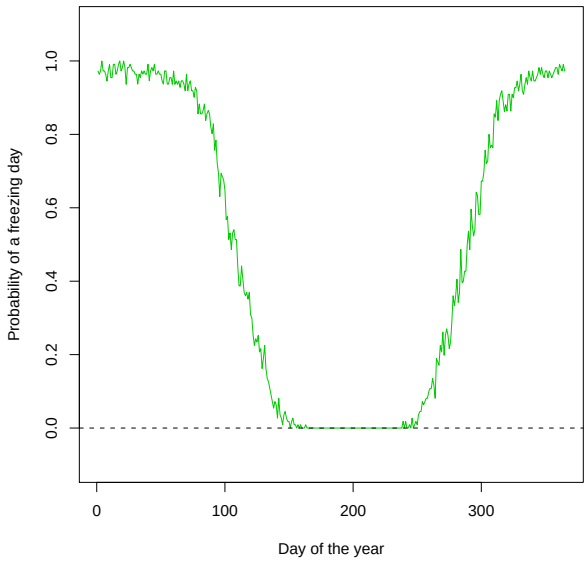

Figure 10.2: The estimated probability of a freezing day for the Medicine Hat site for different days of a year computed using the historical data.

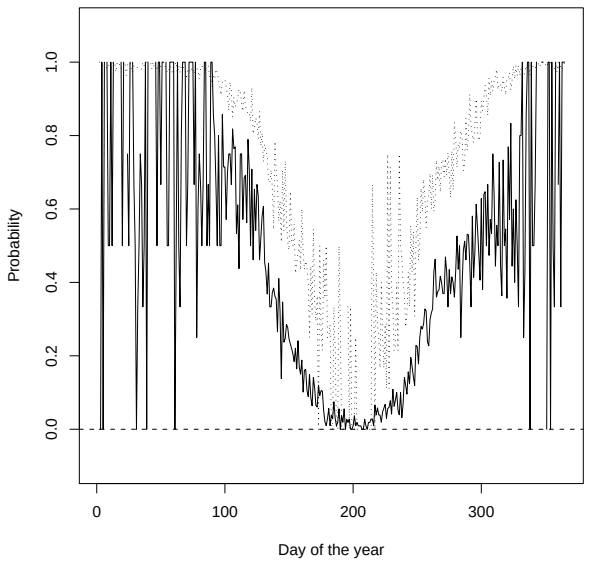

Figure 10.3: The estimated 1st–order transition probabilities for the 0-1 process of extreme minimum temperatures for the Banff site. The dotted line represents the estimated probability of "$e(t) = 1$ if $e(t-1) = 1$" ($\hat{p_{11}}$) and the dashed, "$e(t) = 1$ if $e(t-1) = 0$" ($\hat{p_{01}}$).

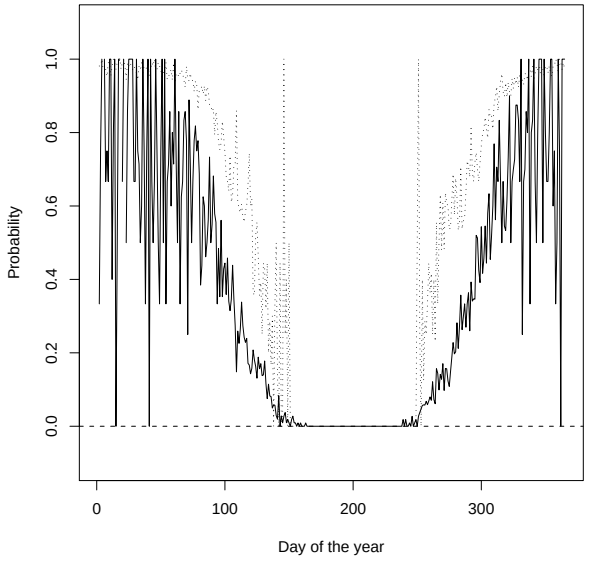

Figure 10.4: The estimated 1st–order transition probabilities for the 0-1 process of extreme minimum temperatures for the Medicine Hat site. The dotted line represents the estimated probability of "$e(t) = 1$ if $e(t-1) = 1$" ($\hat{p_{11}}$) and the dashed, "$e(t) = 1$ if $e(t-1) = 0$" ($\hat{p_{01}}$).

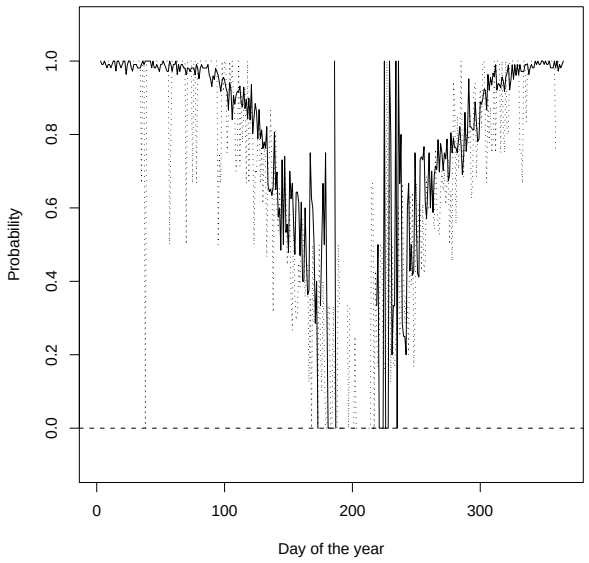

Figure 10.5: The estimated 2nd–order transition probabilities for the 0-1 process of extreme minimum temperature for the Banff site with $\hat{p}_{111}$ (solid) compared with $\hat{p}_{011}$ (dotted) both calculated from the historical data.

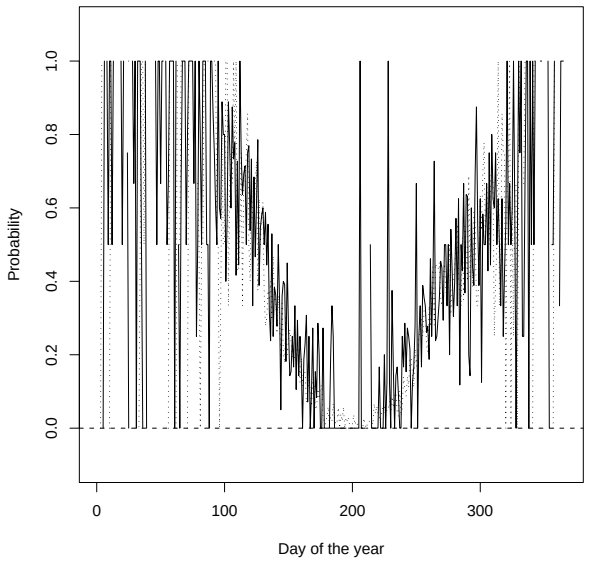

Figure 10.6: The estimated 2nd–order transition probabilities for the 0-1 process of extreme minimum temperatures for the Banff site with $\hat{p}_{001}$ (solid) compared with $\hat{p}_{101}$ (dotted) calculated from the historical data.

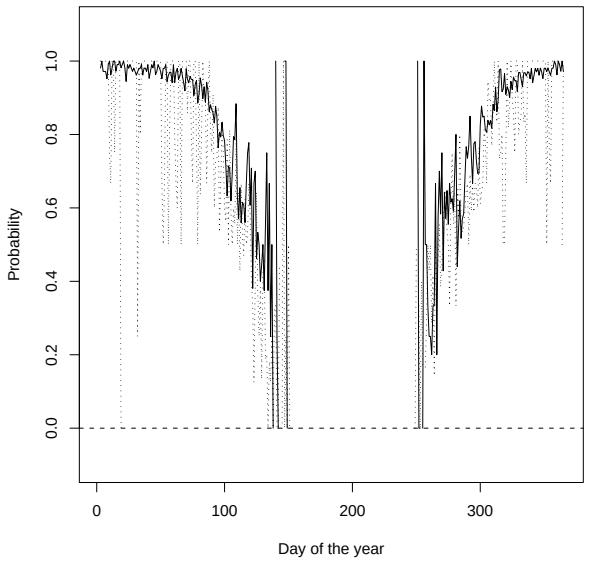

Figure 10.7: The estimated 2nd–order transition probabilities for the 0-1 process of extreme minimum temperatures for the Medicine Hat site with $\hat{p}_{111}$ (solid) compared with $\hat{p}_{011}$ (dotted) calculated from the historical data.

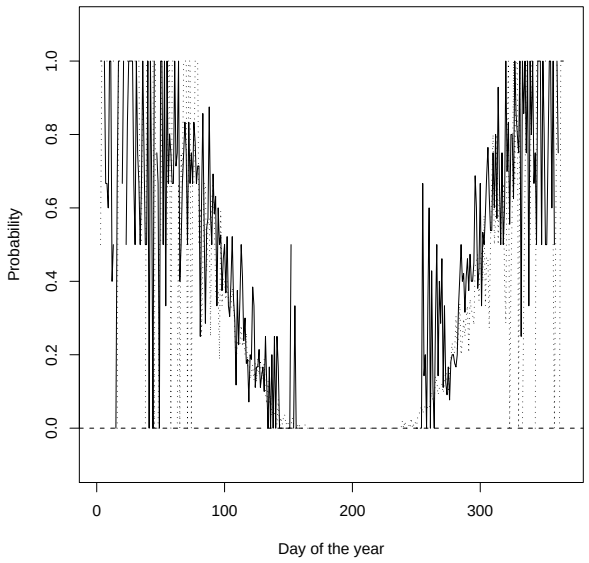

Figure 10.8: The estimated 2nd–order transition probabilities for the 0-1 process of extreme minimum temperatures for the Medicine Hat site with $\hat{p}_{001}$ (solid) compared with $\hat{p}_{101}$ (dotted) calculated from the historical data.

Table 10.1 compares models with a constant and N^k as the covariate process. The optimal model picked by the BIC criterion is the model with the covariates $Z_{t-1} = (1, N^{11})$.

Model: Z_{t-1}	BIC	parameter estimates
$(1, N^1)$	1251.7	(-2.144, 4.260)
$(1, N^2)$	1166.5	(-2.501, 2.490)
$(1, N^3)$	1142.9	(-2.653, 1.755)
$(1, N^4)$	1121.6	(-2.773, 1.371)
$(1, N^5)$	1111.2	(-2.852, 1.125)
$(1, N^6)$	1093.1	(-2.932, 0.961)
$(1, N^7)$	1087.4	(-2.977, 0.835)
$(1, N^8)$	1081.7	(-3.015, 0.739)
$(1, N^9)$	1077.1	(-3.047, 0.663)
$(1, N^{10})$	1066.5	(-3.089, 0.605)
$(1, N^{11})$	**1056.4**	(-3.130, 0.557)
$(1, N^{12})$	1059.5	(-3.135, 0.511)
$(1, N^{13})$	1062.3	(-3.140, 0.472)
$(1, N^{14})$	1072.8	(-3.126, 0.437)
$(1, N^{15})$	1080.9	(-3.118, 0.406)
$(1, N^{16})$	1091.9	(-3.102, 0.379)
$(1, N^{17})$	1104.2	(-3.083, 0.354)
$(1, N^{18})$	1112.1	(-3.075, 0.334)
$(1, N^{19})$	1118.6	(-3.068, 0.315)
$(1, N^{20})$	1126.5	(-3.058, 0.299)

Table 10.1: BIC values for models including N^k for the extreme minimum temperature process $e(t)$ at the Medicine Hat site.

Model: Z_{t-1}	BIC	parameter estimates
(1)	2539.9	(-0.0251)
$(1, e^1)$	1251.7	(-2.144, 4.260)
$(1, e^2)$	1473.6	(-1.856, 3.683)
$(1, e^1, e^2)$	1157.7	(-2.501, 3.085, 1.896)
$(1, e^1, e^2, e^1e^2)$	1162.4	(-2.586, 3.389, 2.190, -0.593)
$(1, mt^1)$	963.7	(0.109, -0.400)
$(1, mt^1, mt^2)$	954.0	(0.091, -0.329, -0.082)
$(1, COS, SIN)$	984.0	(-0.070, 4.292, 1.324)
$(1, COS, SIN, COS2, SIN2)$	984.2	(-0.502, 4.505, 1.399, -0.464, -0.493)
$(1, COS, SIN, COS2)$	986.7	(-0.258, 4.359, 1.335, -0.353)
$(1, COS, SIN, SIN2)$	984.4	(-0.217, 4.365, 1.360, -0.402)
$(1, mt^1, mt^2, mt^3)$	940.7	(0.062, -0.319, -0.009, -0.094)
$(1, mt^1, mt^2, mt^1mt^2)$	943.4	(0.211, -0.339, -0.084, -0.0091)
$(1, e^1, COS, SIN)$	901.5	(-1.008, 1.840, 3.325, 1.013)
$(1, mt^1, COS, SIN)$	**855.3**	(-0.074, -0.234, 2.394, 0.746)
$(1, mt^1, mt^2, COS, SIN)$	861.9	(-0.076, -0.247, 0.023, 2.504, 0.785)

Table 10.2: BIC values for several models for the extreme minimum temperature $e(t)$ at the Medicine Hat site.

Table 10.2 compares several models some of which include seasonal terms and continuous variables. The optimal model is $(1, mt^1, COS, SIN)$, which has the temperature of the previous day and seasonal terms. The model $(1, e^1, COS, SIN)$ has a larger BIC but is preferable to all models other than $(1, mt^1, COS, SIN)$ and $(1, mt^1, mt^2, COS, SIN)$. Note that it is not possible to compute the probability of events in the long-term future using $(1, mt^1, COS, SIN)$, since we do not know mt except for perhaps the present time. Hence the optimal applicable model seems to be $(1, e^1, COS, SIN)$.

10.3 rth–order Markov models for extreme maximum temperatures

This section finds appropriate models for the binary process of extremely hot temperature $E(t)$ as defined above. To define a hot day, we use the 95th percentile of data from 25 stations over Alberta that had daily MT data from 1940 to 2004. The 95th percentile turns out to be $q = 27$ (deg C). Once we used the fast algorithm developed in Chapter 7 to pick the quantile and once we used an exact method; the algorithm gave us the approximate value $q = 26.7$, which is very close to the exact value. (See Table **??** for more details on the computation.)

10.3.1 Exploratory analysis for extreme maximum temperatures

This section uses explanatory data analysis techniques to study the binary process $E(t)$. Again we use two stations for this purpose, the Banff and Medicine Hat sites that have data from 1895 to 2006. The transition probabilities are computed using the historical data considering years as independent observations. The results are summarized as follows:

- Figures 10.9 and 10.10 plot the probabilities of a hot day over the course of a year for the Banff and Medicine Hat stations respectively. A regular seasonal pattern is seen. Medicine Hat seems to have a much longer period of hot days.
- Figures 10.11 and 10.12 plot the estimated transition probabilities, $\hat{p}_{01}$ and $\hat{p}_{11}$ for Banff and Medicine Hat. If the chain were a 0th–order Markov chain then these two curves would overlap. This is not the case so Markov chain of at least 1st–order seems necessary. In the $\hat{p}_{01}$ curve for both Banff and Medicine Hat, large fluctuations are seen in the middle of the year, which corresponds to the warm season. This is not surprising because there are very few pairs in the data with a hot day followed by a not–hot day in the warm season in Alberta.
- In Figure 10.12, $\hat{p_{11}}$ is missing for a period over the cold season. This is because no hot day is observed during this period in the cold season and hence $\hat{p}_{11}$ could not be estimated.
- Figures 10.13 and 10.14 give the plots for the 2nd–order transition probabilities. They overlap heavily and hence a 2nd–order Markov chain does not seem to be necessary.

10.3.2 Model selection for extreme maximum temperature

Here, we use the following abbreviations:

- $E^k(t) = E(t-k)$. Was it an extreme day k days ago?
- $MT^k(t) = MT(t-k)$, the actual maximum temperature k days ago.
- N^k, COS, SIN, COS, $SIN2$ and $COS2$ as previous sections.

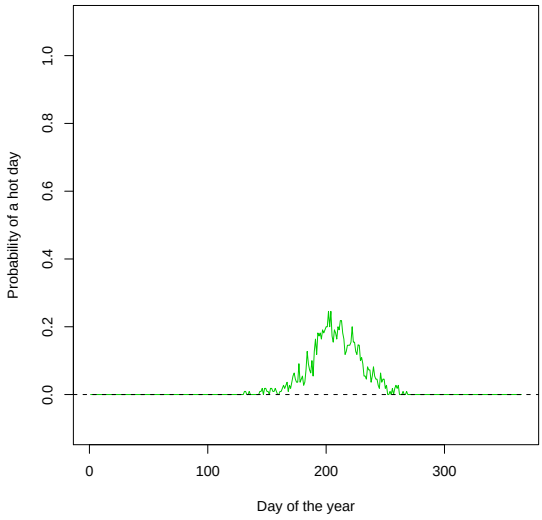

Figure 10.9: The estimated probability of a hot day (maximum temperature ≥ 27 (deg C)) for different days of the year for the Banff site calculated from the historical data.

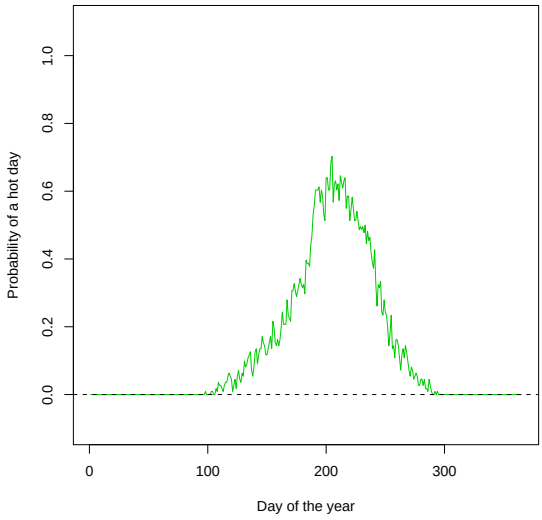

Figure 10.10: The estimated probability of a hot day (maximum temperature ≥ 27 (deg C)) for different days of the year for the Medicine Hat site calculated from the historical data.

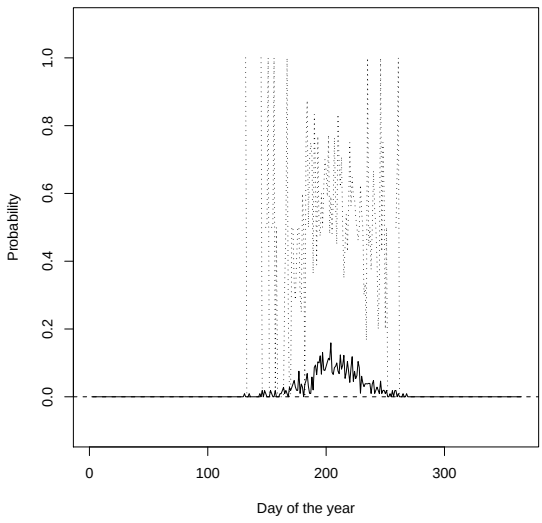

Figure 10.11: The estimated 1st–order transition probabilities for the binary process of extremely hot temperatures for the Banff site. The dotted line represent the estimated probability of "$E(t) = 1$ if $E(t-1) = 1$" ($\hat{p_{11}}$) and the dashed, "$E(t) = 1$ if $E(t-1) = 0$" ($\hat{p_{01}}$).

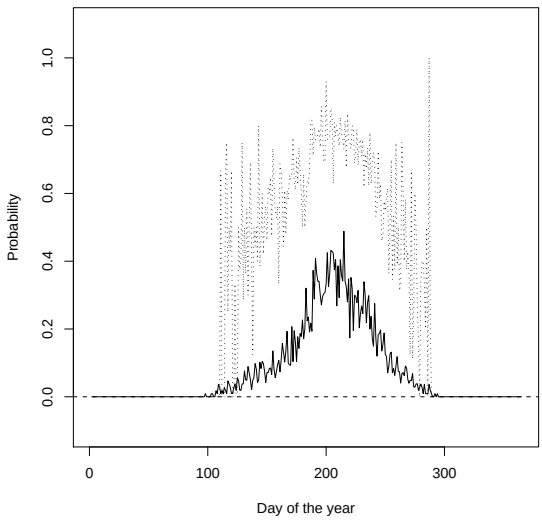

Figure 10.12: The estimated 1st–order transition probabilities for the binary process of extremely hot temperatures for the Medicine Hat site. The dotted line represents the estimated probability of "$E(t) = 1$ if $E(t-1) = 1$" ($\hat{p_{11}}$) and the dashed, "$E(t) = 1$ if $E(t-1) = 0$" ($\hat{p_{01}}$).

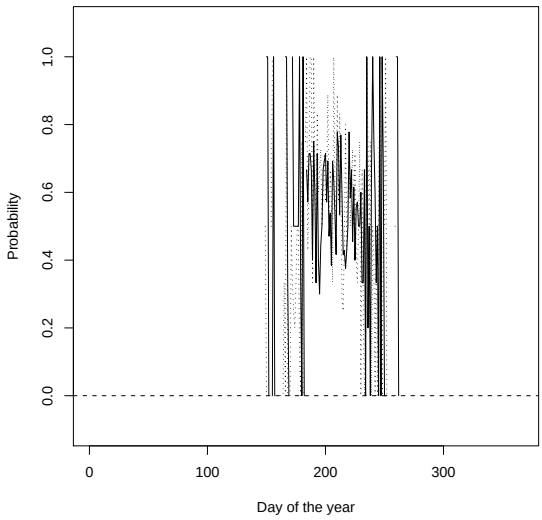

Figure 10.13: The estimated 2nd–order transition probabilities for the binary process of extremely hot temperatures for the Banff site with $\hat{p}_{111}$ (solid) compared with $\hat{p}_{011}$ (dotted) calculated from the historical data.

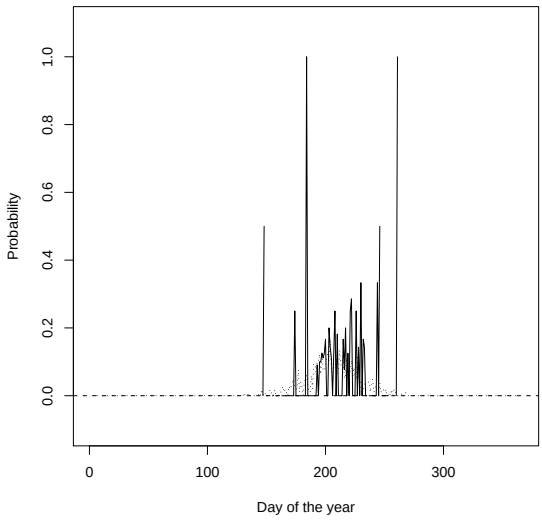

Figure 10.14: The estimated 2nd–order transition probabilities for the binary process of extremely hot temperatures for the Banff site with $\hat{p}_{001}$ (solid) compared with $\hat{p}_{101}$ (dotted) calculated from the historical data.

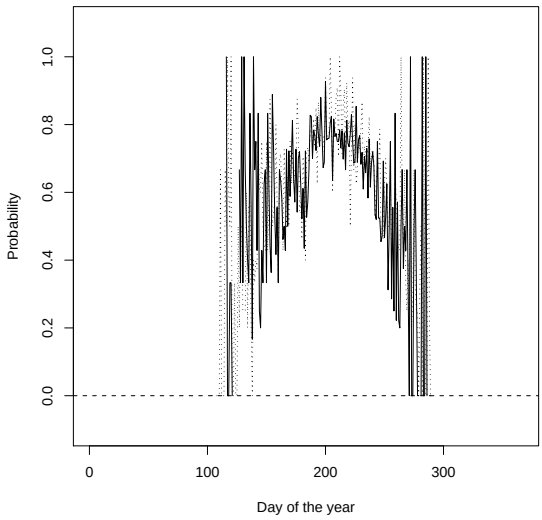

Figure 10.15: The estimated 2nd–order transition probabilities for the binary process of extremely hot temperatures for the Medicine Hat site with $\hat{p}_{111}$ (solid) compared with $\hat{p}_{011}$ (dotted), calculated from the historical data.

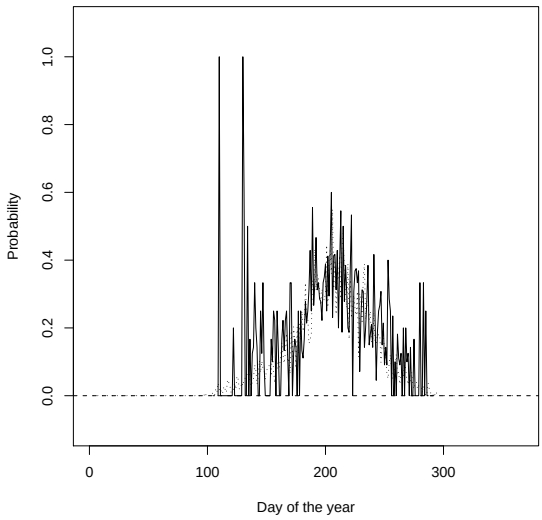

Figure 10.16: The estimated 2nd–order transition probabilities for the binary process of extremely hot temperatures for the Medicine Hat site with $\hat{p}_{001}$ (solid) compared with $\hat{p}_{101}$ (dotted) calculated from the historical data.

Table 10.3 compares several models containing N^k. The optimal model turns out to be $(1, N^{11})$ which is the same as the result for the extreme minimum temperature process $e(t)$.

Model: Z_{t-1}	BIC	parameter estimates
$(1, N^1)$	955.7	(-2.95, 3.82)
$(1, N^2)$	965.9	(-3.00, 2.16)
$(1, N^3)$	942.5	(-3.11, 1.60)
$(1, N^4)$	921.8	(-3.20, 1.29)
$(1, N^5)$	926.8	(-3.23, 1.05)
$(1, N^6)$	931.6	(-3.24, 0.89)
$(1, N^7)$	932.5	(-3.26, 0.78)
$(1, N^8)$	939.0	(-3.26, 0.69)
$(1, N^9$	931.6	(-3.29, 0.63)
$(1, N^{10})$	925.9	(-3.31, 0.57)
$(1, N^{11})$	**911.7**	(-3.35, 0.49)
$(1, N^{12})$	917.5	(-3.34, 0.46)
$(1, N^{13})$	922.8	(-3.33, 0.42)
$(1, N^{14})$	926.0	(-3.32, 0.39)
$(1, N^{15})$	932.1	(-3.31, 0.37)
$(1, N^{16})$	941.7	(-3.29, 0.34)
$(1, N^{17})$	951.5	(-3.28, 0.31)
$(1, N^{18})$	955.3	(-3.27, 0.29)
$(1, N^{19})$	960.6	(-3.26, 0.28)
$(1, N^{20})$	968.3	(-3.25, 0.26)
$(1, N^{21})$	975.3	(-3.23, 0.25)
$(1, N^{22})$	981.8	(-3.22, 0.24)
$(1, N^{23})$	986.0	(-3.22, 0.23)
$(1, N^{24})$	991.6	(-3.21, 0.22)
$(1, N^{25})$	997.0	(-3.21, 0.21)
$(1, N^{26})$	1002.8	(-3.20, 0.20)
$(1, N^{27})$	1009.5	(-3.19, 0.19)
$(1, N^{28})$	1014.4	(-3.18, 0.19)

Table 10.3: BIC values for models including N^k for the extremely hot process $E(t)$.

Table 10.4 compares several models. We observe that major reductions are seen if we use MT^k instead of E^k. The optimal model turns out to be $(1, MT^1, COS, SIN)$ which is combination of seasonal terms and the temperature of the day before.

Model: Z_{t-1}	BIC	parameter estimates
(1)	1520.3	(-1.774)
$(1, E^1)$	955.8	(-2.95, 3.82)
$(1, E^2)$	1170.5	(-2.581, 2.924)
$(1, E^1, E^2)$	941.3	(-3.034, 3.179, 1.099)
$(1, E^1, E^2, E^1E^2)$	929.0	(-3.202, 3.895, 2.137, -1.877)
$(1, MT^1)$	683.8	(-10.040, 0.362)
$(1, MT^1, MT^2)$	689.1	(-10.135, 0.333, 0.034)
$(1, COS, SIN)$	830.8	(-5.484, -5.616, -2.452)
$(1, COS, SIN, COS2, SIN2)$	837.5	(-4.343, -4.255, -0.993, 0.113, 1.016)
$(1, COS, SIN, COS2)$	837.9	(-5.850, -6.231, -2.406, -0.292)
$(1, COS, SIN, SIN2)$	830.0	(-4.481, -4.492, -0.978, 1.011)
$(1, MT^1, MT^2, MT^3)$	669.2	(-10.885, 0.338, -0.061, 0.120)
$(1, MT^1, MT^2, MT^1MT^2)$	681.9	(-21.003, 0.763, 0.452, -0.0162)
$(1, E^1, COS, SIN)$	731.3	(-4.963, 2.005, -4.096, -1.685)
$(1, MT^1, COS, SIN)$	**649.9**	(-10.281, 0.283, -2.829, -1.079)
$(1, MT^1, MT^2, COS, SIN)$	657.3	(-10.109, 0.294, -0.011, -2.609,-1.072)

Table 10.4: BIC values for several models for the extremely hot process $E(t)$.

10.4 Probability of a frost–free period for Medicine Hat

This section shows how the approach developed above can be used in applications. We use the developed methodology to compute two probabilities:

- π_1 : The probability of no frosts in the first week of October at the Medicine Hat site.
- π_2 : The probability of at least 5 days without frost in the first week of October at the Medicine Hat site.

The first day of October is the 275th day of the year in a leap year and the 274th day of the year in a non–leap year. We compute the probabilities for the week between 274th day and 281th day which corresponds to the first week of October in a non–leap year. We prefer this option to computing the probability for the actual first week of October, since this corresponds better to the natural cycles. Of course with a little modification one could compute the probability for the first week of October, for example by introducing a probability of 1/4 for being in a leap year.

Figure 10.17 plots the probability of a frost for each day of years since 1985. Only years with more than 355 days of data are considered. The

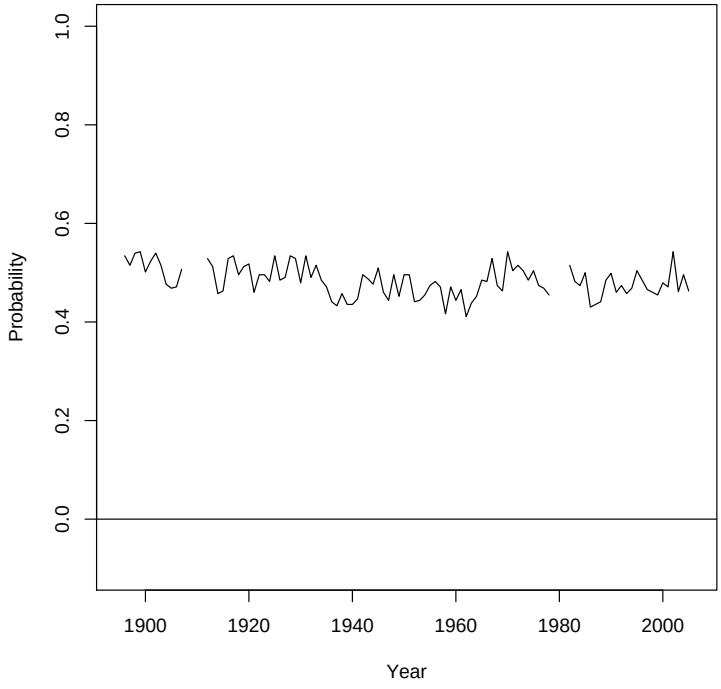

Figure 10.17: Medicine Hat's estimated mean annual probability of frost calculated from the historical data.

figure shows that the probability of a frost is fairly consistent over the years, so we assume a constant probability of frost for all years. Table 10.5 compares models with various N^k. The optimal model is $(1, N^{11})$. Table 10.6 includes two seasonal terms as well as N^k. The optimum this time $(1, N^1, COS, SIN)$, showing that in the presence of seasonal terms, the short–term past modeled by N^k is not necessary.

Model: Z_{t-1}	BIC
$(1, N^1)$	5072.2
$(1, N^2)$	4634.8
$(1, N^3)$	4465.9
$(1, N^4)$	4407.4
$(1, N^5)$	4366.0
$(1, N^6)$	4357.4
$(1, N^7)$	4356.2
$(1, N^8)$	4342.6
$(1, N^9)$	4330.5
$(1, N^{10})$	4329.1
$(1, N^{11})$	**4328.4**
$(1, N^{12})$	4332.4
$(1, N^{13})$	4330.8
$(1, N^{14})$	4345.1
$(1, N^{15})$	4362.9
$(1, N^{16})$	4385.7
$(1, N^{17})$	4407.1
$(1, N^{18})$	4420.1
$(1, N^{19})$	4440.1
$(1, N^{20})$	4463.7

Table 10.5: BIC values for models including N^k for the extremely cold process $e(t)$ at the Medicine Hat site.

Model: Z_{t-1}	BIC
$(1, N^1, COS, SIN)$	**3601.3**
$(1, N^2, COS, SIN)$	3654.8
$(1, N^3, COS, SIN)$	3693.9
$(1, N^4, COS, SIN)$	3735.2
$(1, N^5, COS, SIN)$	3763.1
$(1, N^6, COS, SIN)$	3791.0
$(1, N^7, COS, SIN)$	3813.5
$(1, N^8, COS, SIN)$	3826.2
$(1, N^9, COS, SIN)$	3834.9
$(1, N^{10}, COS, SIN)$	3843.6
$(1, N^{11}, COS, SIN)$	3849.8
$(1, N^{12}, COS, SIN)$	3855.5
$(1, N^{13}, COS, SIN)$	3857.4
$(1, N^{14}, COS, SIN)$	3862.9
$(1, N^{15}, COS, SIN)$	3868.1
$(1, N^{16}, COS, SIN)$	3873.7
$(1, N^{17}, COS, SIN)$	3877.9
$(1, N^{18}, COS, SIN)$	3878.6
$(1, N^{19}, COS, SIN)$	3880.5
$(1, N^{20}, COS, SIN)$	3882.8

Table 10.6: BIC values for several models including N^k and seasonal terms for the extremely cold process $e(t)$ at the Medicine Hat site.

Model: Z_{t-1}	BIC	parameter estimates
(1)	10122.4	(-0.0858)
$(1, e^1)$	5072.2	(-2.13, 4.18)
$(1, e^1, e^2)$	4598.2	(-2.530, 2.977, 2.00)
$(1, e^1, e^2, e^1e^2)$	4582.8	(-2.65, 3.41, 2.43, -0.855)
$(1, COS, SIN)$	3916.870	(-0.3, 4.301, 1.139)
$(1, COS, SIN, COS2, SIN2)$	3865.6	(-0.746, 4.643, 1.253 -0.550 -0.504)
$(1, e^1, COS, SIN)$	3601.3	(-1.116, 1.760, 3.332, 0.856)
$(1, e^1, COS, SIN, COS2, SIN2)$	**3566.7**	(-1.49, 1.71, 3.65, 0.96, -0.48, -0.42)
$(1, e^1, e^2, COS, SIN)$	3601.6	(-1.22, 1.66, 0.33, 3.19, 0.810)
$(1, e^1, e^2, COS, SIN, COS2, SIN2$ $, COS3, SIN3)$	3571.7	(-1.8, 1.7, 4.4, 1.3, -0.78, -0.74, 0.2, 0.4)
$(1, mt^1, COS, SIN, COS2, SIN2)$	**3356.4**	(-0.66, -0.22, 2.85, 0.73, -0.56, -0.42)

Table 10.7: BIC values for several models for the extremely cold process $e(t)$ at the Medicine Hat site.

Covariate	Theoretical sd	Experimental sd
1	0.090	0.093
e^1	0.097	0.100
COS	0.125	0.139
SIN	0.060	0.059
$COS2$	0.089	0.094
$SIN2$	0.081	0.077

Table 10.8: Theoretical and simulation estimated standard deviations for extremely cold process $e(t)$ at the Medicine Hat site.

Table 10.5 compares various models. The winner is

$$(1, mt^1, COS, SIN, COS2, SIN2).$$

However, it is not possible to compute the desired probabilities using this model since we do not know mt^1 (perhaps except at the start of the chain). Among all other models, the optimal is

$$(1, e^1, COS, SIN, COS2, SIN2),$$

which we use to compute the probabilities.

We compute the standard deviations once using simulations by generating chains from the fitted model with the above covariates, and once by computing the partial information matrix, G_N. The results are given in Table 10.8. The variance–covariance matrix calculated using partial likelihood theory is given below:

$$\begin{pmatrix} 0.0082 & -0.0043 & -0.0038 & -0.0011 & 0.0050 & 0.0030 \\ -0.0043 & 0.0094 & -0.0042 & -0.0013 & 0.0002 & 0.0003 \\ -0.0038 & -0.0042 & 0.0158 & 0.0038 & -0.0052 & -0.0037 \\ -0.0011 & -0.0013 & 0.0038 & 0.0037 & -0.0011 & -0.0017 \\ 0.0050 & 0.0002 & -0.0052 & -0.0011 & 0.0079 & 0.0015 \\ 0.0030 & 0.0003 & -0.0037 & -0.0017 & 0.0015 & 0.0066 \end{pmatrix}$$

We also find the variance–covariance matrix using simulations. To do that we generate 50 chains over time using the estimated parameters. The variance–covariance matrix using the simulations is given by:

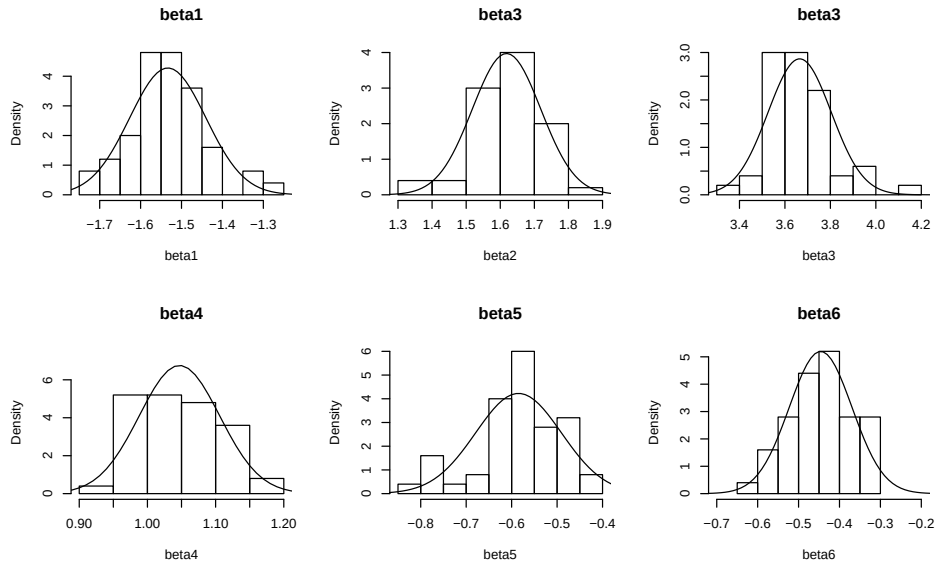

Figure 10.18: Normal curved fitted to the distribution of 50 samples of the estimated parameters.

$$\begin{pmatrix} 0.0087 & -0.0035 & -0.0054 & -0.0012 & 0.0047 & 0.0021 \\ -0.0035 & 0.0101 & -0.0058 & -0.0009 & 0.0026 & 0.0012 \\ -0.0054 & -0.0058 & 0.0194 & 0.0032 & -0.0086 & -0.0032 \\ -0.0012 & -0.0009 & 0.0032 & 0.0035 & -0.0011 & -0.0018 \\ 0.0047 & 0.0026 & -0.0086 & -0.0011 & 0.0089 & 0.0016 \\ 0.0021 & 0.0012 & -0.0032 & -0.0018 & 0.0016 & 0.0059 \end{pmatrix}$$

We see that the simulated variance–covariance matrix has close values to the partial likelihood, all entries having the same sign. We also look at the distribution of the estimators using the 50 samples. Figure 10.18 shows the parameter estimates approximately follow a normal distribution.

To estimate the desired probabilities, we generate samples (10000) from the parameter space using the mean of the parameters and variance–covariance matrix from a multivariate normal. To fix ideas suppose we want to compute the probability of no frost between (and including) the 274th day and the 280th day of the year. For every vector of parameters, we then compute the probability of observing $(0, 0, 0, 0, 0, 0, 0)$ exactly once given it was below zero on the 273th day and once it was above zero. In other words we

compute

$$P(e(274) = 1, \cdots, e(281) = 1 | e(273) = 1),$$

and

$$P(e(274) = 1, \cdots, e(281) = 1 | e(273) = 0).$$

We also use the historical data to estimate $p_0 = P(e(273) = 1)$. Then the desired probability would be

$$\begin{aligned} P(e(274) = 1, \cdots, e(281) = 1) &= \\ p_0 P(e(274) = 1, \cdots, e(281) = 1 | e(273) = 1) &+ \\ (1 - p_0) P(e(274) = 1, \cdots, e(281) = 1 | e(273) = 0) & \end{aligned}$$

Then in order to get a 95% confidence intervals we use $(q(0.025), q(1 - 0.025))$, where q is the (left) quantile function of the vector of the probabilities.

Using the historical data, we obtain $p_0 = P(e(274) = 1) = 0.2432432$. Then for every parameter generated from the multivariate normal with mean and the above variance–covariance matrix we can estimate the two probabilities π_1 and π_2. We sample 10000 times from the multivariate normal, compute 10000 probabilities and take the 0.025th and 0.975th (left) quantiles to get the following confidence intervals for π_1 and π_2 respectively:

$$(0.28, 0.40),$$

and

$$(0.74, 0.85).$$

If we use the simulated variance–covariance matrix, we'll get the following confidence intervals for π_1 and π_2

$$(0.28, 0.40),$$

and

$$(0.75, 0.85),$$

which are very similar to the aforementioned intervals.

10.5 Possible applications of the models

To understand the potential applications of these models and results I contacted Dr. Nathaniel Newlands from AAFC (Agriculture and Agi-food Canada). He give the following insightful comments.

"Forecasted (probability of precipitation) is a leading indicator used by crop insurance companies. Probabilities of this kind (agroclimate) are typically most useful in early growing season by farmers in deciding planting dates and deciding on irrigation scheduling and ordering fertilizer and other kinds of inputs. Frost probability in latter growing season is critically important in deciding when to harvest crops before they have a higher potential for weather damage. So, essentially at the start and end of growing season, frost, precipitation (sometimes as a water stress index) and temp extremes are all informative for farmers and other decision makers in ag industry.

I would generally say that a broader set of probabilities like these are of special interest to the government side as they look for improving and/or developing new models, web portals and other tools to aid a wide array of the decision makers in the agricultural industry with their business decisions. Farmers (depending on what region of Canada they are in) are used to dealing with reoccurring weather and now climate change events, so often their viewpoint and decision needs are far more regionally specific than government which tries to balance regional with national needs and levels of risk to changing agroclimate.

The crop insurance industry is probably the most specific user of such information. For example, they base their insurance quotes for the event of precipitation on some specific times of the year."

Chapter 11

Conclusions and future research

11.1 Introduction

This chapter summarizes the work and draws conclusions from the the statistical analysis and the theory developed in the previous chapters. We also point out a few topics for future research as a continuation of the work done in this thesis.

11.2 Summary

This thesis has presented statistical techniques we have developed to model precipitation and temperature over time. The dataset we use is the historical weather data published by Environment Canada [10]. A Python code was provided to extract the data from the binary format and the Python module is available in [23]. [See the appendices for more information regarding the dataset, the Python module and other resources.] Then we performed an exploratory analysis of the data. See the conclusions section of Chapter 2 for details. In order to model the 0-1 precipitation process over time, rth-order Markov chains are a natural choice. We found a representation theorem for such chains using the conditional probabilities and used it to pick appropriate models for precipitation and dichotomized temperatures in the next chapters. In order to dichotomize a continuous process (temperature) one can use quantiles as thresholds. The climate data are often very large in size and hence computing quantiles is not possible due to memory or space limitations. We propose an algorithm that uses smaller partitions of the data in order to approximate the quantiles and provides a measure of goodness of such approximations. Thinking about the quantiles led us to an extension of the traditional definition of "quantiles" to the "left–" and "right–quantiles" and we showed by various theorems that this definition is more intuitively appealing and practically useful. For example a

symmetric relation holds with the new definition which we used in various applications. In order to assess the goodness of approximating quantiles, we introduced the "probability loss function", which we showed is invariant under monotonic transformations. We used this loss function in various applications such as picking optimal probability index vectors to summarize data vectors or assigning quantiles to a random sample in order to make a quantile-quantile plot. Then we used this loss function, to define a distance between random variables and showed that this distance is also invariant under monotonic transformations. We also pointed out how the probability loss function and the distance defined by it could be used to estimate parameters of a distribution. Chapter 10 uses the above methods to find appropriate models for extremely high and low temperatures. For example, we show how these models can be used to build confidence intervals for the probability of a frost-free period.

11.3 Future research

In this section, we suggest a few lines of research that are continuations of this thesis work.

11.3.1 rth-order Markov chains

Chapter 3 developed a consistency theorem for the conditional probabilities of a discrete–time categorical stochastic process and a representation theorem for rth-order Markov chains. We expressed the conditional probabilities of such chains as a linear combination of monomials of past times and used partial likelihood to estimate the parameters in the binary case. We propose the following extensions to this work:

- Find a similar consistency theorem for general (not only categorical) discrete–time categorical processes and a representation theorem for rth-order Markov chains.
- We used partial likelihood only to estimate the parameters in the binary case; an extension is needed to chains with larger number of states.
- We pointed out in Chapter 3 that we can add other covariates to the linear terms to get non-stationary chains. We can also add spatial components to build spatial-temporal models. However, estimating

the parameters in this case needs an extension of the theory due to the possible dependence over space.

- A Bayesian method can be deployed to estimate the parameters of these models.

11.3.2 Approximating quantiles and data summaries

We provided a general framework for summarizing data, combining summaries and making inference about the original data. We propose the following research topics:

- Suppose a data vector x is given which is partitioned to $x^1, \cdots, x^m$ of lengths $n_1, \cdots, n_m$. We are allowed to read the partitions separately and save $k_1, \cdots, k_m$ data points from these partitions.

  1. What information regarding $x^1, \cdots, x^m$ (of length $k_1, \cdots, k_m$) should be saved to optimally approximate $lq_x(p)$ for a fixed p?
  2. What information regarding $x^1, \cdots, x^m$ (of length $k_1, \cdots, k_m$) should be saved to optimally approximate $lq_x(p)$ for all $p \in E \subset [0,1]$?
  3. Suppose pre-defined summaries of $x^1, \cdots, x^m$ are given which are not necessarily optimal. How can we optimally infer about $lq_x(p)$ or $lq_x(p)$ for all $p \in E \subset [0,1]$?
  4. Suppose a fixed memory space is given. Find an optimal (fastest) algorithm which gives approximations of accuracy ϵ (in the probability loss sense).

- Suppose a random sample $X_1, \cdots, X_n$ is given. We can build distribution-free confidence intervals for quantiles of the underlying distribution. (See [15].) Now suppose we have created a summary of this random sample in a certain way. Build confidence intervals based on these summaries.

11.3.3 Parameter estimation using probability loss and quantile distances

Chapter 9 developed a framework to estimate parameters of distributions. We also introduced the quantile distances in order to measure the distance between random variables and showed its invariance under monotonic transformations. We propose the following extensions:

- Given a random sample $X_1, \cdots, X_n$ what is the best estimate of $lq_X(p)$ using the probability loss function. What are the properties of that estimator? Is it consistent?
- What are the suprema of $LQD_{\delta_X}(X,Y)$ and $LQD_{\delta_X+\delta_Y}(X,Y)$ over the space of all random variables?
- What is the relation between $LQD_{\delta_X}(X,Y)$ and $LQD_{\delta_Y}(X,Y)$?
- Do $LQD_1(X,Y) = LQD_{\delta_X}(X,Y)$ or $LQD(X,Y) = LQD_{\delta_X+\delta_Y}(X,Y)$ satisfy the triangle inequality?
- Chapter 9 was a theoretical chapter. A lot of simulation studies and analysis of real data is needed to support the theory and get new ideas.

Bibliography

[1] R. Agrawal and A. Swami. A one-pass space-efficient algorithm for finding quantiles. In *in Proc. 7th Intl. Conf. Management of Data (COMAD-95*, 1995.

[2] H. Akiake. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, pages 716–723, 1974.

[3] K. Alsabti, S. Ranka, and V. Singh. A one-pass algorithm for accurately estimating quantiles for disk-resident data. In *VLDB '97: Proceedings of the 23rd International Conference on Very Large Data Bases*, pages 346–355, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[4] T. W. Anderson and L. A. Goodman. Statistical inference about markov chains. *Ann. Math. Statist.*, pages 89–110, 1957.

[5] M. S. Bartlett. The frequency goodness of fit test for probability chains. *Proc. Cambridge Philos. Soc.*, pages 86–95, 1951.

[6] J. Besag. Spatial interactions and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society series B*, pages 192–225, 1974.

[7] P. Billingsley. *Probability and measure*. John Wiley and Sons, 1985.

[8] R. W. Blum and J. W. John. Time bounds for selection. *J. Comput. Sys. Sci.*, 7:448–461, 1973.

[9] L. Breiman. *Probability*. SIAM, 1992.

[10] Environment Canada. The climate cds. *http://www.weatheroffice.ec.gc.ca*, 2007.

[11] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury, 2001.

[12] E. H. Chin. Modeling daily precipitation process with markov chain. *Water resources research*, (6):949–956, 1977.

[13] W. K. Ching, E. S. Fung, and K. M. NG. Higher-order markov chain models for categorical data sequences. *Naval Research Logistics*, pages 557–574, 2004.

[14] N. Cressie and L. Subash. New models for markov random fields. *Journal of applied probability*, pages 877–884, 1992.

[15] H. A. David and H. N Nagaraja. *Order Statistics (3rd Edition)*. Wiley, 2003.

[16] P. Embrechts, C. Klppelberg, and T. Mikosch. *Modelling extremal events for insurance and finance*. Springer, 2001.

[17] J. E. Freund and B. M. Perles. A new look at quartiles of ungrouped data. *The American statistician*, pages 200–203, 1987.

[18] K. R. Gabriel and J. Neumann. A markov chain model for daily rainfall occurance at tel aviv. *Quart. J. Roy. Met. Soc.*, pages 90–95, 1962.

[19] M. Greenwald and S. Khanna. Space-efficient online computation of quantile summaries. In *In SIGMOD*, pages 58–66, 2001.

[20] E. J. Hannan. The estimation of the order of an arma process. *Ann. Statist.*, pages 1071–1081, 1980.

[21] L. Hao and D. Q. Naiman. *Quantile Regression*. Quantitative Applications in the Social Sciences Series. SAGE publications, 2007.

[22] D. M. A. Haughton. On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, (1):342–355, 1988.

[23] R. Hosseini. Python module for canadian climate data. *http://bayes.stat.ubc.ca/∼reza/python*, 2009.

[24] R. Hyndman and Y. Fan. Sample quantiles in statistical packages. *The American Statistician*, 1996.

[25] R. J. Hyndman and Y. Fan. Sample quantiles in statistical packages. *The American Statistician*, pages 361–365, 1996.

[26] R. Jain and I. Chlamtac. The p2 algorithm for dynamic calculation of quantiles and histograms without storing observations. *Commun. ACM*, 28(10):1076–1085, 1985.

[27] B. Kedem and K. Fokianos. *Regression Models for Time Series Analysis*. Wiley Series in Probability and Statistics, 2002.

[28] D. E. Knuth. *Sorting and Searching*, volume 3. Addison-Wesley, 1973.

[29] R. Koenker. *Quantile Regression*. Cambridge university press, 2005.

[30] E. L. Lehmann and G. Casella. *The theory of point estimation*. Springer-Verlag, 1998.

[31] L. P. Llorente. *Statistical Inference Based on Divergence Measures*. CRC Press, 2006.

[32] G. S. Manku, S. Rajagopalan, and B. G. Lindsay. Approximate medians and other quantiles in one pass and with limited memory. pages 426–435, 1998.

[33] G. S. Manku, S. Rajagopalan, and B. G. Lindsay. Random sampling techniques for space efficient online computation of order statistics of large datasets. In *In SIGMOD*, pages 251–262, 1999.

[34] E. Mekis and W. D. Hogg. Rehabilitation and analysis of canadian daily precipitation time series. *Atmosphere Ocean*, pages 53–85, 1999.

[35] S. E. Moon, S. Ryo, and J. Kwon. International journal of climatalogy. pages 1009–116, 1993.

[36] J. I. Munro and M. S. Paterson. Selection and sorting with limited storage. *Theoretical computer science*, 12:253–258, 1980.

[37] B. Öksendal. *Stochastic Differential Equations: An Introduction with Applications.* Springer, 2003.

[38] E. Parzen. Nonparametric statistical data modeling. *Journal of the American Statistical Association*, 74:105–121, 1979.

[39] M. Paterson. Progress in selection. pages 368–379, 1997.

[40] A. E. Raftery. A model for higher order markov chains. *J. R. Statist. B.*, (3):528–539, 1985.

[41] T. Rychlik. *Projecting statistical functionals*. Springer, 2001.

[42] G. Schwartz. Estimating the dimension of a model. *Ann. Statist.*, pages 461–464, 1978.

[43] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, 1980.

[44] R. Shibita. Selection of the order of an autoregressive modelby akiake's information criterion. *Biometrika*, pages 117–126, 1976.

[45] H. Tong. Determination of the order of a markov chain by akiake's information criterion. *J. Appl. Prob.*, pages 488–497, 1975.

[46] H. Tong and P. Gates. On markov chain modelling to some weather data. *Journal of applied meteorology*, pages 1145–1151, 1976.

[47] L.A. Vincent, X. Zhang, B. R. Bonsal, and Hogg W.D. Homogenization of daily temperatures over canada. *Journal of Climate*, pages 1322–1334, 2002.

[48] W. Wong. Theory of partial likelihood. *The Annals of Statistics*, (1):88–123, 1986.

[49] F. F. Yao. On lower bounds for selection problems. Technical report, Cambridge, MA, USA, 1974.

Appendix A

Climate review

A.1 Organizations and resources

- WMO: The World Meteorological Organization (WMO) is a specialized agency of the United Nations. It is the UN system's authoritative voice on the state and behavior of the Earth's atmosphere, its interaction with the oceans, the climate it produces and the resulting distribution of water resources.

- Environment Canada: Environment Canada's mandate is to preserve and enhance the quality of the natural environment; conserve Canada's renewable resources; conserve and protect Canada's water resources; forecast weather and environmental change; enforce rules relating to boundary waters; and coordinate environmental policies and programs for the federal government.

- The Meteorological Service of Canada: The Meteorological Service of Canada is Canada's source for meteorological information. The Service monitors water quantities, provides information and conducts research on climate, atmospheric science, air quality, ice and other environmental issues, making it an important source of expertise in these areas.

- Natural Resources Canada

- Agriculture and Agri–Food Canada: Agriculture and Agri-Food Canada (AAFC) provides information, research and technology, and policies and programs to achieve security of the food system, health of the environment and innovation for growth. AAFC, along with its portfolio partners, reports to Parliament and Canadians through the Minister of Agriculture and Agri-Food and Minister for the Canadian Wheat Board.

- Alberta Agriculture Food and Rural Development

- Statistics Canada
- AMS: The American Meteorological Society promotes the development and dissemination of information and education on the atmospheric and related oceanic and hydrologic sciences and the advancement of their professional applications. Founded in 1919, AMS has a membership of more than 11,000 professionals, professors, students, and weather enthusiasts. AMS publishes nine atmospheric and related oceanic and hydrologic journals (in print and online) sponsors more than 12 conferences annually, and offers numerous programs and services.
- GeoBase is a federal, provincial and territorial government initiative that is overseen by the Canadian Council on Geomatics (CCOG). It is undertaken to ensure the provision of, and access to, a common, up-to-date and maintained base of quality geospatial data for all of Canada. Through the GeoBase portal, users with an interest in the field of geomatics have access to quality geospatial information at no cost and with unrestricted use.

A.2 Definitions and climate variables

- Atmosphere: Gaseous envelope which surrounds the Earth. Definition source: International Meteorological Vocabulary, WMO - No. 182
- Troposphere: Lower part of the terrestrial atmosphere, extending from the surface up to a height varying from about 9 km at the poles to about 17 km at the equator, in which the temperature decreases fairly uniformly with height. Definition source: International Meteorological Vocabulary, WMO - No. 182
- Meteorology: Study of the atmosphere and its phenomena. Definition Source: International Meteorological Vocabulary, WMO - No. 182
- Climatology: Study of the mean physical state of the atmosphere together with its statistical variations in both space and time as reflected in the weather behavior over a period of many years. Definition Source: International Meteorological Vocabulary, WMO - No. 182
- Hydrology: (1) Science that deals with the waters above and below the land surfaces of the Earth, their occurrence, circulation and distribution, both in time and space, their biological, chemical and phys-

ical properties, their reaction with their environment, including their relation to living beings. (2) Science that deals with the processes governing the depletion and replenishment of the water resources of the land areas, and treats the various phases of the hydrological cycle. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Basic topography: General geometrical configuration of the distribution of geopotential height on an isobaric surface or on a thickness chart, or of atmospheric pressure on a constant–height chart (e.g., mean sea–level surface chart). Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Weather: State of the atmosphere at a particular time, as defined by the various meteorological elements. Term Source: International Meteorological Vocabulary, WMO - No. 182

- Climate: Synthesis of weather conditions in a given area, characterized by long–term statistics (mean values, variances, probabilities of extreme values, etc.) of the meteorological elements in that area. Definition source: International Meteorological Vocabulary, WMO - No. 182

- Paleoclimate: Climate of a prehistoric period whose main characteristics may be inferred, for example, from geological and paleobiological (fossil) evidence. Definition source: International Meteorological Vocabulary, WMO - No. 182

- Climate change: (1) In the most general sense, the term "climate change" encompasses all forms of climatic inconstancy (i.e., any differences between long–term statistics of the meteorological elements calculated for different periods but relating to the same area) regardless of their statistical nature or physical causes. Climate changes may result from such factors as changes in solar emission, long–period changes in the Earth's orbital elements (eccentricity, obliquity of the ecliptic, precession of the equinoxes), natural internal processes of the climate system, or anthropogenic forcing (e.g. increasing atmospheric concentrations of carbon dioxide and other greenhouse gases). (2) The term "climate change" is often used in a more restricted sense, to denote a significant change (i.e., a change having important economic, environmental and social effects) in the mean values of a meteorological element (in particular temperature or amount of precipitation)

in the course of a certain period of time, where the means are taken over periods of the order of a decade or longer. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Climate model: Representation of the climate system based on the mathematical equation governing the behavior of the various components of the system and including treatments of key physical processes and interactions,cast in a form suitable for numerical approximation (generally now making use of electronic computers). Definition source: International Meteorological Vocabulary, WMO - No. 182

- Precipitation: Hydrometeor consisting of a fall of an ensemble of particles. The forms of precipitation are: rain, drizzle, snow, snow grains, snow pellets, diamond dust, hail and ice pellets. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Rainfall: Amount of precipitation which is measured by means of a rain gauge. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Atmospheric pressure: Pressure (force per unit area) exerted by the atmosphere on any surface by virtue of its weight; it is equivalent to the weight of a vertical column of air extending above a surface of unit area to the outer limit of the atmosphere. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Humidity: Water vapor content of the air. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Climatic season: A long spell of weather which characterizes part of the year and which occurs with some approach to regularity, especially in low latitudes. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Growing season: Season during which meteorological conditions are favorable to the growth of plants. Definition Source: International Meteorological Vocabulary, WMO-No.182

- Dry season: Period of the year characterized by the (almost) complete absence of rainfall. The term is mainly used for low latitude regions. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Rainy season: In the lower latitudes, an annually recurring period of high rainfalls preceded and followed by relatively dry periods. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Flood: (1) The overflowing by water of the normal confines of a stream or other body of water, or the accumulation of water by drainage over areas which are not normally submerged. (2) Controlled spreading of water over a particular region. Definition Source: International Meteorological Vocabulary, WMO - No. 182 Term Note

- Drought: (1) Prolonged absence or marked deficiency of precipitation. (2) Period of abnormally dry weather sufficiently prolonged for the lack of precipitation to cause a serious hydrological imbalance. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Drought index: An index which is related to some of the cumulative effects of a prolonged and abnormal moisture deficiency. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Climate system: System consisting of the atmosphere, the hydrosphere (comprising the liquid water distributed on and beneath the Earth's surface, as well as the cryosphere, i.e. the snow and ice on and beneath the surface), the surface lithosphere (comprising the rock, soil and sediment of the Earth's surface), and the biosphere (comprising Earth's plant and animal life and man), which, under the effects of the solar radiation received by the Earth, determines the climate of the Earth. Although climate essentially relates to the varying states of the atmosphere only, the other parts of the climate system also have a significant role in forming climate, through their interactions with the atmosphere. Definition Source: International Meteorological Vocabulary, WMO-No.182

- Wind: Air motion relative to the Earth's surface. Unless otherwise specified, only the horizontal component is considered. Definition Source: International Meteorological Vocabulary, WMO-No.182

- Humidity: Definition Water vapor content of the air. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Statistical model: (1) Mathematical model which has been derived from the statistical analysis of relevant meteorological variables. (2)

Numerical model, usually of the general circulation, which predicts certain statistical properties of the atmosphere rather than the full three-dimensional, time-dependent, distribution of each variable. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Statistical forecast: Definition Objective forecast based on a statistical examination of the past behavior of the atmosphere, using regression formulae, probabilities, etc. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Probability forecast: Definition Objective forecast based on a statistical examination of the past behavior of the atmosphere, using regression formulae, probabilities, etc. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Circulation model: Simplified representation of atmospheric flow used to study its principal characteristics. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- El Niño: An anomalous warming of ocean water off the west coast of South America, usually accompanied by heavy rainfall in the coastal region of Peru and Chile. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Hurricane: (1) Name given to a warm core tropical cyclone with maximum surface wind of 118 km h-1 (64 knots, 74 mph) or greater (hurricane force wind) in the North Atlantic, the Caribbean and the Gulf of Mexico, and in the Eastern North Pacific Ocean. (2) A tropical cyclone with hurricane force winds in the South Pacific and South-East Indian Ocean. Definition Source: International Meteorological Vocabulary, WMO - No. 182

- Green house effect: Warming of the lower layers of the atmosphere due to its different absorption properties for long- and short-wave radiation. Definition Source: International Meteorological Vocabulary, WMO - No. 182

A.3 Climatology

A.3.1 General circulations

Forces that cause variety of land forms on the Earth can be categorized into two types:

- Inside forces: Volcanoes, earth quakes and etc.
- Outside forces: Forces that are conveyed by atmosphere to the Earth's surface. Sun is the most important factor in causing such forces in different forms.

Although, the first type is of great importance and is not totally independent of the second type, here we only focus on the second type.

Weather is defined to be day-to-day variations to the state of atmosphere. In order to understand the weather, we need to understand how such forces interact and the factors that cause such variations.

The climate system is composed of three parts:

- a radiative energy flow system
- a circulation system
- water cycle

We will explain these in the following.

The Sun is the most important source of energy driving the climate system. The atmosphere reflects about 31 percent of the energy to the space. It also absorbs (ozone, water vapor and carbon dioxide) 23 percent of the energy from Sun before it reaches the Earth's surface. Finally, the Earth's surface absorbs about 46 percent. The Earth's surface radiates back some of this energy with longer wavelengthes which in turn is absorbed by the atmosphere. In fact atmosphere is able to absorb long wavelengthes better. The presence of greenhouse gases (ozone, water vapor and carbon dioxide) in the atmosphere can cause the greenhouse effect by absorbing more energy from the long wavelengthes of energy. Also some of the heat from the earth goes back to the atmosphere indirectly by the evaporated water.

Near the Equator the solar radiation reaches the Earth's surface with a steeper angle and shorter path through the atmosphere compared to the poles. This explains why it is warmer at the Equator than at the poles.

Atmospheric circulation are created as a natural response to the difference of temperature between the Equator and the poles. However, other factors also have an effect: the Earth's rotation, the force of gravity, the temperature of the ocean and land, and the presence of topographical features such as mountains, plants ice and so on.

A.3.2 Topography of Canada

A listing of main features comprises the Western Cordillera, the Prairies, the Great Lakes, the Canadian Shield, the Gulf of St. Lawrence and the Arctic Islands. We only review the Prairies which are the most suitable lands for farming.

The Prairies extend eastward from the Rocky Mountains sloping down towards the great Canadian Shield. The elevations range from 1500 m in the west to about 250 m in Manitoba. The slope however is not even but is broken by steps, the Manitoba Escarpment and the Missouri Coteau. Minor hill rows tend to run parallel to these; the Cypress Hills however are an exception. A chain of large lakes in Manitoba marks the extent of a giant inland lake during glacial times. The rivers run from the Rockies toward the northeast, some into the Arctic Ocean, others into Hudson Bay. They are often cut deeply into the flat or slightly rolling, generally featureless plain.

A.4 Some interesting facts about Canadian geography and weather

- Total Area of Canada:
  The total area of Canada is 9,984,670 square kilometers. Of this, 9 093,507 square kilometers is land and 891,163 square kilometers is fresh water. Canada's area is the second largest in the world (after Russia which has a total area of 17,075,000 square kilometers). On Canadian territory, the longest distance North to South (on land) is 4,634 kilometers from Cape Columbia on Ellesmere Island, Nunavut to Middle Island in Lake Erie, Ontario. The longest distance East to West is 5,514 kilometers from Cape Spear, Newfoundland and Labrador, to the Yukon Territory–Alaska boundary.

- Boundary:
  The total length of the Canada–United States boundary is 8890 kilometers.

- Landmass and Freshwater:
  Approximately 40% of Canada's landmass and freshwater is north of 60 degrees North latitude. Between them, the Northwest Territories and Nunavut contains 9.2% of the world's total freshwater. The area of Canada north of the treeline is 2,728,800 square kilometers or 27.4% of the total area of the country.

- The Great Lakes:
  The Great Lakes (Superior, Michigan, Huron, Erie and Ontario) are the largest group of freshwater lakes in the world. They have a total surface area of 245,000 square kilometers, of which about one third is in Canada. Lake Michigan is entirely within the USA.

- Coastline:
  Canada has the world's longest coastline: 202 080 kilometers.

- Hailstorm:
  At the time it happened, the most expensive natural catastrophe in terms of property damage was a violent hailstorm that struck Calgary (photo of Calgary) on September 7th, 1991. Insurance companies paid about $400 million to repair over 65,000 cars, 60,000 homes and businesses, and a number of aircraft.

- Tornado:
  The Regina Tornado of June 30th, 1912, rated as F4 (winds of 330 to 416 kilometres per hour) was the most severe tornado so far known in Canada. It killed 28 people, injured hundreds and demolished much of the downtown area.

- Most Severe Flood:
  The most severe flood in Canadian history occurred on October 14th to 15th, 1954 when Hurricane Hazel brought 214 millimeters of rain in Toronto region in just 72 hours.

- Manitoulin Island:
  The world's largest island in a freshwater lake is Manitoulin Island in Lake Huron, 2765 square kilometers.

- Mount Logan:
  The highest mountain in Canada is Mount Logan, Yukon Territory, 5959 meters.

- Medicine Hat:
  Medicine Hat is the driest city with 271 days without measurable precipitation. [Source: Phillips, D. 1990. The Climate of Canada. Catalogue No. En56-1/1990E. Ottawa: Minister of Supply and Services of Canada.]

Appendix B

Extracting Canadian Climate Data from Environment Canada dataset

B.1 Introduction

In this document, some instructions are given to use the climate data provided by environment Canada [10]. The data we are using are contained in a file, which can be downloaded from the environment Canada website:

http://www.weatheroffice.ec.gc.ca.

"The National Climate Data and Information Archive, operated and maintained by Environment Canada, contains official climate and weather observations for Canada" (quoting from the website).

Environment Canada has published a series of climate data CDs: 1993, 1996, 2002, 2007. The newest version is the 2007 CD. The Environment Canada website also includes some other useful information, as a glossary of some useful terms in climate literature and also some information about the files. In particular, the glossary includes the definition of precipitation:

Precipitation: The sum of the total rainfall and the water equivalent of the total snowfall observed during the day.

On the 2007 CD, data are stored in a binary format in several files. The CD includes two softwares to use the data, "cdcd" and "cdex" along with manuals to use the softwares. "cdcd" is to view the data and "cdex" is to extract the data. "cdex" can only extract the data for one climate station at a time in certain formats which are not necessarily convenient to use in R (a well known statistical software) or other statistical softwares. In these formats the longitude, latitude and elevation are missing. Hence, to get the data in our desired way, we need to read the binary files using another

program. Bernhard Reiter has written a code using Python to get the data, which is available online at

http://www.intevation.de/∼bernhard/archiv/uwm/canadian_climate_cdformat/.

However, this code fails to get the data for a large proportion of the stations. We have modified the code to get the data for all stations. The modified code [23] is available at

http://bayes.stat.ubc.ca/∼reza/python.

After getting the data, we need to write the data in our desired formats. We have also included many new functions in Python for different extraction purposes.

There are 7802 stations from all over Canada. The available variables are:

1. maximum temperature
2. minimum temperature
3. one–day rainfall
4. one–day snowfall
5. one–day precipitation
6. snow depth on the ground

These data are available, both daily and monthly. For each station the data are available for different intervals of time.

The data are saved in 8 directories on the CD labeled $1, 2, \cdots, 8$. They correspond to different territories of Canada.

```
1 --> British Columbia

2 --> Yukon territories, Nunavut and North west territories

3 --> Alberta

4 --> Saskatchewan
```

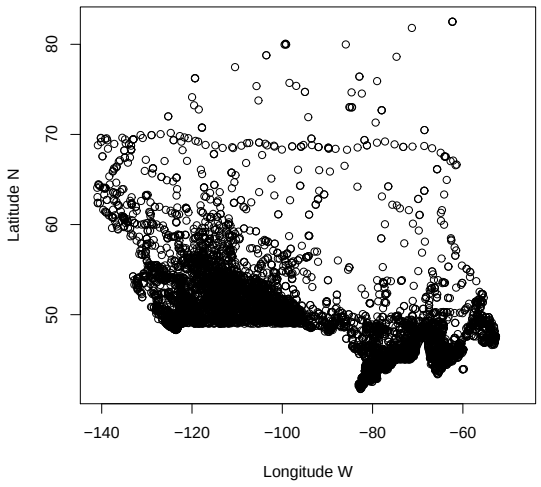

Figure B.1: Canada site locations

`5 --> Manitoba`

`6 --> Ontario`

`7 --> Quebec`

`8 --> Nova Scotia, New found land and Labrador`

Each directory contains a number of data files and index files. For example, directory 3 which correspond to Alberta contains the following files:

$$DATA.301, DATA.302, \cdots, DATA.308$$

and

$$INDEX.301, INDEX.302, \cdots, INDEX.308.$$

Each DATA file corresponds to the data of a region in Alberta and the corresponding INDEX file contains the information about the stations in the given region. In Figure B.1, you can see the location of available stations over Canada.

B.2 Using Python to extract data

In the following, we illustrate getting the data using the python module:

"Reza_canadian_data.py"

After opening the python interface, let us import some necessary packages and tell python where the data are stored. Using sys.path.append specify the directory where Reza_canadian_data.py is stored as shown below. Also, define Topdirectory to be where the data are stored.

```
>>>import sys
>>>sys.path.append("D:\School\Research\Climate\Python_code")
>>>Topdirectory="D:\Data"
>>>from Reza_canadian_data import *
>>>stations=get_station_list(Topdirectory)
```

Once you did that you can call the command get_station_list from Reza_canadian_data to get the list of the stations available on the CD. Let us see how many stations we have access to:

```
>>> len(stations)
7802
```

Let us pick a random station, say the 3000–th station and find out its id and index.

```
>>> s=stations[2436]
>>> s.stationnumber
'3025480'
>>> s.index_record
('5480', 'RED DEER A            ', 'YQF', 5211, 11354, 905, 1938,
1938, 1938, 1938, 1938, 1938, 1955, 2007, 2007, 2007, 2007, 2007,
2007, 2007, 9904)
>>> len(s.index_record)
21
```

The command "stationnumber" gives back the id of the given station on the CD. The stations in the same district start with the same numbers. For example the stations in Alberta all start with 30 and so Red Deer is in Alberta. You can use cdcd to see the list of the stations and id numbers to figure out which ids correspond to which districts.

The index_record command reads the information available for the given station. There are many values available and it is hard to understand what they mean. As you see the index has 21 components. Here is the explanation of each component:

1. The last four digits of the id
2. station name
3. Airport is the three–character airport identifier that some stations have (e.g., "YWG" for Winnipeg); if none exists for this station then the field is left blank
4. latitude
5. longitude
6. elevation
7. The first available year for max temperature
8. The first available year for min temperature
9. The first available year for mean temperature
10. The first available year for rainfall
11. The first available year for snowfall
12. The first available year for snow depth
13. The first available year for precipitation
14. The last available year for Max temperature
15. The last available year for min temperature
16. The last available year for mean temperature
17. The last available year for rainfall
18. The last available year for snowfall
19. The last available year for precipitation
20. The last available year for snow depth

21. Starting Record Number: This record is a header that contains information about the station

Hence, for example this station name is Red Deer. It has the data for precipitation from 1938 to 2007. Whenever, 9999 is recorded as the first and 55537 as the last available year for a variable, that variable is missing. As mentioned before the available data for a given station are maximum temperature, minimum temperature, one–day rainfall, one–day snowfall, one–day precipitation and snow depth. These are coded in Reza_canadian_data.py as

```
"MT" "mint" "rain" "snow" "precip" "snow_ground"
```

We have used the following procedure in python interface to create a file "stations.txt", which has the information for all the available stations. In every row the information for a stations is given. There are 22 columns, the first one is the stations id and the other 21 are as described above. Whenever, the station was not an airport station, the airport identifier was recorded as NA. Notice, how using the "if" command in below, we have separated the case where the airport identifier is blank from the case that there is an airport identifier.

```
stations=get_station_list(Topdirectory)

f=open('stations.txt','w') for s in stations:
    ind=s.index_record
    if ind[2]=='   ':
      f.write(str(s.stationid)+','+str(ind[0])+','+str(ind[1])
      +','+'NA'+','+str(ind[3])+','+str(ind[4])+','+str(ind[5])
      +','+str(ind[6])+','+str(ind[7])+','+str(ind[8])
      +','+str(ind[9])+','+str(ind[10])+','+str(ind[11])
      +','+str(ind[12])+','+str(ind[13])+','+str(ind[14])
      +','+str(ind[15])+','+str(ind[16])+','+str(ind[17])
      +','+str(ind[18])+','+str(ind[19])+','+str(ind[20])
      +'\n')
    else:
      f.write(str(s.stationid)+','+str(ind[0])+','+str(ind[1])
      +','+str(ind[2])+','+str(ind[3])+','+str(ind[4])
      +','+str(ind[5])+','+str(ind[6])+','+str(ind[7])
```

```
      +','+str(ind[8])+','+str(ind[9])+','+str(ind[10])
      +','+str(ind[11])+','+str(ind[12])+','+str(ind[13])
      +','+str(ind[14])+','+str(ind[15])+','+str(ind[16])
      +','+str(ind[17])+','+str(ind[18])
      +','+str(ind[19])+','+str(ind[20])+'\n')

f.close()
```

One of the useful commands in Reza_canadian_data.py is get_data. Let us use this command to get some data.

```
data=s.get_data(1995,``precip")
>>> len(data)
3
>>> len(data[0])
366
>>> len(data[1])
366
>>> len(data[2])
108
```

As you see the data object created has three components. The first two components each have 366 entries and the third one has 108 components. The first component of the data is the data values for each day of the year, the amount of precipitation. The second component includes the flag associated with each daily values. The third component correspond to monthly values, number of missing days for a given month and etc. Let us look at the first two components. We print the value of precipitation for the first 60 days of the year:

```
>>> for precip in data[0][0:60]:
        print "%5.1f" % precip,
  0.0   0.2   0.0   0.0   0.0   0.5   0.0   0.0   0.0   2.0   0.8
  0.0   0.2   0.4   2.2   0.4   0.6   0.0   0.0   0.0   0.0   0.0
  0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
  0.0   0.0   0.0   0.0   0.2   0.0   0.4   0.0   0.0   0.0   0.0
  0.0   0.2   1.2   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
  0.0   3.0   0.0   0.0 -999.9
```

Everything looks OK other than the last value which is -999.9. Every missing value in the dataset is shown by -999.9. In fact to see the status of a data point look the corresponding flag which is given in the second component of the data. Let us look at the flag for the first 60 days of the year as well:

```
>>> for flag in data[1][0:60]:
        print "%5.1f" % flag,


  0.0   0.0   0.0   0.0   2.0   0.0   2.0   2.0   2.0   0.0   0.0
  2.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
  0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   2.0
  2.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   2.0   2.0   2.0
  0.0   0.0   0.0   2.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
  2.0   0.0   2.0   0.0  14.0
```

We need to know what each flag means. Note that the flag corresponding to -999.9 is 14. A description of the flags is given below:

```
0 ->  Observed value
1 ->  Estimated
2 ->  Trace. Value is reported 0
3 ->  Precipitation occurred, amount uncertain; value is 0
4 ->  Precipitation may or may not have occurred; value is 0
5 ->  Accumulated amount (from past days possibly)
6 ->  Accumulated and estimated
7 to 13 -> unused
14 -> This is used to denote Feb 29 in a non-leap year
15 -> Missing data
```

In summary only a data point with 0 flag (no flag) is valid. The flag 2.0 corresponds to "trace" (as called by Environment Canada) which is a precipitation under 0.2 (mm) that can not be measured accurately and the value is reported as zero. In the above example the flag corresponding to -999.9 is 14 which is to denote Feb 29 in a non–leap year as explained above and this makes sense since the 60th day of the year corresponds to Feb 29th. There are 13 points flagged 2.0 and so we have many "trace" values. For more information regrading the flags and data format refer to 'Reza_canadianinfo.txt"

In order to extract and interpret the data, one needs to read each data point as well as the flag corresponding to the data point.

B.3 New functions to write stations' data

In the Python package, "Reza_canadian_data.py", we have also introduced some functions to write the data for a given station including the stations' information as longitude, latitude and elevation. Using these commands has the advantage that we do not need to worry about the flags anymore. Whenever, the data is missing we will get NA (instead of -999.9) and also for trace values (precipitation occurrence with value smaller than 0.2 (mm)) for precipitation, we get "trace". The command for getting the data for a given stations is "write_station(, ,)". We need three entries for this command: station number, the list of all stations (We can get that by the command stations=get_station_list(Topdirectory) as shown above.) and value:

("MT","mint","rain","snow","precip").

For example consider:

```
>>>write_station(2436,stations,'MT')
```

The output for this command (if the data are available) is a text file. There is also an output in the terminal. If the data are available the output is "success" and the name of the text file created. If the data are not available then the output is simply "failure". A statement is also printed depending on the data being available or not. If the data are available, the number of years the data are available is reported and also the name of the file created. For the above example, we get:

```
The data file 3025480-MT.txt created. It should contain 69 years.
('success', '3025480-MT.txt')
```

If the data are not available, we get:

```
There was not any years containing more than 100 days. No file
created. ('failure','none')
```

Also note that, we write data for a year only if it contains more than 100 days of data. You can modify this easily by modifying the write function in the module.

The data files are named by the id followed by "MT", "mint" or "precip" which stand for Max temperature, min temperature and precipitation respectively. For example, since the id for ABBOTSFORD's id is "1100030" the file containing the data for maximum temperature for ABBOTSFORD is called "1100030-MT.txt"

Each row of the data files corresponds to a year. The first entry is the year and then 366 entries corresponding to the observed daily values for the given year. Whenever the actual year has 365 days only, the value corresponding to Feb 29th is recorded as NA (60th day of the 366 year).

Note that we can use this command in a "for" loop to write a bunch of stations. To keep track of the stations that have data available, we make a list of the created files. In the following we have done that with stations_list which contains the number of the stations that have data available. stations_list contains the name of the created files.

```
list=733,4034,2517,7467,6744,1518,2113,7269 subset=list value='MT'
stations_list=[] stations_files=[]

for i in subset:
        snum=i
        d=write_station(snum,stations,value)
        if d[0]=='success':
                stations_list.append(i)
                stations_files.append(d[1])
```

B.4 Concluding remarks

The software described in this report can be used to generate datasets suitable for analysis with R and other standard datasets. Moreover the tutorials and demonstrations should help users understand the process for doing so.

Appendix C

Algorithms and Complexity

In this appendix, we include the definitions for the complexity of the algorithms. For a more detailed treatment see [28] for example.

Definition We say that $f(x) = o(g(x))$ if $\lim_{x\to\infty} f(x)/g(x)$ exists and is equal to 0.

Definition We say that $f(x) = O(g(x))$ if $\exists C; x_0$ such that

$$|f(x)| < Cg(x), \ \ \forall x > x_0.$$

Definition We say that $f(x) = \Theta(g(x))$ if there are constants $c_1 \neq 0; c_2 \neq 0; x_0$ such that for all $x > x_0$ it is true that $c_1 g(x) < f(x) < c_2 g(x)$.

Definition We say that $f(x) = \Omega(g(x))$ if there is an $\epsilon > 0$ and a sequence $x_1, x_2, x_3, \cdots$

$$x_n \to \infty \text{ as } \ n \to \infty,$$

such that $\forall j : |f(x_j)| > \epsilon g(x_j)$.

Appendix D

Notations and Definitions

We follow the widely used conventions throughout the thesis. Latin uppercase letters, often X, Y, Z, sometimes with subscripts such as s, t, are used for random variables.

List of notation and abbreviations:

$\mathbb{R}$	The real numbers: $(-\infty, \infty)$
$\mathbb{N}$	The natural numbers: $1, 2, \cdots$
$\mathbb{Z}$	The integer numbers: $\cdots, -2, -1, 0, 1, 2, \cdots$
$\sim$	Distributed as
$\approx$	Approximate to
Σ	A σ-field
(Ω, Σ, P)	A probability space over the set Ω and σ-algebra Σ
X	A random variable; $X : (\Omega, \Sigma, P) \to (\mathbb{R}, \mathcal{B})$ (a measurable function from Ω to $\mathbb{R}$ with Borel σ-field)
F_X	The distribution function of the random variable X
$X\|Y$	Random variable X conditional on random variable Y
$\hat{\alpha}$	Estimate of α
$N(\mu, \sigma^2)$	Univariate normal distribution with mean μ and variance σ^2
$\{X_t\}_{t \in T}$	A stochastic process over the space T
station	Gauged site where measurements are available
$i.i.d$	Independently and identically distributed
$E(X)$	Expectation of random variable X
$Var(X)$	Variance of random variable X
$Cov(X, Y)$	Covariance of random variables X and Y
LHS	Left hand side
RHS	Right hand side
iff	If and only if
$\emptyset$	The empty set

$sort(x)$	The sorted version of the vector x
$stack(x, y)$	Put (concatenate) the two vectors x and y together (starting from x and ending with y) to make a single vector
$length(x)$	The length (dimension) of a vector x
$\underset{a}{argmin} f(a)$	The set of elements of domain of f that minimize f.
MT	Maximum temperature during a day
mt	Minimum temperature during a day
PN	Precipitation amount (or occurrence) during a day
COS and SIN	The deterministic process $\cos(\omega t)$, and $\sin(\omega t)$, where t denotes time and $\omega = \frac{2\pi}{366}$.
AIC and BIC	Akaike information and Bayesian information criterion
$A \subset B$	A is a subset of B. It is possible that $A = B$
$p_n \uparrow p$	The sequence p_n is non–decreasing and tends to p
$p_n \downarrow p$	The sequence p_n is non–increasing and tends to p
$X_{(i)}$ or $X_{i:n}$	The ith order statistics of a random sample $X_1, \cdots, X_n$
$\{x\|P(x)\}$	The set defined by elements that satisfy the property $P(x)$

Definitions

- F_X or F_X^c: The distribution function of a random variable; $P(X \leq x)$.
- F_X^o: The open distribution function; $F_X^o(x) = P(X < x)$.
- G_X^o: The open right distribution function; $G_X^o(x) = P(X > x)$.
- G_X^c: The closed right distribution function; $G_X^c(x) = P(X \geq x)$.
- $lq_X(p)$: The left quantile function; $lq_X(p) = \inf\{x|F_X(x) \geq p\}$.
- $rq_X(p)$: The right quantile function; $rq_X(p) = \inf\{x|F_X(x) > p\}$.
- $(\Omega, P, \{X_\theta\}_{\theta\in\Theta})$: A statistical space consisting of a set Ω and a probability measure P on Ω and a set of random variables $\{X_\theta\}_{\theta\in\Theta}$ indexed by parameter θ in the parameter space Θ a subset of the Euclidean space.
- $1_A(x)$ is the standard indicator function formally defined as

$$1_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}.$$

- δ_X: The probability loss function associated with the random variable X,
$$\delta_X(a, b) = P(a < X < b) + P(b < X < a).$$