Novel likelihood-based inference for symbolic data analysis

Author:
Lin, Huan

Publication Date:
2018

DOI:
https://doi.org/10.26190/unsworks/20863

Novel likelihood-based inference for symbolic data analysis

Huan Lin

Supervised by: Prof. Scott A. Sisson

A thesis in fulfilment of the requirements for the degree of
Doctor of Philosophy

School of Mathematics and Statistics
Faculty of Science

November 2018

Thesis/Dissertation Sheet

Surname/Family Name	:	**Lin**
Given Name/s	:	**Huan**
Abbreviation for degree as give in the University calendar	:	**PhD**
Faculty	:	**Faculty of Science**
School	:	**School of Mathematics and Statistics**
Thesis Title	:	**Novel likelihood-based inference for symbolic data analysis**

Abstract 350 words maximum: (PLEASE TYPE)

Symbolic data analysis (SDA) is a relatively new branch in statistics. It has emerged from the need to consider data that contain information which cannot be satisfactorily represented and modelled within classical data models. SDA is a new paradigm which extends the classical data models to take into account more complete and complex information and serves as an alternative solution to tackle "big data" problems by reducing and summarising data of massive size to "classes" of interest. SDA organises multiple unstructured data tables to a single coherent data table containing symbolic-valued variables, often recorded in the forms of intervals or histograms. There has been a considerable amount of research in this area with many of the existing methods developed based on a uniformity within a symbol assumption. It has shown that this uniformity assumption is unrealistic in solving real-world problems.

Likelihood functions are fundamental in statistical inference and to date; two likelihood-based methods for SDA have been introduced. However, while these methods have shown to be beneficial, there are a number of current methodological weaknesses that limit their potential to become an invaluable tool in a modern statistician's toolkit. To this end, we propose new models for performing likelihood-based inference for SDA. Our approach overcomes the need to assume uniformity within a symbol (intervals or histogram bins) assumption which is conventional in SDA literature. Instead, our approach allows for a natural way of specifying the underlying distribution of the data from which symbolic variables are obtained. As a result, our approach enables statistical inference to be made at the underlying data level which may be more desirable, from the point of view of the statistical analyst. In addition, our approach offers an opportunity for statistical analysts to use higher-dimensional symbols to address complex real-world problems. The new models are demonstrated by simulated case studies. In addition, the proposed symbolic likelihood function for histograms has been applied to improve the analytical results of an existing model in measuring aerosol particle number concentration.

COPYRIGHT STATEMENT

‘I hereby grant the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all proprietary rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.
I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstract International (this is applicable to doctoral theses only).
I have either used no substantial portions of copyright material in my thesis or I have obtained permission to use copyright material; where permission has not been granted I have applied/will apply for a partial restriction of the digital copy of my thesis or dissertation.'

Signed ……………………………………………...........................

Date ……………………………………………...........................

AUTHENTICITY STATEMENT

‘I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis. No emendation of content has occurred and if there are any minor variations in formatting, they are the result of the conversion to digital format.’

Signed ……………………………………………...........................

Date ……………………………………………...........................

ORIGINALITY STATEMENT

‘I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.’

Signed ………………………………………………..............

Date ………………………………………………..............

INCLUSION OF PUBLICATIONS STATEMENT

UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

Publications can be used in their thesis in lieu of a Chapter if:

- The student contributed greater than 50% of the content in the publication and is the "primary author", ie. the student was responsible primarily for the planning, execution and preparation of the work for publication
- The student has approval to include the publication in their thesis in lieu of a Chapter from their supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis

Please indicate whether this thesis contains published material or not.

☐ *This thesis contains no publications, either published or submitted for publication (if this box is checked, you may delete all the material on page 2)*

☐ *Some of the work described in this thesis has been published and it has been documented in the relevant Chapters with acknowledgement (if this box is checked, you may delete all the material on page 2)*

☐ *This thesis has publications (either published or submitted for publication) incorporated into it in lieu of a chapter and the details are presented below*

CANDIDATE'S DECLARATION

I declare that:

- I have complied with the Thesis Examination Procedure
- where I have used a publication in lieu of a Chapter, the listed publication(s) below meet(s) the requirements to be included in the thesis.

Name	Signature	Date (dd/mm/yy)

Postgraduate Coordinator's Declaration (to be filled in where publications are used in lieu of Chapters)

I declare that:

- the information below is accurate
- where listed publication(s) have been used in lieu of Chapter(s), their use complies with the Thesis Examination Procedure
- the minimum requirements for the format of the thesis have been met.

PGC's Name	PGC's Signature	Date (dd/mm/yy)

School of Mathematics and Statistics
The Red Centre, Centre Wing
Kensington Campus
UNSW Sydney, NSW 2051
Australia

Graduate Research School
Lvl 2 South Wing Rupert Myers Building
Gate 14 Barker Street Entrance
Kensington Campus
UNSW Sydney, NSW 2051
Australia

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Professor Scott Sisson for his continuous support and guidance throughout my PhD journey. I can still vividly remember the first meeting we had. It was at that time I knew that he would be a tremendous mentor for me. The past three and a half years have witnessed many ups and downs in my study but under whatever circumstances, Scott has always been patient and encouraging. His passion for research has had a profound influence on my decision to forge a career in academia. Scott's humour has always lifted the mood during our weekly meetings and provided me with the energy to start a week of hard work. Scott is truly the funniest supervisor and one of the smartest people I know. Under his guidance, I have acquired a strong foundation in symbolic data analysis and Bayesian methods. In addition, I have gained the confidence to communicate as a statistician.

I am truly grateful to work with Dr Boris Beranger who is like my secondary supervisor and mentor. He has taught me more than I could ever give him credit for here. He has always been very generous in sharing his valuable experience and shown me, by his example, what an independent and good researcher (and person) should be.

Also special thanks to the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) for providing financial support and for funding my overseas conferences. More importantly, I would like to thank ACEMS for offering an excellent platform for graduate students to collaborate and network with senior and world-renowned researchers. I would like to thank Dr Julian Caley for providing the global species abundance data for Chapter 3 and also for offering expert advice in translating my statistical findings to ecological ones. I owe thanks to Professor Kerrie Mergensen for offering me the opportunity to work with Dr Sam Clifford for the project in Chapter 5 and thank Dr Sam Clifford for his providing of the datasets and initial guidance. I also appreciate ACEMS for organising a series of mentor workshops to provide academic and non-academic support and for organising an annual student retreat for us to showcase our research.

In June 2017, I was fortunate to attend the sixth symbolic data analysis workshop held in Ljubljana, Slovenia where I got to meet some of the enthusiastic senior researchers from other countries. I am so grateful for this experience; in particular, because of the closeness of the symbolic data analysis community has further stimulated my interest in statistics.

To my fellow PhD students with whom I share the same office, thank you for always -“being there”- in the office, you have created a studious atmosphere and motivated me to study harder. To my fellow PhD colleagues who are also under the supervision of Professor Scott Sisson, thank you for discussing new research ideas with me. Special thanks to Dr Xuhui Fan for providing suggestions and help when I was stuck.

To all the professional and academic staff, I have encountered and worked with at UNSW, thank you for your assistance and for providing me with a wonderful opportunity to work as a part-time statistics tutor for the school. I truly treasure this tutoring experience as it has helped me to further develop my statistics and communication skills. I have also discovered a passion in teaching from this experience.

Last but certainly not least, I cannot thank my family enough. My father and mother have been supportive since the first day of my life. Your unconditional and unselfish love is my biggest mental strength that pushes me to work harder and to complete this challenging journey. Words simply cannot express my greatest gratitude to you. I hope I am on the way to becoming the person you are proud to call your daughter. To all my friends in Australia, back home and in New Zealand, thank you for not letting the location or the time drift us apart. I am deeply thankful and honoured to call you my friends. This is the end of my PhD study but it is the beginning of a brand-new chapter in my life. Thank you all for riding the journey with me.

Contents

Chapter 1

Introduction

Both Bayesian and Frequentist statisticians are accustomed to analysing "classical" data- that is, data whose realisations are single points in Euclidean space $\mathcal{X} \subseteq \mathbb{R}^p$. Data of this kind are generally organised into a $n \times p$ data matrix where each of n individuals (in rows, often called "statistical units") takes one single value which might be observed or missing for each of p variables (in columns). These variables can be described by single point quantitative (i.e numerical) and/or qualitative (i.e categorical) values from which exploratory analysis and statistical inference can be performed to extract knowledge from the data. A classical data table of credit card expenditures by a coterie of individuals can be seen in Table 1.1 taken from Billard and Diday (2006). Symbolic data analysis (henceforth, SDA),

i	Name	Month	Food	Social	Travel	Gas	Clothes
1	Jon	February	23.65	14.56	218.02	16.79	45.61
2	Leigh	May	28.47	8.99	141.60	21.74	86.04
3	Leigh	July	24.13	15.97	190.40	35.71	20.02
4	Tom	July	30.86	9.55	193.14	24.26	95.68
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 1.1 – A partial list of credit card expenditures on a range of each individual's itemised expenses (in dollars) for food, social entertainment, travel, gas and clothes over a 12-month period.

introduced by Diday (1987), provides a new way of thinking in modern Data Science by extending "classical data" to a set of classes of individual entities. In the SDA paradigm, classes of a given population are considered to be units of interest to be studied. Table 1.2 illustrates such an example where the analysis of interest lies in the class of person-months and its corresponding spending patterns. In order to take into the account of variation within each class, each observation is now in an interval-valued format. The intervals correspond to the range of classical values observed in that person-month in the original classical data. Other common representations of classes include histograms, distributions, set of categories or numbers sometimes weighted and the like. SDA is particularly relevant in the era of big data as it can tackle complex data challenges by first building the symbolic data table where the rows are "classes", representing subsets of individual entities possessing a common characteristics. The columns contain variables taking symbolic

Name-Month	Food	Social	Travel	Gas	Clothes
Jon-January	[20.81, 29.38]	[9.74, 18.86]	[192.33, 205.23]	[13.01,24.42]	[44.28,53.82]
Jon-February	[21.44, 27.58]	[10.86, 18.01]	[214.98, 229.63]	[16.08,22.96]	[50.51,63.57]
⋮	⋮	⋮	⋮	⋮	⋮
Tom-January	[23.28, 30.00]	[8.67, 18.31]	[193.53, 206.53]	[26.28,35.61]	[15.51,25.66]
Tom-February	[20.61, 28.66]	[10.66, 17.20]	[195.53, 203.83]	[25.43,34.18]	[12.99,24.88]
⋮	⋮	⋮	⋮	⋮	⋮
Leigh-January	[25.59,35.33]	[7.07,19.00]	[194.12,207.05]	[17.75,23.07]	[61.47,75.43]
Leigh-February	[31.30,40.80]	[9.05, 24.44]	[212.76, 227.43]	[13.81,25.08]	[71.43,85.58]
⋮	⋮	⋮	⋮	⋮	⋮

Table 1.2 – Credit card use by person-months.

values. A symbolic data table constructed from classes aggregates complex data from multiple unstructured data tables to a data table with a compact structure. It consequently reduces the data set into a manageable size. In addition, SDA is useful to researchers whose scientific interests are units of second-level generalisation of the individual entities, represented by these "classes". For example, a loan company could define different classes of borrowers, e.g. high risk, medium risk, low risk based on their financial profiles. An environmentalist could define different classes of air particulate matter, e.g. PM_{10}, $PM_{2.5}$ and ultra-fine particles based on the particle diameters. As demonstrated in Table 1.3 extracted from Diday (2016), to assess the likelihood of a soccer team winning the French Cup, it is perhaps more sensible to analyse at the "class" level defined by teams in the French Cup instead of assessing an individual player within a "class". In official statistics, confidentiality issues renders the data custodians to release data of individual entities and only "class" level data are made available to the public. While thinking by classes in data can produce a smaller data set containing symbolic data, some data are naturally "symbolic". Blood pressure, for example, is usually recorded in an interval-valued format due to its continuous fluctuations.

French Cup Teams	Weight	National Country	Age
Paris	[73,85]	France, Argentina, Senegal	(0) 0.3, (1) 0.7
Lyon	[68,90]	France, Brazil, Italia	(0) 0.3, (1) 0.65, (2) 0.5
Marseille	[77,85]	France, Brazil, Algeria	(1) 0.4, (1) 0.52, (3) 0.08
Bordeaux	[80,90]	France, Argentina	(0) 0.4, (1) 0.6

Table 1.3 – Symbolic Data Table where "Classes" are Teams of the French Cup and four variables taking symbolic values of Interval, Sequence of Categories and Histogram. These symbolic data describe the classical data of the players in each soccer team.
Age variable represents the frequency of the age players being in the intervals [less than 20], [20,25], [25,30], [more than 30],respectively, coded as: (0), (1), (2), (3).

The term 'symbolic' stresses the fact that the values that symbolic data take are of a different nature to classical data. They are essentially distributional in nature and cannot be reduced to numbers without losing a significant amount of information. Statistical techniques for analysing classical data have been developed to analyse univariate or multivariate single-valued variables however, without modifications, these are not appropriate for handling distributional variables which have complex internal structures. To this end, there has been a considerable development within SDA research to develop methods to accommodate data with intrinsic variability and to thereby extend the scope of classical data analysis methods to a broader definitions of data.

To date, descriptive statistics for a variety of symbolic variables such as symbolic sample means, sample variances and covariances have been established. See e.g. Bertrand and Goupil (2000), Gioia and Lauro (2005), Billard and Diday (2006), Billard (2007), Billard (2008) and a critical review can be found in Irpino (2013). Since then, a large number of non-parametric descriptive approaches have been developed such as principal component analysis, clustering and linear regressions based on least squares, resulting in a large number of publications in journals and conference presentations.

In contrast, parametric inference methodologies remain largely unexplored. Le-Rademacher and Billard (2011) proposed a symbolic likelihood function and gave examples for interval and histogram-valued symbolic variables. Brito and Duarte Silva (2012) used this approach to propose probabilistic models for interval data assuming a joint multivariate Normal or skew-Normal distribution for the midpoints and log-ranges of the interval variables. This basic structure has been used for a variety of multivariate analyses for interval data including analysis of variance, discriminant analysis (Silva and Brito, 2015), model-based clustering (Brito et al., 2015) and outlier detection (Silva et al., 2017). An alternative likelihood-based inferential method that is able to incorporate both intra- and inter- symbol variations was proposed by Zhang and Sisson (2016) who demonstrated applications using interval-valued and histogram-valued symbolic variables.

Continuing forward from current developments in SDA, this thesis contributes a new method to construct symbolic likelihood functions for interval-valued and histogram-valued data. This method addresses some potential issues in the likelihood-based methods of Le-Rademacher and Billard (2011) and Brito and Duarte Silva (2012), including whereby inferences can only be made at the “class” level, overlooking the generative process of the underlying data. The proposed method also alleviates the need for assuming within ‘class’ uniformity within intervals and histogram bin data. This assumption is endemic in the SDA literature with many methods dependent on it being true. Of course, uniformity will not be true in practice, invalidate many existing SDA methods.

Zhang and Sisson (2016) construct likelihood functions for interval-valued symbols which can be considered as a first attempt to move beyond the issues implicit to the other likelihood-based SDA approaches, and provide an inferential framework that aligns a generative process with an aggregation process. Similar to Zhang and Sisson (2016)’s proposal, our new method considers both the generative process from which the classical data are observed and an aggregation process of transforming classical data points to defined “classes” taking symbolic values. In this way, inferences can be made at both the “class” level and at the underlying classical data level, for which the latter may be of greater interest to the analysts. In this manner, our likelihood-based approach can be shown to reduce to the regular classical data likelihood function when the symbolic variables reduce to classical data. In summary, this method aims to achieve firstly, explicitly recognising and accommodating the data generative and aggregation processes; Secondly, departing from the restrictive and unrealistic uniformity within "classes" assumption; Thirdly, provide more direct interpretations on the underlying classical data aggregated within individual entities.

The contents of this thesis are structured as follows, Chapter 2 presents a comprehensive literature review of SDA methodology. It begins with a detailed summary of different kinds of symbolic data with particular emphasis placed on interval-valued and histogram-valued symbolic variables, explaining reasons why current techniques for classical data analysis fail to adequately account for symbolic data. This is then followed by a summary of the current state of developments in SDA in the areas of exploratory data analysis, cross-sectional and time series, linear regressions and likelihood-based inference. A brief discussion about the research challenges within SDA is presented and how these motivate the introduction of a new likelihood-based approach in Chapter 4.

In spite of the growing attention in biodiversity conservation (Caley et al., 2014), ecological data such as estimates for species abundance is highly variable and cannot be determined with certainty. Typically, data of species estimates are recorded in one of three different forms. Some might be point-estimates, representing experts' most informed best estimate of the number of species based on a detailed analysis or study. Others might be provided as intervals, representing a plausible range of species counts. The remainder might be supplied with both a point estimate and a possible range recorded as an interval. Due to this inconsistent form of data, a previous analysis by Caley et al. (2014) was unable to appropriately capture all of the information in the species diversity data, thereby reducing the authors' ability to derive satisfactory estimations for global species richness and for individual species taxa. In Chapter 3, a novel meta analysis application for estimating the global and individual taxa species abundance is presented, adopting the existing likelihood-based SDA approach of Brito and Duarte Silva (2012). This approach, together with a Bayesian hierarchical model that is more complex than those previously considered, allows us to statistically reconcile the three different data formats, and also enforce logical consistency in estimating species abundance within and between different species categories.

While the above analysis makes no assumption of the information of within interval-valued data, it could be symmetric or asymmetric, the information about the internal distribution within each interval might be relevant in a given analysis.

Acknowledging this need and as previously discussed, in Chapter 4, we introduce our new method to construct symbolic likelihood functions for interval- and histogram-valued data. This method is capable of capturing both intra- and inter- symbol relationships in addition to ensuring that the choice of the internal distribution within a "class" taking symbolic values is at the researchers' discretion. It accordingly offers significantly more flexible and accurate modelling.

In Chapter 5, we apply our new symbolic likelihood framework to a particulate matter data analysis. Ultrafine particles (UFPs), whose diameters are less than 100 nm are ubiquitous in urban air and are acknowledged to have adverse risk to climate, visibility and human health. Due to their negligible mass compared with larger-sized particles (such as PM10 and PM 2.5), UFPs are commonly evaluated through measurements of particle number concentration (PNC) Hussein et al. (2005). To track the dynamic evolution of PNC and its impact on children's health, a measurement campaign by the International Laboratory for Air Quality and Health (ILAQH) entitled "Ultrafine Particle Emissions from Traffic and Child Healt" (UPTECH) was conducted in Brisbane where PNC levels were measured continuously over a 2-week period at 25 primary schools in the Brisbane Metropolitan Area. To better understand aerosol dynamic processes, PNC measurements are collected at every 5 minutes at each school site, resulting in 12 observations per hour.

Previous analysis performed by Clifford et al. (2012a) constructed a Bayesian spatio-temporal model for PNC using only the hourly-averaged data. This is potentially suboptimal use of the collected data and motivates us representing each 5-minute PNC measurement as a member of a "class" defined by hour. Now these "classes" of hourly PNC can no longer be summarised by a single point. In Chapter 5, these classes are chosen to be histograms constructed from data quantiles. The modelling of temporal effects for these histogram-valued symbolic data is consistent with the approach taken by Clifford et al. (2012a). Specifically we adopt a Bayesian time-varying finite mixture model to account for potential multi-modality induced from heterogeneity in the underlying sequences of histograms. Allowing the parameters of the mixture model to vary over time provides some insights into how the underlying distribution evolves over time.

In Chapter 6, we conclude with a discussion. We make some comments about our new way of constructing symbolic likelihood functions and how it can contribute to the analysis of large and complex datasets, and discuss potential future research directions.

Chapter 2

Literature Review of Symbolic Data Analysis

2.1 Introduction

Chapter 1 briefly discusses how Symbolic Data Analysis (SDA) provides a systematic way of thinking and representing classical data, aggregated to new kinds of data "points", called 'symbolic' data organised by "classes". This Chapter goes into details. Firstly the concepts of symbolic data are formally presented. This is followed by discussing the similarities and differences with classical data. Symbolic data are distinctive in their own right and thus motivate the need to establish a new analysis framework. Special attention is given to quantitative interval-valued and histogram-valued symbolic data. We review the descriptive statistics basics and some methods of analysing these data, while briefly summarising methods for other types of symbolic variables, before highlighting some weaknesses of these current approaches. This thesis will tackle some of these problems.

2.2 From Classical to Symbolic Data

"Big data" has become a buzzword in modern Data Science. It commonly refers to structured or unstructured data with immense volumes and complexity. As summarised by Diday (2016), SDA is a new tool in Data Science that offers an efficient and effective way of knowledge extraction from data. It can be used to tackle data of big volumes by aggregating the micro-data (individual observations) to a much smaller number of group level symbols ("classes") (where $m \ll n$, m is the symbolic data size while n is the size of the classical data). At the same time, addressing data complexity as thinking by classes requires a transformation of multiple unstructured data tables and unpaired variables to a symbolic data table where the rows are classes and columns are symbolic variables. The following example, extracted from Diday and Noirhomme-Fraiture (2008), demonstrates how two seemingly unrelated data tables with different individuals and different variables can be merged into a symbolic data table by using a common "class". In Table 2.1, individual entities are schools which are described by 1 quantitative variable (the number of pupils per school) and 3 qualitative variables (town, school type and school level). In contrast, the individual entities in Table 2.2 are hospitals and are described by 1 quantitative variable (the number of coded beds per hospital) and 2 qualitative variables (town and coded speciality). In both tables, only the variable 'town' is common. Using the idea from SDA, the two tables can be aggregated into a single table (2.3) by thinking of 'towns' as "classes". It becomes immediately apparent that the towns then constitute "classes" defined by the same quantitative variables (No.of pupils and No.of beds), which are no longer single values but intervals. The qualitative variables such as the type of school of Table 2.3 is transformed into a new variable whose values are several categorical with weights defined by the frequencies of school types in each town. In this way, we obtain these new kinds of data, called 'symbolic' data, presenting variability between the individual entities within a class. It is also worth noting that some data are naturally "symbolic" , in the sense that there is inherent variability within data. It is briefly discussed in Billard and Diday (2003b) that pulse rate, systolic blood pressure and diastolic blood pressure are types of data that are generally recorded as intervals. In addition, birds may be characterised by colours e.g., Bird 1 ={black}, Bird 2 ={black and blue} and Bird 3 ={half yellow, half red},... That is, the variable colour of an individual bird takes not just one category of colours but could be a list of all possible colours or a list with corresponding proportion of each colour for that individual bird. However, this thesis focuses on symbolic data that are obtained from an aggregation process where individual entities are generalised to different "classes".

School	Town	No. of pupils	Type	Level
Janres	Paris	320	Public	1
Condorcet	Paris	450	Public	3
Chevreul	Lyon	200	Public	2
St Helene	Lyon	380	Private	3
St Sernin	Toulouse	290	Public	1
St Hilaire	Toulouse	210	Private	2

Table 2.1 – A Classical Data Table of schools in different towns in France

School	Town	Coded no. of beds	Coded speciality
Janres	Paris	750	5
Condorcet	Paris	1200	3
Chevreul	Lyon	650	3
St Helene	Lyon	720	2
St Sernin	Toulouse	520	6
St Hilaire	Toulouse	450	2

Table 2.2 – A Classical Data Table of hospitals in different towns in France

Suppose a $n \times p$ data matrix $X = (X_{ij})$, where each cell (i,j) contains x_{ij}, the observed value of variable $j, j = 1,2,3,...p$ for individual $i \in \Omega = \{1,2,3,...n\}$. Note n is the number of the observations while p is the number of variables and both n and p can be extremely large, which is fairly common in the era of "big data". In addition, let the domain of X_j be $\mathcal{X}_j$ then $X = (X_1, X_2, ...X_p)$ takes values in $\mathcal{X} = \times_{j=1}^{p} \mathcal{X}_j$. In classical data analysis,

Town	No. of pupils	Type	Level	Coded no.of beds	Coded speciality
Paris	[320,450]	100% Public	$\{1,3\}$	[750,1200]	$\{3,5\}$
Lyon	[200,380]	50% Public, 50% Private	$\{2,3\}$	[650,720]	$\{2,3\}$
Toulouse	[210,290]	50% Public, 50% Private	$\{1,2\}$	[450,520]	$\{2,6\}$

Table 2.3 – A Symbolic Data Table of schools and hospitals constructed from "classes"–towns in France

variables can be either quantitative (numeric), e.g., age with $\mathcal{X}_{age} = \{x >= 0\} = \mathcal{X}_+$ as a continuous random variable; or with $\mathcal{X}_{age} = \{0, 1, 2, ...\} = \mathcal{N}_0$ as a discrete random variable. Qualitative variables could be $\mathcal{X}_{grade}$ = {High Distinction, Distinction, Credit, Fail} or coded $\mathcal{X}_{grade} = \{1, 2, 3, ...\}$ respectively. It can also be an indicator variable, e.g., X = Passing the exam with domain $\mathcal{X}$ = {No, Yes}. In the classical data setting, there is exactly one realised value for each x_{ij} in X, for example, an individual's X_{age}= 24 whose X_{grade} is High Distinction and $X_{passing}$ = Yes. In summary, a classical data point is a single point in the p-dimensional space $\mathcal{X}$.

In this notation setting, let us further define $S_j, j = 1, 2, 3, ...p$, be the j^{th} symbolic variable, with a particular variable S_j assuming a value ξ_{ij} for the i^{th} individual. The values that symbolic variables take are of a different nature to classical variables as they are not restricted to just a particular point value. Generally, a symbolic data point ξ_{ij} is a distribution of some kind with an internal distributional structure. Given this unique characteristic, SDA has to deal with the internal variation of each observation in addition to the variation between observations. In contrast, a classical observation with its single point value has no internal variation and classical data analysis is developed to deal with variation between observations only. As a result, classical data analysis is not appropriate to analyse symbolic data which leads to the emergence of SDA aiming at extending the classical data models to account for data with growing complexity. However, similar to their classical counterparts, they can still be classified as either quantitative (numeric) or qualitative (categorical) variables. Firstly, we introduce quantitative symbolic variables with special attention given to intervals (Section 2.3.1) and histograms (Section 2.3.2) before introducing more general quantitative symbolic variables in Section 2.3.3. In Sections 2.4.1 and 2.4.2, we discuss some qualitative symbolic variables. In Sections 2.5.1, 2.5.2 and 2.5.3 we present a short description of some of the current SDA approaches and methods to analyse symbolic data. Section 2.6 concludes this Chapter with a brief discussion, pointing out some of the key issues which remain to be solved in current SDA methodologies. This thesis hopes to address some of them.

2.3 Quantitative Symbolic Variables

Let us define the following notation to be used consistently throughout this chapter.

- ξ_{ij} be an entry assumed by a symbolic variable S_j for the i^{th} observation, where $i = 1...n, j = 1...p$.
- Given S_j is an interval-valued symbolic variable, its realisation is denoted as ξ_{ij} =

$(l_{ij}, u_{ij}) \subseteq \mathbb{R}$, where u_{ij} is the upper bound and l_{ij} is the corresponding lower bound, with $u_{ij} \geq l_{ij}$ and $l_{ij}, u_{ij} \in \mathbb{R}$.

- Given S_j is a K-bin histogram-valued symbolic variable, its realisation is represented as $\xi_{ij} = ((l_{ij1}, u_{ij1}; p_{ij1}), ..., (l_{ijK}, u_{ijK}; p_{ijK}))$. u_{ijk} is the upper bound of the k^{th} subinterval and l_{ijk} is the corresponding lower bound, with $u_{ijk} \geq l_{ijk}$, $l_{ijk}, u_{ijk} \subseteq \mathbb{R}$, $\sum_{k=1}^{K} p_{ijk} = 1$ and $0 \leq p_{ijk} \leq 1$ for $k = 1...K$.

2.3.1 Interval-Valued Symbolic Variables

An interval-valued variable is one of the most commonly seen quantitative symbolic variables, whose values are finite subsets of its domain. Continuing from the example above, an aggregating process transforms individual entities to a "class", defined by, say, university students who study MATH1041 at UNSW. The equivalent symbolic version of quantitative variables S_{age} may now be recorded as an interval $\xi_{age} = (20, 28) \subseteq \mathbb{R}$. In other possible cases, the same interval variable may be obtained for an single individual whose age is not precisely known, or whose age has varied over a duration of time where a longitudinal study was undertaken and thus producing interval-ranged data. An interval ξ_{age}, for example, is represented by its lower and upper bounds. Given it is obtained from an aggregation process from which individual entities who share the same pre-defined characteristics are considered as a "class", then the lower bound may represent the youngest individual in this "class" and the upper bound may represent the oldest individual in this "class". Alternatively, the two bounds can be two order statistics underlying the individual entities. For modelling purposes, Brito and Duarte Silva (2012) suggested to use an equivalent parametrisation: the midpoint ($\frac{l_{ij}+u_{ij}}{2}$) and *log* range ($log(u_{ij} - l_{ij})$) of the interval. It is not hard to see that a classical quantitative variable say, $X_{age} = 20$ is a special case of an interval variable, whenever $u_{age} = l_{age}$. It can be interpreted as there is no variability between individuals within this "class", or there is no uncertainty in this individual's age.

Bertrand and Goupil (2000) first defined the empirical density function, sample mean and sample variance for interval-valued symbolic variables, many examples can be found in Billard and Diday (2006). Billard (2007, 2008) obtained the sample covariance for interval-valued data. All of these summary statistics are based on a single most important assumption that the micro-data within random intervals and rectangles is uniformly distributed. However, as it is shown in Kosmelj et al. (2014)'s analysis and also in real data analyses in this thesis, the uniformity assumption is rarely satisfied.

Since a uniformity assumption is assumed for a point S_{ij} in S_j over the observed interval (l_{ij}, u_{ij}), we have

$$P(\xi_{ij} \leq \xi) = \begin{cases} 0, & \xi \leq l_{ij} \\ \frac{\xi - l_{ij}}{u_{ij} - l_{ij}}, & l_{ij} \leq \xi \leq u_{ij} \\ 1, & u_{ij} \leq \xi. \end{cases}$$

Definition 2.3.1. For an interval-valued random variable S_j with n observations, the empirical density is

$$f_{S_j}(\xi) = \frac{1}{n} \sum_{i:\xi \in \xi_{ij}} \left(\frac{1}{u_{ij} - l_{ij}}\right). \tag{2.1}$$

Note that the summation in Equation (2.1) is only summing over the i^{th} observation for which $\xi \in \xi_{ij}$. In addition to a uniformity assumption within an interval, it further assumes that each interval-valued symbol is equally likely to be observed with probability $\frac{1}{n}$.

Definition 2.3.2. For an interval-valued random variable S_j, the symbolic sample mean $\bar{S_j}$ is

$$\bar{S_j} = \frac{1}{2n} \sum_{i=1}^{n} (l_{ij} + u_{ij}). \tag{2.2}$$

Definition 2.3.3. For an interval-valued random variable S_j, the symbolic sample variance $\widetilde{S_j^2}$ is

$$\widetilde{S_j^2} = \frac{1}{3n} \sum_{i=1}^{n} (l_{ij}^2 + l_{ij}u_{ij} + u_{ij}^2) - \frac{1}{4n^2}\left[\sum_{i=1}^{n} (l_{ij} + u_{ij})\right]^2. \tag{2.3}$$

Detailed derivations of the above two statistics can be found in Bertrand and Goupil (2000) and Billard and Diday (2006). It can be easily verified that these two statistics reduce to the classical sample mean and sample variance when an interval variable shrinks to a single point $l_{ij} = u_{ij} = x_{ij}$ for all observations. Billard (2007, 2008) illustrated that sample variance defined in Equation (2.3) is a function of the total sum of squares (SST) of the interval-valued observations $S_i, i = 1...n$, and that the SST can be further divided into the sum of the internal variation, called the within sum of squares (SSW), and the external variation, called the between sum of squares (SSB). To be more specific, we can write

$$nS_j^2 = SST = SSB + SSW \tag{2.4}$$

where

$$SSB = \sum_{i=1}^{n} (\bar{S_{ij}} - \bar{S_j})^2. \tag{2.5}$$

and

$$SSW = \frac{1}{3} \sum_{i=1}^{n} [(l_{ij} - \bar{S_{ij}})^2 + (l_{ij} - \bar{S_{ij}})(u_{ij} - \bar{S_{ij}}) + (u_{ij} - \bar{S_{ij}})^2] \tag{2.6}$$

where

$$\bar{S_j} = \sum_{i=1}^{n} \frac{(l_{ij} + u_{ij})}{2n}. \tag{2.7}$$

$$\bar{S_{ij}} = \frac{(l_{ij} + u_{ij})}{2}. \tag{2.8}$$

Given $(u_{ij} - \bar{S_{ij}}) = (\bar{S_{ij}} - l_{ij}) = \frac{(l_{ij}+u_{ij})}{2}$, it follows that Equation (2.6) becomes

$$SSW = \frac{1}{12} \sum_{i=1}^{n} (u_{ij} - l_{ij})^2. \tag{2.9}$$

It is worth noting that Equation (2.9) is the formula for the sample variance of a random variable coming from a uniform distribution. The result in Equation (2.9) is consistent with the assumption that the values within an interval (l_{ij}, u_{ij}) are uniformly distribution. In other words, given $S_{ij} \overset{i.i.d}{\sim} uniform(l_{ij}, u_{ij})$, for $i = 1...n, j = 1..p$, then

$$Var(S_{ij}) = \frac{(u_{ij} - l_{ij})^2}{12}. \tag{2.10}$$

This is followed by the total within interval variance of the n observations $S_1, ...S_n$ is the sum of the variances in Equation (2.10), which is

$$Var(S_{ij}) = \frac{1}{12}\sum_{i=1}^{n}(u_{ij} - l_{ij})^2. \tag{2.11}$$

Let us consider interval-valued random variables S_1 and S_2 and suppose that $S_{i1} = (a_{i1}, b_{i1})$, $S_{i2} = (c_{i2}, d_{i2}), i = 1...n$. In addition, let us define a rectangle $A_{i,12} = [(a_{i1}, b_{i1}), (c_{i2}, d_{i2})]$. Billard and Diday (2003b, 2006) provided the empirical joint density function as

Definition 2.3.4. For bivariate interval-valued random variables S_1 and S_2, their empirical joint density function is

$$f(\xi_1, \xi_2) = \frac{1}{3n}\sum_{i:\xi \in \xi_{ij}^i} \frac{I_{ij}(\xi_1, \xi_2)}{\|A_{i,12}\|}, \tag{2.12}$$

where I_{ij} is the indicator function indicating whether (ξ_1, ξ_2) is in the rectangle $A_{i,12}$ or not and where $\|A_{i,12}\|$ is the area of the rectangle. Note that the summation in Equation (2.12) is summing over the i^{th} observation for which $\xi \in \xi_{ij}^i$.

Definition 2.3.5. For interval-valued random variables S_1 and S_2, Billard (2007, 2008) introduced the definition of the empirical covariance function $Cov(S_1, S_2)$. Assuming the realisations for S_{i1} and S_{i2} are (a_i, b_i) and (c_i, d_i) respectively.

$$Cov(S_1, S_2) = \frac{1}{6n}\sum_{i=1}^{n}[2(a_{i1}-\bar{S_1})(c_{i2}-\bar{S_2})+(a_{i1}-\bar{S_1})(d_{i2}-\bar{S_2})+(b_{i1}-\bar{S_1})(c_{i2}-\bar{S_2})+2(b_{i1}-\bar{S_1})(d_{i2}-\bar{S_2})] \tag{2.13}$$

with $\bar{S_1}$ and $\bar{S_2}$ defined as in Equation(2.2).
Given the definitions of sample variance and covariance, it is straightforward to define the correlation coefficient for interval-valued variables S_1 and S_2.

Definition 2.3.6. Let S_1 and S_2 be two interval-valued random variables. Then the sample correlation function $\gamma(S_1, S_2)$ with is

$$\gamma_{(S_1,S_2)} = \frac{Cov(S_1, S_2)}{\widetilde{S_1}\widetilde{S_2}} \tag{2.14}$$

where the covariance $Cov(S_1, S_2)$ is defined in Equation 2.13 and their corresponding standard deviations $\widetilde{S_1}, \widetilde{S_2}$ are obtained by taking the square root of the sample variances defined in Equation 2.3.

2.3.2 Histogram-Valued Symbolic Variables

When single-valued quantitative variables are aggregated into intervals, the information inside the interval is lost. One way to keep more information in the symbol is to incorporate the empirical distributions over a set of K non-overlapping subintervals. Let us define a K-bin histogram-valued symbolic variable $S_{ij}, i = 1...n, j = 1...p$, as the i^{th} histogram-valued observation of the j^{th} random variable, with a realisation ξ_{ij},

$$\xi_{ij} = \{[l_{ij1}, u_{ij1}), p_{ij1}; [l_{ij2}, u_{ij2}), p_{ij2}; ...[l_{ijK}, u_{ijK}), p_{ijK}\},$$

where $([l_{ijk}, u_{ijk}), p_{ijk})$ refers to the k^{th} subinterval of ξ_{ij} with relative frequency $\sum_{k=1}^{K} p_{ijk} = 1$. It is worth noting that interval-valued variables are particular cases of histogram-valued variables, when $K = 1$, $S_{ij} = [(l_{ij}, u_{ij}), p_{ij} = 1]$. Consequently, a single-valued quantitative variable is also a special case, e.g., $X_{ij} = [(l_{ij}, u_{ij}), p_{ij} = 1], l_{ij} = u_{ij}$.

Definition 2.3.7. For a histogram-valued random variable S_j, the empirical density is

$$f_{S_j}(\xi) = \frac{1}{n} \sum_{k:\xi \in \xi_{ijk}} (p_{ijk} \frac{1}{u_{ijk} - l_{ijk}}) \tag{2.15}$$

Note the summation in Equation (2.15) is summing over the k^{th} subinterval for which $\xi \in \xi_{ijk}$.

Definition 2.3.8. For a histogram-valued random variable S_j, the symbolic sample mean $\bar{S}_j$ is

$$\bar{S}_j = \frac{1}{2n} \sum_{i=1}^{n} \sum_{k=1}^{K} p_{ijk}(u_{ijk} + l_{ijk}). \tag{2.16}$$

Definition 2.3.9. For a histogram-valued random variable S_j, the symbolic sample variance $\widetilde{S_j^2}$ is

$$\widetilde{S_j^2} = \frac{1}{3n} \sum_{i=1}^{n} \sum_{k=1}^{K} p_{ijk}(l_{ijk}^2 + l_{ijk}u_{ijk} + u_{ijk}^2) - \frac{1}{4n^2} [\sum_{i=1}^{n} \sum_{k=1}^{K} p_{ijk}(l_{ijk} + u_{ijk})]^2. \tag{2.17}$$

Similar to interval-valued symbolic variables, summary statistics for histogram-valued symbols are derived assuming that the distribution of the micro-data within histogram bins is uniform. Detailed derivations of the above expressions can be found in Billard and Diday (2003b).

2.3.3 More General Quantitative Symbolic Variables

This class of quantitative symbolic variables include functional variables, where for each observation a function is recorded. Special attention was paid to considering cumulative probability function as symbols by Diday and Vrac (2005) and Cuvelier et al. (2009).

2.4 Qualitative Symbolic Variables

2.4.1 Categorical Multi-Valued Variables

The values of a multi-valued modal variable are finite sets of categories. Consider a variable S_j representing the brands of cars owned by a households, with domain $\{\mathcal{S}_{car}\}$ =\{Chevrolet, Ford, Toyota, Volvo,...\}. Then i^{th} household might be observed to have the value ξ_i, car = \{Toyota, Volvo\}, i.e., the household has two models of car.

2.4.2 Categorical Multi-Valued Modal Variables

Compared to multi-valued symbolic variables, a multi-valued modal variable not only takes values over finite sets of categories, but also with weights, frequencies or probabilities, indicating how frequent or likely that category is for this observation. For example, the above multi-valued variable can be elaborated to be ξ_i, car = \{Toyota $(\frac{1}{3})$, Volvo $(\frac{2}{3})$\} for a household with 3 cars.

2.5 Methods for the Analysis of Symbolic Data

Section 2.2 compares and contrasts the classical and symbolic variables with detailed descriptions, examples and formulas presented in Sections 2.3.1, 2.3.2, 2.3.3, 2.4.1 and 2.4.2 for different types of symbols. A conclusion which can be drawn from the above sections is that unlike classical data on p random variables which are single points in a $p-$dimensional space $\mathbb{R}^p$, symbolic data have internal variability due to assuming values as a hypercube or a Cartesian product of distributions in p-dimensional space or a mixture of both. The standard statistical framework has been developed to deal with variables with single-point values and no internal variability. For symbolic data, these results are no longer appropriate. The need to develop new methods that go beyond the traditional statistical framework to account for this intrinsic variability has led to the development of new techniques to model symbolic variables.

2.5.1 Methods For the Analysis of Interval-Valued Symbolic Data

Based on symbolic descriptive statistics for intervals, exploratory analysis such as principal component analysis (PCA) has first been addressed by Diday and Bock (2000) based on either the midpoints or the vertices of intervals. Montanari et al. (2005) propose a method based on both midpoints and ranges of interval-valued variables. While Gioia and Lauro (2006) proposed a method based on interval algebra and optimisation. Most recently, Le-Rademacher and Billard (2013) provides a method based on the symbolic variance and covariance matrix defined in Equation (2.3) and Equation (2.13) respectively. However, all these methods are less than ideal in the sense that they just map a distribution to a classical vector and then use the standard PCA methods. The method is not really adapted for interval-valued symbolic data themselves.

Linear regression based on least squares estimation has been heavily studied for interval-valued symbolic variables. The first linear regression model is proposed by Billard and

Diday (2000) where the authors built a classical linear regression model using centre points of the observed intervals and used the fitted model to predict the lower and upper bounds of the interval. It is known as the centre model. Assume $S_1, S_2 ... S_p$ are p independent interval-valued variables, and S_y is the dependent interval-valued variable. Let $S_i^c = (S_{i1}^c, S_{i2}^c, ..., S_{ip}^c)$ and S_{yi}^c, be the centre points of the interval-data with the observed values (a_{ij}, b_{ij}) and $(c_i, d_i), i = 1...n, j = 1..p$ respectively for the independent and dependent interval-valued symbolic variables, where the centre points can be calculated as

$$S_{ij}^c = \frac{(a_{ij} + b_{ij})}{2}, S_{yi}^c = \frac{(c_i + d_i)}{2}, i = 1...n, j = 1...p.$$

Then the fitted univariate linear regression model is

$$S_y^c = S^c \beta^c + \epsilon^c, \tag{2.18}$$

where $S_y^c = (S_{y1}^c, S_{y2}^c, ..., S_{yn}^c)', S^c = (S_1^c, S_2^c, ..., S_n^c)', \beta^c = (\beta_0, \beta_1, ...\beta_p)$ and $S_i^c = (1, S_{i1}^c, S_{i2}^c, ..., S_{ip}^c)'$ for $i = 1...n$ and ϵ^c is the error of the centre model.

Based on the classical regression framework, the least squares estimator of β^c is given by

$$\hat{\beta^c} = ((S^c)' S^c)^{-1} (S^c)' S_y^c.$$

Suppose we have a set of new covariates $S^{new} = (S_1^{new}, S_2^{new} ... S_p^{new})$, where $S_j^{new} = [S_{jL}^{new}, S_{jU}^{new}], j = 1...p$. Then the predicted interval $\hat{S_y} = [\hat{S_{yL}}, \hat{S_{yU}}]$ can be obtained by

$$\hat{S_{yL}} = (S_L^{new}) \hat{\beta^c}, \hat{S_{yU}} = (S_U^{new}) \hat{\beta^c}.$$

Neto et al. (2004) incorporated both the centre points and ranges of the interval-valued data into modelling, but they were modelled independently. The centre model is the same as defined in Equation(2.18) and is augmented by a range model with ranges defined as

$$S_{ij}^r = (b_{ij} - a_{ij}), S_{yi}^r = (d_i - c_i), i = 1...n, j = 1...p.$$

Then the range model is

$$S_y^r = S^r \beta^r + \epsilon^r, \tag{2.19}$$

where $S_y^r = (S_{y1}^r, S_{y2}^r, ..., S_{yn}^r)', S^r = (S_1^r, S_2^r, ..., S_n^r)', \beta^r = (\beta_0, \beta_1, ...\beta_p)$ and $S_i^r = (1, S_{i1}^r, S_{i2}^r, ..., S_{ip}^r)'$ for $i = 1...n$ and ϵ^r is the error of the range model.

Based on classical regression, the least squares estimator of β^r can be obtained from

$$\hat{\beta^r} = ((S^r)' S^r)^{-1} (S^r)' S^r.$$

The predicted $\hat{S_y} = [\hat{S_{yL}}, \hat{S_{yU}}]$ given a set of new covariates $S_j^{new} = [S_{jL}^{new}, S_{jU}^{new}], j = 1...p$, can then be obtained by combing the estimates from both the centre and the range model.

$$\hat{S_{yL}} = \hat{S_y}^c - \frac{\hat{S_y}^r}{2}, \hat{S_{yU}} = \hat{S_y}^c + \frac{\hat{S_y}^r}{2},$$

where $\hat{S_y}^c = S^{new}\hat{\beta}^c$ and $\hat{S_y}^r = S^{new}\hat{\beta}^r$.

As acknowledged in Brito and Duarte Silva (2012), the lower and upper bounds (or equivalently their reparametrisations of centres and ranges) of an interval-valued variable are two quantities related to only one variable and should not be modelled separately. Unfortunately, neither of the above models succeeds in meeting this criterion. In fact, all of the above models have the undesirable characteristic that the predicted upper bounds can be smaller than the predicted lower bounds whenever the estimated slope becomes negative. To resolve this issue, Neto and de Carvalho (2010) and Giordani (2015) introduced some forms of constraints to ensure the logical predictions of the intervals. All the above methods approximate the internal variations by considering only the ranges. As a result, Xu (2010) developed a linear regression model based on symbolic sample covariance as defined in Equation (2.13) to better approximate the internal variation. Maia and de Carvalho (2008) proposed a least squares model based on absolute deviation to provide robust estimators in the presence of outliers. Lima Neto and dos Anjos (2015) proposed to represent the lower and upper bounds of the interval-valued response variable as a bivariate random vector and considered copula theory to induce a general bivariate distribution for this random vector. Most recently, Dias and Brito (2016) proposed to represent intervals by quantile functions and thereby considering the distributions within them. In their paper, two specific distributions namely a uniform distribution and a symmetric triangular distribution are studied. In essence, however, all the above methods are non-probabilistic, in the sense that the estimation of regression parameters are based on the minimisation of an error criterion (eg. least squares). The above approaches are fine in model estimation but without probabilistic modelling, it is impossible to use inference techniques on the parameter estimates, such as hypothesis tests, confidence intervals etc. Neto et al. (2009) recognised this need and considered the probabilistic aspects related to the regression models for interval-valued variables, while Ahn et al. (2012) advocated using Monte Carlo resampling to fit a linear regression model to interval-valued data. In this way, one can derive approximate sampling distributions of the estimated regression coefficients.

Clustering is a multivariate technique whose aim is to allocate entities to homogeneous classes based on observed values in a set of variables. A significant number of clustering methods have been studied in the domain of interval-valued symbolic data. Clustering symbolic data can be broadly classified into two main categories: methods based on dissimilarities, by defining appropriate measures of distance measures for interval-valued data (Chavent and Lechevallier, 2002; de Souza and De Carvalho, 2004; de Carvalho et al., 2006; De Carvalho and Tenório, 2010) and conceptual clustering methods where classes are usually described by necessary and sufficient criteria based on generalisation processes (Brito, 2002; Brito et al., 2008).

Time series analysis is an important branch of statistics which aims to make forecasts or uncover the dynamic data generating process. Teles and Brito (2005) first examined interval-valued time series data by fitting univariate ARIMA processes to the interval bounds. Maia et al. (2008) fitted the same process to the midpoints and range of the intervals and used it for prediction. Ohter authors including Gallardo and Jiménez (2008), García-Ascanio and Maté (2010) and Ai et al. (2008) defined interval stochastic processes, interval-valued time series and weak stationarity based on interval-valued moments.

In terms of parametric modelling, Le-Rademacher and Billard (2011) proposed the first likelihood-based approach for interval-valued and histogram-valued symbolic variables. The essence of their approach is to map each symbol to a random vector that uniquely defines the symbol, and then specifies a standard statistical likelihood model for each of the observed symbols. For example, univariate micro-data $X_i, i = 1...n$ may represent log-debt for n individuals where n can be very large. Based on their socio-economic profiles, these individuals can be aggregated into m risk groups (where $m \ll n$) represented by interval-valued symbols $S_j = [l_j, u_j], j = 1...m$ whose interval bounds are determined by the minimum and maximum values of log-debts among all individuals in that risk group. Now assume a uniform distribution assumption within these intervals, Let Θ_{j1} be the internal mean of S_j and Θ_{j2} be the internal variance of S_j, with realisations $\theta_{j1} = \frac{(l_j+u_j)}{2}$ and $\theta_{j2} = \frac{(u_j-l_j)^2}{12}$ respectively. Firstly, assuming the independence between Θ_{j1} and Θ_{j2}, then the approach of Le-Rademacher and Billard (2011) assumes two appropriate distributions, such as

$$\Theta_{j1} \sim N(\mu, \sigma^2), \Theta_{j2} \sim \text{Exp}(\beta). \tag{2.20}$$

Then the joint likelihood function is

$$L(\mu, \sigma^2, \beta; \theta_1, \theta_2, ...\theta_m) = \Pi_{j=1}^{m}[f(\frac{(l_j+u_j)}{2}; \mu, \sigma^2) \times f(\frac{(u_j-l_j)^2}{12}; \beta)], \tag{2.21}$$

where $\theta_j = (\theta_{j1} = \frac{(l_j+u_j)}{2}, \theta_{j2} = \frac{(u_j-l_j)^2}{12}), j = 1...m$. The standard approach for solving the maximum likelihood estimators (MLE) of μ, σ^2, β can then be applied to Equation (2.21). Other reparameterisation of S_m into a function of interval midpoint and *log* range (Brito and Duarte Silva, 2012; Lin et al., 2017) can be adopted. In the paper, the authors (Le-Rademacher and Billard, 2012) also derived MLEs for the case where internal parameters are dependent. Likelihood functions built directly at the symbol level is fine on its own if the interest of the analysis lies solely in inference at the group-level characteristics. However, more often than not, an analyst may be interested in modelling the underlying distribution X_{ij}. Unfortunately, under Le-Rademacher and Billard (2012)'s set up, it is not clear how one can incorporate the knowledge about the distribution of the micro-data into the likelihood function. In an attempt to address this issue, Zhang and Sisson (2016) developed an inferential framework specifically for interval-valued symbolic variables that allows the direct fitting of models for the real-valued micro-data, thereby providing more interpretable results to the data analysts. More importantly, this method provides a natural mechanism that depart from the unrealistic uniformity-within-intervals assumption adopted by a majority of current SDA approaches. This thesis will follow Zhang and Sisson (2016)'s approach to derive a new general construction tool that can be systematically used for building likelihood functions for interval-valued and histogram-valued symbolic variables. Similar to Zhang and Sisson (2016)'s proposal, this approach is able to capture both intra- and inter symbol variations. It does not require a restrictive uniformity within symbols assumption instead it allows for any distributions suggested by the underlying data themselves.

2.5.2 Methods For the Analysis of Histogram-Valued Symbolic Data

Compared to interval-valued symbolic variables, histogram-valued symbolic variables admit more information about the underlying classical data. In spite of their ability in capturing information about the underlying data, the development for histogram-valued symbolic variables is scarce relative to their interval counterparts. Exploratory analysis such as PCA for histogram-valued symbolic variables has been studied by Rodrıguez et al. (2000); Rodrıguez and Pacheco (2004); Ichino (2011) and Le-Rademacher and Billard (2013). It is worth noting that the values of a histogram variable may equivalently be represented by an empirical distribution function or a quantile function and from which analyses can then be performed. Quantile functions, being the inverse of cdf's, are always defined in $[0, 1]$. The dissimilarity between two histograms can be measured through the quantile functions by e.g., the L_2 Wasserstein metric.

Definition 2.5.1. The L_2 Wasserstein-Kantorovich metric, also known as L_2-Mallow's distance, can be used to compare two univariate distributions represented by quantile functions

$$d_w(S_1, S_2) = \sqrt{\int_0^1 (F_1^{-1}(t) - F_2^{-1}(t)dt}. \tag{2.22}$$

Based on minimising this dissimilarity measure, there has been some active development for histogram-valued symbolic variables. Firstly, distance-based clustering for histogram-valued data has been studied by Irpino and Verde (2006); Brito and Chavent (2012). Linear regression based on least squares estimation to minimise the differences between the fitted and the observed histograms (using Equation (2.22)) was proposed by Irpino and Verde (2015). However, using this method, the predicted response would be a valid quantile function (i.e., a monotonic increasing function in [0,1]) if and only if it is a conical combination of quantile functions (i.e., a linear combination with positive coefficients). However, this requirement may fail in the case of multiple linear regression where classical OLS estimates cannot be guaranteed to be all positive. Dias and Brito (2015) addressed this problem by introducing the symmetric quantile distributions in the regression model as new predictor variables. However, these new variables do not have intuitive interpretations. In the time series setting, based on the dissimilarly measure defined in Equation (2.22), Arroyo and Maté (2009) proposed the K-Nearest Neighbours (K-NN) algorithm to forecast time series data aggregated as histograms.

2.5.3 Methods For the Analysis of Other types of Symbolic Variables

Though the majority of the work in SDA has been done for interval-valued or histogram-valued symbolic variables, there has been some progress for categorical multi-valued variables illustrated in Section 2.4.2. Lauro et al. (2008) proposed a generalised canonical analysis to study the relationships between symbolic object descriptors and symbolic objects on a factor plan. Within the framework of conceptual clustering, a "symbolic" hierarchical or pyramidal method was proposed by Brito (1994) and Brito (1995) to cluster multi-valued data of different types, which was further developed to allow for clustering multi-valued modal data (Brito and Polaillon, 2012).

2.6 Conclusion

In this Chapter, we first demonstrated how symbolic data can emerge from real-valued micro-data and we demonstrated that this new kind of data includes classical data as a special case, which occurs when internal variation shrinks to zero. As a result, one should expect that methods in SDA should reduce to standard methods in classical data analysis. Given that most SDA regression models are built on either lower and upper bounds or midpoints and ranges (or log(range)), the convergence to classical data approaches will not hold in the limit as the symbol approaches classical data. Further, the likelihood-based model proposed by Le-Rademacher and Billard (2011) only permits inferential statements to be made at the level of the real-valued random vector which summarises a symbolic-valued random variable. Therefore, using their method, it is awkward to incorporate distributional information regarding the micro-data into the likelihood function. The root of the above problem lies in the fact that current SDA methods are proposed directly at the symbol level and fail to account for the underlying distribution from which symbols are obtained. Another construction flaw inherent in current SDA approaches is the assumption of uniformity within symbols (intervals or within histogram bins), which significantly limits the generalisation of current SDA approaches. To sustain the development in SDA, it is important to address the above issues endemic in the SDA literature and methodology.

The content of this thesis is organised as follows: In Chapter 3, we used the symbolic model proposed by Brito and Duarte Silva (2012) combined with a Bayesian hierarchical model to improve and address a challenging issue present in estimating the global species richness. In Chapter 4, we introduce a new method of model fitting for interval-valued and histogram-valued symbolic data based on fitting to the underlying data rather than fitting to summary statistics of symbols. We also introduce several new methods of symbol construction.In Chapter 5, we apply the proposed symbolic likelihood function to histogram-valued particle number concentration data to estimate their dynamic evolution across time. This thesis concludes with a discussion in Chapter 6.

Chapter 3

Estimating global species richness using symbolic data meta-analysis

3.1 Introduction

Knowing the total number of species on Earth – or at least having a good approximation – has been a key but elusive goal for ecologists for many decades (Caley et al. 2014). Without an adequate baseline against which future changes in biodiversity can be compared, it will not be possible to know with any certainty what and how much biodiversity has been lost anytime in the future, the success or failure of conservation and recovery efforts, or indeed, even whether all species can be named before they become extinct (Costello et al. 2013). It is also well accepted that global biodiversity is threatened by myriad agents (Hunter 2007, Pimm et al. 2014), and that its accelerating loss (Ceballos et al. 2015) will be associated with degraded ecosystem services (Benayas et al. 2009, Mooney et al. 2009). Despite this urgency to understand the world's biological resources and what could be lost, little progress has been made in achieving logically consistent estimates across taxa and ecological realms and estimates that converge through time both in terms of the most likely estimates and increasing confidence around these estimates (Caley et al. 2014). Recently, however, some evidence has begun to emerge that at least for terrestrial arthropods, and some subtaxa thereof, global species richness estimates have begun to narrow (Stork et al. 2015). Nonetheless, while this narrowing of the ranges of these estimates is encouraging, substantial further narrowing is still needed; species richness estimates still vary by more than 1 million to many millions of species depending on the group being considered (Stork et al. 2015). Moreover, this quest for better estimates of species richness has been dominated by the sequential development of estimation methods each considered and applied in isolation. By considering these estimates in isolation from each other, information is inevitably lost that, if combined, might be able to be harnessed to achieve better estimates. For example, where a species richness estimate for a realm (e.g. all marine species: May 1992b) is lower than an estimate for a marine habitat (e.g. coral reefs: Fisher et al. 2015) a source of uncertainty and a logical inconsistency has been clearly identified. Thus, the information inherent in these conflicting estimates may be able to be leveraged in aid of a better estimate, if such information from separate studies could be combined in meaningful ways. Here, we incorporate information across studies that estimate global species richness or components thereof. We do so by developing a Bayesian hierarchical model to estimate species richnesses that are logically consistent across all species. Our approach not only enforces this logical consistency of estimates across the hierarchy, but it also improves estimation accuracy by sharing of information across taxa within the hierarchy.

For our analysis, we used previously published estimates of the total number of species worldwide, as well as estimates of the number of species from various subcategories, including coral reefs, the marine environment, beetles, insects and terrestrial arthropods (Table 1). Although there is a wealth of published species richness estimates for various taxa at regional and subregional scales, we restrict our analyses here to global estimates. We do this because without robust estimates of the spatial turnover of species (beta - diversity) between sufficient locations to provide a reasonable global estimate of beta diversity for each taxon considered, and which are not currently available, there is the danger of counting the same species multiple times, and/or under counting others that are regionally endemic and thereby introducing otherwise avoidable biases in global estimates (May 1992a). The global estimates we used are recorded in three forms. Some estimates (17) are point-estimates, x. These constitute an experts' most informed and best estimate of the number of species based on a detailed analysis or study. These point estimates are presented in the literature with no estimated upper or lower uncertainty bounds. Some estimates (19) are recorded as intervals (a, b) only, representing a plausible range of species counts. These range estimates are presented with no estimate of central tendency, the most likely value. The remaining estimates (9) provide both a point estimate and an interval, which estimates the upper and lower uncertainty bounds. While representing the most complete and information rich estimates, these estimates introduce further complexity because these uncertainty bands can be symmetric or asymmetric around the best estimate.

One approach to combining these data into a joint analysis is to convert the estimates of intervals to a single value, e.g. the midpoint of the range $x = (a + b)/2$, which would then permit a simple analysis on these assumed means combined with other point estimates. However, taking these midpoints to represent the best estimate of central tendency assumes that the uncertainty around this midpoint is symmetrical, and analysis of these single points ignores the often very considerable uncertainty contained in the full interval suggesting the high probability that the midpoint may not be the true value. Caley et al. (2014) followed this approach, and also fitted independent models to the upper and lower interval endpoints to search for evidence that species richness estimates have converged over time. However, assuming independence between these three models could produce self-inconsistent results whereby the lower interval boundary could exceed the upper boundary, and as before with logical inconsistencies of estimates within the hierarchy, information may be lost. As an alternative, we model these data, both single-values and intervals, as part of a unified analysis using techniques from symbolic data analysis (see e.g. Billard and Diday 2006 for a comprehensive review). Symbolic data analysis provides a principled way of performing an analysis for data where the 'datapoints' themselves are distributions, rather than point values. Here, our diversity interval estimates (a, b) can be considered as a distribution defined between a and b, and we can then proceed to coherently analyse them using symbolic techniques. In this setting, for convenience, the univariate interval $(a, b) \subseteq \mathbb{R}$ (where $\mathbb{R}$ is a set of real numbers) commonly mapped to the bivariate random vector $(m, \log r)^\top \in \mathbb{R}^2$, where $m = (a + b)/2$ represents the interval mid-point, and $r = (b - a)$ is the interval range (Brito and Duarte Silva 2012). Performing a statistical analysis on this bivariate vector is equivalent to analysing the univariate random interval (a, b) (Zhang and Sisson 2016). Translating the intervals to midpoint and log range makes no assumption that the distribution within each interval is symmetric – rather it is just a convenient transformation. Single valued estimates x can be also directly expressed in this bivariate vector form, where the midpoint is given by the expert's best guess $m = x$, and the range r, which describes the uncertainty on the estimate, is simply unobserved, so that $(m, \log r)^\top = (x, \mathrm{NA})^\top$. Given this reparameterisation into a common data format, these bivariate random vectors can then be modelled via a Bayesian hierarchical model (e.g. Gelman et al. 2013, chapter 5) that retains information from all types of estimates available: point, and range with and without midpoint.

This model also enforces logically consistent hierarchical relationships among different species categories. For example, the number of insects plus the number of other arthropods must sum to the total number of arthropods. It can also produce estimates of species categories that have not yet been observed or recorded, and can provide estimates of the missing ranges r for the point-estimate only data, thereby allowing the unobserved full interval to be predicted from the model's posterior distribution. It additionally permits determination of the effects on model species richness estimates when including a new measurement in a particular category, thereby facilitating assessments of where best to focus future estimation efforts in order to most efficiently improve species richness estimates. This approach also provides a way to update global estimates as new estimates within the hierarchy become available.

3.2 Methods

The data analysed included 45 previous estimates of global species richness obtained from the literature (Table 1), of which 42 were previously analysed by Caley et al. (2014). Each paper provides a species estimate in one or more of 7 categories: coral reefs (4 estimates), marine species (8), terrestrial species (1), arthropods (10), insects (12), beetles (1) and total global species estimates (9). These estimates are published in interval form (a, b) with no point estimate (19 cases), point estimate form x with no interval estimate (17 cases) and combined interval and point estimate form (a, x, b) (9 cases). In the latter case, the best guess point estimate x exactly coincides with the midpoint of the interval in 4 cases (so that $x = (a + b)/2$), but is different in the remaining 5 cases. As described in the Introduction, we re-parametrise each species richness estimate into the form of a bivariate vector $(m, \log r)^\top$ describing interval midpoint and log range. That is, where interval estimates (a, b) or (a, x, b) are available, these are expressed as $m = (a + b)/2$ and

$r = (b - a)$, and where only a point estimate x is available, then $m = x$, and $\log r$=NA is a missing value.

Note that in the 5 cases where a point estimate and interval are both available, but where the point estimate is not the midpoint of the interval, we ignore the point estimate, and still express the m as the midpoint of (a, b). While this potentially loses information about the possible asymmetric nature of the interval estimates in these 5 cases, we do this for two reasons. First, modelling the midpoints as $m = (a + b)/2$ makes no assumptions about the distribution of the expert's estimate within the interval. This could be symmetric or asymmetric. Rather, it simply states that the midpoint of the interval itself is $m = (a+b)/2$. So treating the estimates (a, x, b) in this way means that the midpoint m has the same interpretation as the interval-only estimates (a, b). By implication, this means that we are stating that for point estimates $m = x$ only, that the point estimate is the centre of the unobserved interval, which then becomes a model assumption. Second, there are very few (5) abundance estimates with a point estimate that is different from the interval mid-point. While it is possible to construct an asymmetric model for a 3-dimensional reparameterisation of the trivariate vector $(a, x, b)^\top$ (e.g. Le-Rademacher and Billard 2011 propose a way of modelling asymmetric interval-valued data assuming a triangular distribution), this would result in a large number of missing values for the rest of the dataset, far more than could be handled with confidence. It is therefore more realistic and conservative for the present analysis to restrict our analyses to bivariate modelling. Trivariate modelling will be more informative in the future as more $(a, x, b)^\top$ format data become available. In the meantime, the results of the bivariate modelling approach adopted here are unlikely to be adversely affected by this small loss of information. For observed data within a given species category $j \in \mathcal{C}$ with $\mathcal{C}$ = {global, other-global, marine, other-marine, arthropods, other-arthropods, coral-reefs, insects, other-insects, beetles} (Fig. 3.1), we can then model the derived $(m, \log r)^\top$ via an appropriate statistical model. A positive association between m and $\log r$ is expected given that these data are species counts and the variability of count data increases with the number being counted. Visually, a bivariate Gaussian distribution could credibly represent these data well, albeit with different location and scale parameters for each category. That is, for each species category j, with observed data $(m_{ij}, \log r_{ij})^\top$ for $i = 1, \ldots, n_j$, we suppose that

$$(m_{ij}, \log r_{ij})^\top \sim N_2(\mu_j, \Sigma_j) \tag{3.1}$$

where

$$\mu_j = (\mu_{mj}, \mu_{rj})^\top \qquad \text{and} \qquad \Sigma_j = \begin{bmatrix} \sigma_{mj}^2 & \rho\sigma_{mj}\sigma_{rj} \\ \rho\sigma_{mj}\sigma_{rj} & \sigma_{rj}^2 \end{bmatrix}.$$

σ_{mj} and σ_{rj} are standard deviation corresponds to the midpoint m_{ij} and log-range $\log r_{ij}$ respectively. Note that we specify the correlation ρ between m_{ij} and $\log r_{ij}$ to be the same across all species categories j. This decision is based on visual inspection of Figure 3.1 (with allowance for small sample sizes), in which the linear dependence appears similar across all categories, and the not unreasonable assumption that any species counting process is similar for all categories. The advantage of making this assumption is that the

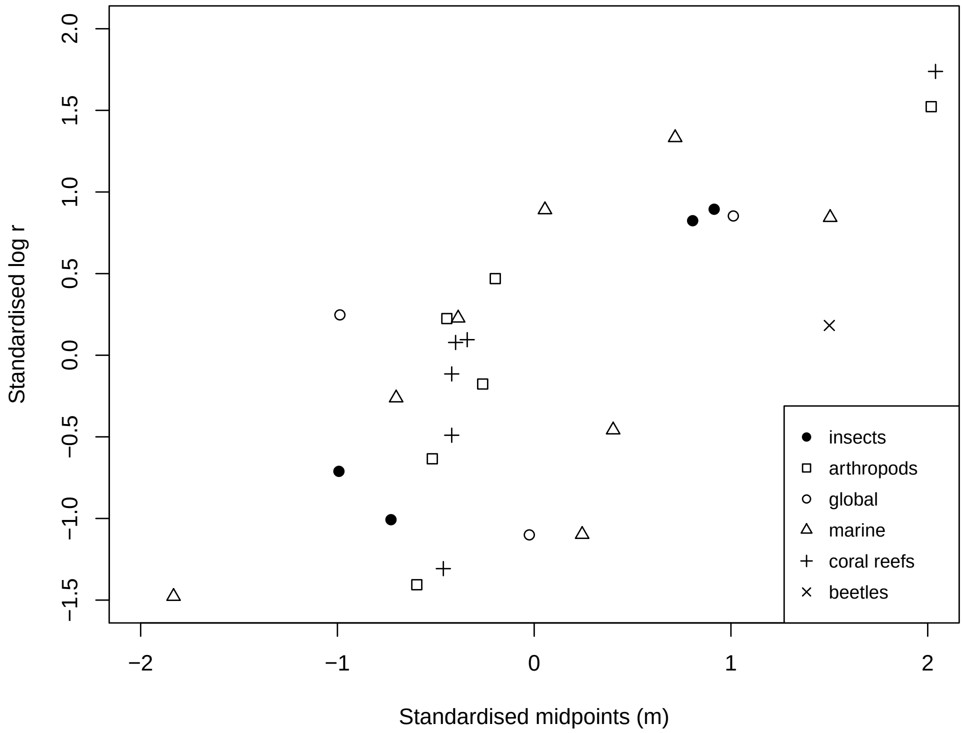

Figure 3.1 – Scatterplot of standardised midpoints (m) versus standardised log range ($\log r$) for the 28 observed intervals. Point types indicate species category.

correlation ρ can be estimated using the observed data from all categories, and thereby, provides a way for the model to share information between categories. This can be particularly useful when estimating missing range values ($\log r$) for single point estimates $(m, \mathrm{NA})^\top$. The above model is appropriate for each species category in isolation, however, there is also a hierarchical relationship between the different categories that is important to account for. Our observed data consist of global species estimates for coral reefs, marine species, insects and arthropods, and these are naturally related hierarchically (Figure 3.2). (Note that in Table 1 there is a single observation for terrestrial category. However, this datapoint is largely inconsistent with other observations in the insects and arthropods categories, and so it was removed as an observation, and as a species category, from our analysis.) This model indicates that, for example, the number of arthropod species comprises the number of insect species plus the number of non-insect arthropod species.

For some species categories we have observed data and for others we have none (Figure 3.2, white and grey boxes, respectively). In the absence of observed data, the model parameters for these categories can be inferred from the hierarchical structure of these data because observations at any level in the hierarchy provide information about values that are possible elsewhere in the hierarchy. For example, the number of arthropod species does not solely consist of the number of insect species. It must also contain a number of other arthropods that have not been estimated individually, or even discovered yet (Fisher et al 2015). Each "other" species category is structurally modelled in the same way as the other categories via equation (3.1), except that there are no observed data. The unknown

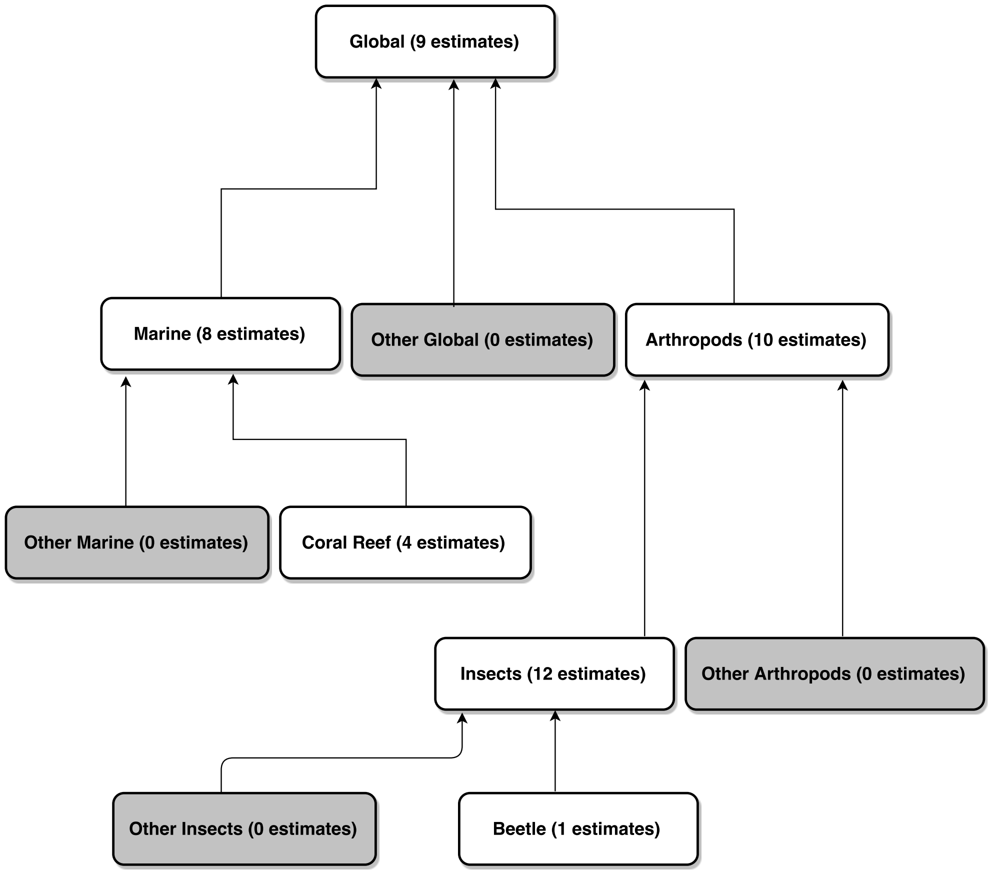

Figure 3.2 – Schematic of the hierarchical structure of species categories analysed. White boxes correspond to categories with observed data. Grey boxes correspond to assumed "other species" categories not observed.

parameters are therefore determined by the mismatch in estimated parameters from its neighbouring categories both sideways and up and down in the hierarchy. For example, the parameters of the other-arthropods category may be inferred from the difference between the number of insect species and the number of arthropod species. The model (3.1) additionally assumes that each data point $(m, \log r)^\top$ is an unbiased estimate of the true species count midpoint (μ_{mj}) and log range (μ_{rj}), and that the interval estimates are exchangeable within each species category (which seems to be supported by Caley et al. 2014). If this assumption holds then the hierarchical structure will allow us to obtain unbiased parameter estimates for the "other" species categories.

For example, the hierarchical relationship between the species categories in Figure 3.2 means that e.g. the sum of the number of beetle and other-insect species should equal the number of insect species, while the sum of the number of insect and other-arthropod species should equal the number of arthropod species, etc.

To enforce this hierarchical structure, we assign the following constraints on the midpoint mean parameter μ_m:

$$\begin{aligned}
\mu_{m_{insects}} &= \mu_{m_{beetles}} + \mu_{m_{other-insects}} \\
\mu_{m_{arthropods}} &= \mu_{m_{insects}} + \mu_{m_{other-arthropods}} \\
\mu_{m_{marine}} &= \mu_{m_{coralreefs}} + \mu_{m_{other-marine}} \\
\mu_{m_{global}} &= \mu_{m_{marine}} + \mu_{m_{arthropods}} + \mu_{m_{other-global}}.
\end{aligned}$$

Further, it is reasonable to suppose that a similar hierarchical structure also holds for the interval log ranges. By noting that if $X \sim (a_X, b_X)$ and $Y \sim (a_Y, b_Y)$ then $X + Y \sim (a_X + a_Y, b_X + b_Y)$ (where $Z \sim (a_Z, b_Z)$ denotes that the random variable Z is distributed between a_Z and b_Z), then, for example, given that the number of insects must equal the number of beetles plus the number of other insects, then in terms of intervals (a, b) we must have

$$\begin{aligned}
a_{insects} &= a_{beetles} + a_{other-insects} \\
b_{insects} &= b_{beetles} + b_{other-insects}.
\end{aligned}$$

This then implies the constraint

$$\mu_{r_{insects}} = \log[2(\mu_{m_{insects}} - \mu_{m_{beetle}} - \mu_{m_{other-insects}}) + \exp(\mu_{r_{beetle}}) + \exp(\mu_{r_{other-insects}})]$$

on the mean log range parameter for insects, $\mu_{r_{insects}}$. Equivalent constraints on μ_r also hold for the arthropods and marine categories. The global category, which is the sum of marine, arthropod and other global categories, is similarly constrained

$$\begin{aligned}
\mu_{r_{global}} = \; & \log[2(\mu_{m_{global}} - \mu_{m_{arthropods}} - \mu_{m_{marine}} - \mu_{m_{other-global}}) \\
& + \exp(\mu_{r_{arthropods}}) + \exp(\mu_{r_{marine}}) + \exp(\mu_{r_{other-global}})].
\end{aligned}$$

In combination, the above constraints mean that the parameters μ_j and Σ_j for the species categories with “children" categories in Figure 3.2 are fully determined by the parameters of their children categories. That is, the only categories with free parameters are beetles, coral reefs and the four “other" species categories, and once these parameters are known, the parameters for the rest of the hierarchical model become fixed. However, estimation of these parameters accordingly means choosing these parameters so that the observed data could credibly have been observed over the entire hierarchy, so that e.g. observed data in the global category will directly influence the parameter estimates of all other species categories. In this way, our model allows the sharing of information in estimating the number of species in one category given the data in all other categories.

Our model is analysed under the Bayesian framework. We adopted the following prior specification for the parameters of the non-fixed “children” categories: $\mu_m \sim N(0, 10000)I(\mu_m > 0)$ represents a diffuse prior constrained to be a positive real number, so that the interval midpoint can be located effectively anywhere above zero. $\mu_r \sim N(-1, 1.5)I(\mu_r < \log(2\mu_m))$ puts most prior weight on smaller ranges, where the constraint ensures that the lower bound of the resulting interval (which depends on both midpoint and range) is also always greater than zero. The standard deviation parameters are specified as $\sigma_m, \sigma_r \sim$ Half-Cauchy$(0, 2.5)$ (that is, a Cauchy(0,2.5) distribution constrained to the positive real line), which is a reasonable default choice for a scale parameter in the absence of specific information (Gelman et al. 2006). Writing P as the correlation matrix associated with each Σ_j, we specify $P \sim LKJ(1)$ so that the prior for P is uniform over all correlation matrices. Markov chain Monte Carlo simulation from the resulting posterior distribution was implemented in the Stan software package (Carpenter et al. 2015).

3.3 Results

3.3.1 Overall Species Estimates

Figure 3.3 illustrates the observed data and resulting posterior interval summaries from the fitted hierarchical model, plotted on the log scale, for each species category, with different species categories shown by different colours. Table 2 enumerates some of these posterior quantities.

The posterior distributions of model parameters incorporate the hierarchical constraints as discussed above (Figure 3.3). This means that, for example, the number of beetles (orange lines) plus the number of other insects (red) can be roughly seen to sum to the number of insects (magenta) both as a midpoint (filled circles) and as an interval (note the log scale). Secondly, within each species category the posterior mean intervals (thick lines) and midpoints (filled circles) are mostly consistent with the observed data in each category. This is most clearly seen for insects (magenta), where the length of the posterior mean interval is roughly the average of the observed insect interval lengths, and the posterior midpoint mean is located roughly at the centre of all the observed point estimates (open circles) and observed interval midpoints. However, this is less apparent for other categories. For example, while the posterior mean midpoint for coral reefs (purple) appears to be well located at the centre of the 4 observed intervals, the posterior mean interval range is considerably shorter than the ranges of the observed data. This is not an error, but is actually a direct and beneficial outcome of the hierarchical model. By combining the information from four separate estimates and analysing these within this hierarchical framework, a central point can be estimated with a much narrower range.

If the 4 coral reef interval observations are analysed in isolation using the model (3.1) but without taking into consideration the hierarchical structure surrounding these estimates, the resulting posterior mean interval ranges are more consistent with the ranges of the observed intervals,as seen in Figure 3.4 (leftmost thick line). As this simplified (non-hierarchical) model analysis appears to be performing correctly, the difference in outcomes with the full (hierarchical) model for global species richness (rightmost thick line, Figure 3.4) compared to the separate estimates, must then be due to the imposition of the hierarchical structure in the analysis. This structure enforces that the interval and midpoint estimates in any category are not just informed by the estimates for that category, but also by the estimates in other categories given their direct and known relationships. In the case of coral reefs, a narrow posterior mean interval, as opposed to the quite wide independent interval estimates, is consistent with the richness estimates that are estimated from the data in other categories, most immediately in the marine and other-marine categories. As a result and given the extra information borrowed from the other categories, the model is able to drastically revise its certainty regarding credible interval ranges for coral reefs based on the data in these related species categories. This would not have been possible without the hierarchical model.

In general, this hierarchical analytical approach is sensible, and preferable to non-hierarchical ones, both because it allows for a pooling of information over categories which can lead to more precise within-category estimates, and because it can enforce parameter estimation that satisfies known model constraints, such as species richness consistently increases upwards through the taxonomic hierarchy, and thereby automatically generate estimates consistent with these constraints. An additional benefit of building a symbolic hierarchical model (that is, in this case one that combines interval and point estimate data) is that the known positive relationship between interval midpoint and range (Figure 3.1) also permits a sharing of information between these two quantities. This means that more informed estimates of (say) midpoints are obtained than if a hierarchical model was to be constructed on midpoints alone, which is the standard hierarchical model format. This benefit is in addition to the precision gained by incorporating both interval and point estimate data into the analysis in the first place.

One caveat regarding these benefits is that we are assuming that every interval estimate (a, b) or point estimate x is independent and an unbiased estimate of the true quantity for the given category. If this is not the case, then errant observations will not only affect the parameter estimates for their own species category, but they will also influence those in other categories. Accordingly, there needs to be a strong emphasis on ensuring quality and consistency of the data analysed in this way, especially as new global richness estimates become available and the results of this model updated. As an example of this, note that we excluded the one "terrestrial" observed point estimate (Table 1; May (1992b)) as it was inconsistent with the hierarchical structure, and to include it would likely have negatively affected the remaining category parameter estimates.

Finally, note that in these analyses we have only presented the posterior mean interval or posterior mean midpoint as point estimates of these quantities. However, there are in fact full joint posterior distributions associated with them. For example, Table 2 presents both the posterior mean and 95% highest posterior density (HPD) intervals for the interval midpoints. Note that in many cases these HPD intervals also encompass the mean lower and upper bounds of the associated interval estimates. This is not inconsistent – the joint posterior distribution for these quantities enforces the constraint $\mu_{aj} \leq \mu_{mj} \leq \mu_{bj}$ absolutely. However, just presenting the posterior marginal mean of each of these parameters hides the fact that there is some uncertainty associated with each of these parameters, beyond the mean values presented here. For example, the two dashed intervals in Figure 3.4 illustrate 95% HPD intervals for the upper and lower interval estimates for coral reefs. The posterior means of these intervals are then used to construct the posterior mean interval estimates, as indicated by the horizontal dashed lines.

3.3.2 Estimates over time

The final aspect we examine is how the parameters of the hierarchical model evolve as more data are included in the analysis over time as new estimates of global species richness become available, either for some taxonomic subset or for all species. We study this to evaluate how the nature of individual diversity estimates have changed over time, and also to show how the addition of data for one species category under the model affects the species richness estimates for the same or other categories. That there should be some local or global effect is clear due to the nature of the hierarchical model. For example, we might suspect that observing data within one category will have the largest impact on parameter estimates in that category, but there may also be a smaller effect on parameter estimates in other categories through the effect of information sharing through the hierarchical relationships. In principle, understanding the nature of these changes should be informative in deciding where best to focus efforts in obtaining future data to expedite progress towards agreed and more precise estimates of global species richness within or among any species category.

In the following, we arrange our observed data according to the year the interval or point estimate was published, and construct four (nested) datasets consisting of the data published in the years 1952–1991 (9 observations), 1952–98 (19 observations), 1952–2007 (31 observations) and 1952–2015 (all 44 observations). These year ranges were chosen to include a roughly equal number of new diversity estimates in each successive time period. We fit our hierarchical symbolic model to the data in each dataset, and observe how the parameter estimates evolve over time as more data are included in the analysis in subsequent periods. The results are summarised in Figure 3.5.

For arthropods, as more data are observed over time, the location and variability of the interval midpoint are reduced substantially. This reduction in variability is expected in the presence of a (relatively) large number of observed datapoints for arthropods (10), but also occurs due to the large amount of information in the neighbouring parent global category (9) and the child category insects (12) (see Figure 3.2), making arthropods one of the most well informed species categories in the hierarchy. The reduction in the midpoint estimate over time can be primarily attributed to changes from very large early published estimates of arthropod diversity (specifically, $x = 30$ (million) from Erwin (1982), and $(a, b) = (10, 80)$ (million) from Stork (1988)), to a consistently lower sequence of eight later published estimates with a mean observed midpoint of ≈ 6.82 (million) (Table 1). Secondary influences on the midpoint estimate are due to the need for consistency with the rest of the hierarchical model.

Arthropod interval range estimates also decrease over time, but less dramatically than for the midpoint. In part, this discrepancy between the effects of mid points and ranges is a result of there being less observed data for ranges than for midpoints for the relevant categories (Figure 3.3: 4/10 arthropod, 3/9 global and 4/12 insect diversity estimates have unobserved ranges). The other-arthropods category is wholly determined by the insects

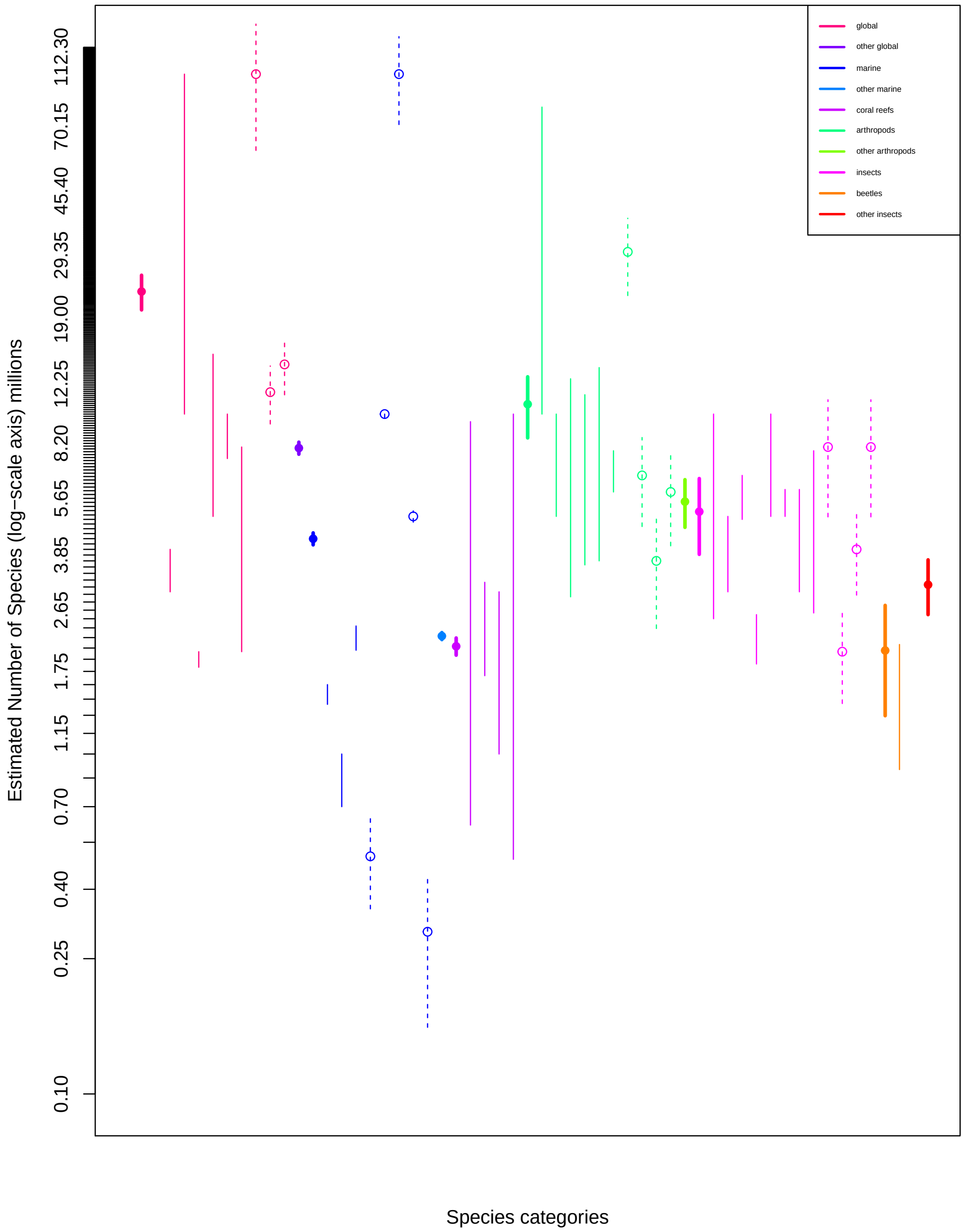

Figure 3.3 – Observed data and posterior interval estimates for each species category (as indicated by colour). Open circles and thin lines illustrate observed point (x) and interval (a, b) data. Thick lines indicate posterior means of interval for each category, obtained by inverting the mapping $(m, \log r)^\top \to (a, b)^\top$ back to the (a, b) parameterisation for the parameters $(\mu_{mj}, \mu_{rj})^\top$ of each category. (I.e. we transform the posterior for $(\mu_{mj}, \mu_{jr})^\top$ to the posterior for $(\mu_{aj}, \mu_{bj})^\top$ where $\mu_{aj} = \mu_{mj} - \exp(\mu_{rj})/2$ and $\mu_{bj} = \mu_{mj} + \exp(\mu_{rj})/2$). The illustrated interval is that obtained from the posterior mean of the lower (μ_{aj}) and upper (μ_{bj}) endpoints of this interval. The filled circle indicates the posterior mean of the interval midpoint (μ_{mj}). Dashed lines indicate posterior predicted posterior mean of interval where only point data x is observed.

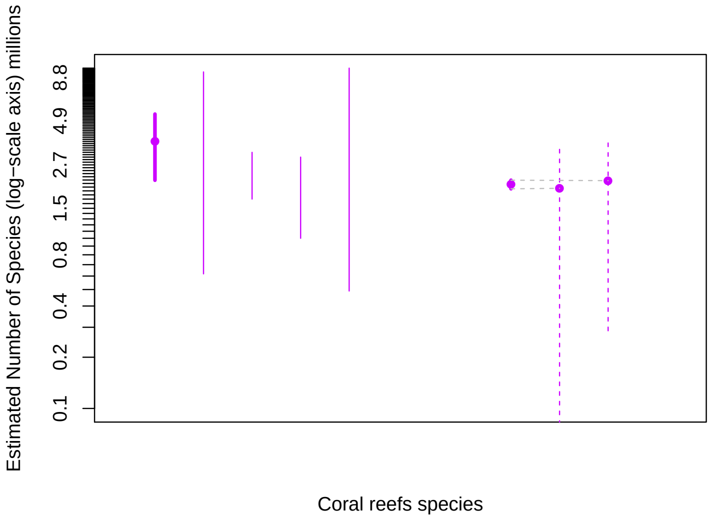

Figure 3.4 – As for Figure 3.3, except for a separate analysis of coral reefs with no hierarchical structure (solid lines). The leftmost thick line illustrates the resulting (non-hierarchical) posterior mean interval, whereas the rightmost thick line illustrates the same interval using the full hierarchical model. The two dotted intervals represent 95% HPD intervals of the lower and upper interval endpoints, based on the full hierarchical model, illustrating considerable uncertainty. (Filled circles represent posterior means.)

and arthropods categories. As these are both well estimated in the presence of large numbers of observed data, both midpoint and range of the other-arthropods category are particularly well informed, and naturally follow the information within arthropods, despite there being no direct observations in this category. This naturally provides a precise and hierarchically consistent estimate of the likely interval for the diversity of all unobserved arthropods. In contrast, a less data-rich section of the hierarchy involves the beetles category (1 observation) along with the parent insects (12 observations) and the unobserved other-insects category (see Figure 3.2). Here, although the estimates for insects become more precise over time, because there is only one observation for beetles in the last time point, there is nothing to distinguish between the beetles and other-insects categories before this datapoint is observed. Hence, for the first three time periods the midpoint and range estimates for beetles and other-insects are highly similar, with their precise values each determined as "half" the estimates for insects. Only when the single beetle estimate is included in the final time point can some difference be discerned in the beetle midpoints, although with such a small amount of data this still makes differentiating between beetles and other-insects difficult. More direct observations in the beetles category would help resolve this lack of distinction between beetles and other-insects. Similarly, other such observations as the hierarchy is populated by further estimates, other opportunities will arise for prioritising effort in making new observations of species richness within categories

As with other species categories, the global species richness estimates clearly get more precise over time as the number of direct estimates in the global category increases, and also as the number of observations at other categories (which determine the global category) also increase.

As with arthropods, the drastic reduction in the location of the global species midpoint is primarily driven by two early and very large global diversity estimates (of $(a, b) = (10, 100)$(million) from Ehrlich and Wilson (1991) and $x = 100$ (million) from May (1992b)), whereas the subsequent six estimates are more consistent and smaller. Part of the explanation for the changes in the nature of these estimates, and those in other categories, could arise from an increase in the sophistication of the methods used to estimate species richness in various categories. Similarly, in time further species estimate fluctuations could arise from the application of genetic data with the potential to split species currently considered to be one and synonymies others.

The bottom left panel of Figure 3.5 illustrates the changes in the global correlation (ρ) between interval midpoint (μ_{mj}) and range (μ_{rj}) across all species categories. Clearly as the number of full interval observations (a, b) increases, the correlation between midpoint and range is better estimated. The full dataset analysis correlation is estimated to be moderately strong with a posterior mean of 0.57 and a 95% HPD interval of (0.25,0.81). Finally, the bottom right panel of Figure 3.5 shows the predicted missing ranges of the observed point estimate $x = 30$ (million) taken from Erwin (1982). When there are little data, despite there being a positive correlation with the observed interval midpoints, the predicted range is strongly influenced by the prior distribution, which places most density on smaller interval ranges. When the arthropod midpoint and range parameters become better estimated, along with their correlation (ρ), the predicted interval associated with the $x = 30$ point estimate becomes more realistic, and more accurately incorporates information from around the hierarchical model. The final mean predicted interval for this datapoint is (22.7,37.3).

3.3.3 Prior sensitivity analysis

A prior sensitivity analysis was performed to test the robustness of the posterior distribution to the choice of the prior. For the prior $\mu_m \sim N(0, \tau)I(\mu_m > 0)$ where we specified $\tau = 10,000$, we alternatively consider also setting τ to 1000, 100, 10 and 1, expression a level of informativeness ranging from the most uninformative ($\tau = 10,000$) to the most informative ($\tau = 1$). Figure 3.6 shows the posterior density for $\mu_{m_{global}}$ under each prior specification (all other priors are held the same). As can be seen, when τ is large, the resulting posterior is invariant to the prior specification. When τ becomes smaller (i.e. $\tau = 10, 1$) and the prior becomes more informative and in conflict with the information in the data, the posterior is more strongly affected. As such, we are satisfied that the choice of $\tau = 10,000$ represents a minimal and appropriate hyper prior choice. Similar conclusions can be drawn when varying the priors of the other μ_m parameters–see Figure A.1–and when changing the value of the normal standard deviation component of the prior for $\mu_r \sim N(-1, 1.5)I(\mu_r < \log(2\mu_m))$ from 1.5 to 2.5 and 5 (Figure A.2 and Figure A.3).

Similarly we perform a sensitivity analysis to assess the robustness of the posterior to the prior for $\sigma_m, \sigma_r \sim$ Half-Cauchy$(0, A)$, where we specified $A = 2.5$ as a convenient weakly informative choice, following (Gelman et al., 2003). We consider the alternative choices of $A = 1.25$ and 5, with larger values of A resulting in a uniform prior density (in

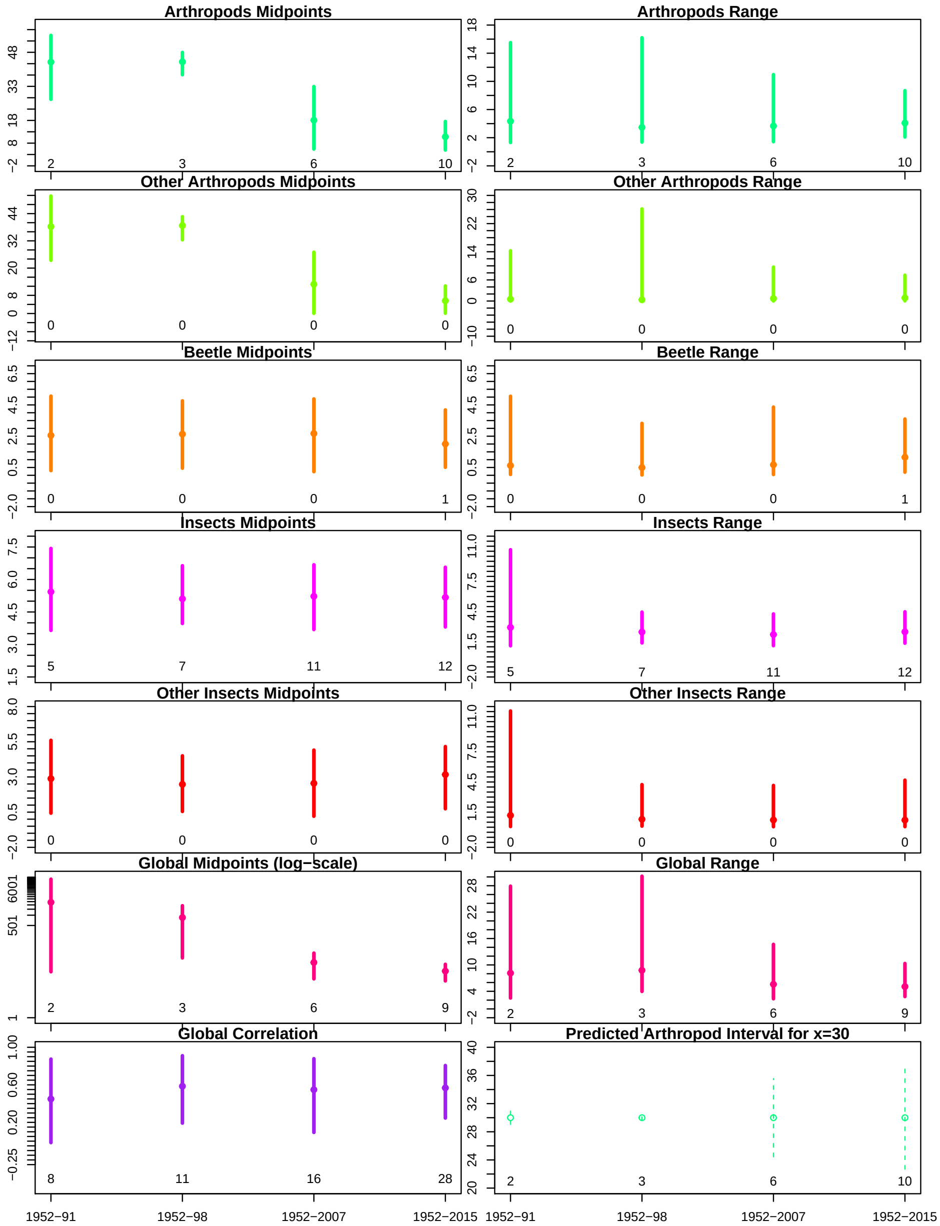

Figure 3.5 – Posterior means (filled circles) and 95% high density credible intervals for interval midpoints (μ_{mj}; left panels) and ranges (μ_{rj}; right panels) measured in millions, estimated from data from four different time periods 1952–1991, 1952–1998, 1952–2007 and 1952–2015. Panels show results for [top to bottom] arthropods, other-arthropods, beetles, insects, other-insects and global, with the number under each graphic indicating the number of directly observed estimates in each category for each time point. The bottom left panel shows the corresponding correlation between all midpoints and ranges (ρ) for each dataset. The bottom right panel illustrates the predictive mean interval for the associated observed arthropod point estimate of $x = 30$ taken from Erwin (1982).

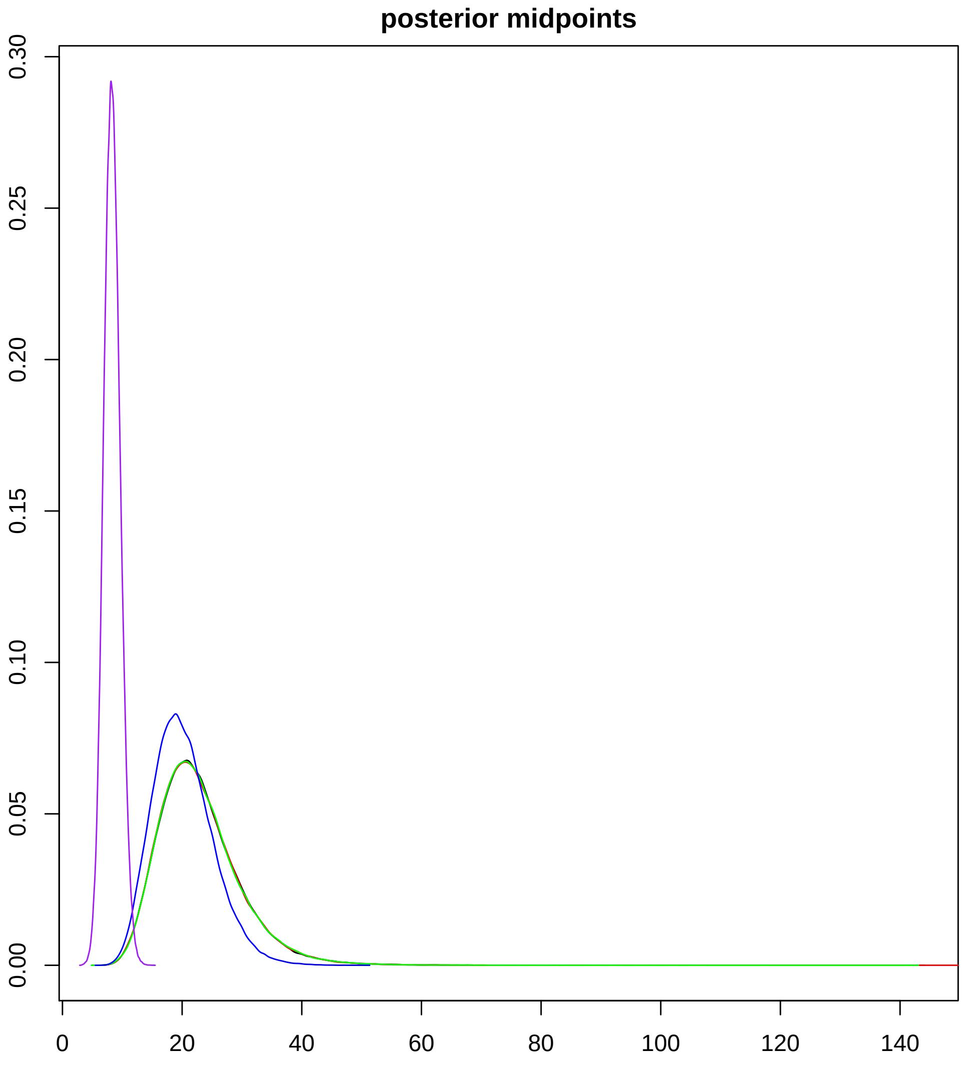

Figure 3.6 – Posterior distribution for $\mu_{m_{global}}$ when fitted with different scale values in the hyperprior distribution $\mu_{m_{global}} \sim N(0, \tau)I(\mu_m > 0)$. The black line represents $\tau = 10,000$, the green line represents $\tau = 1000$, the red line represents $\tau = 100$, the blue line represents $\tau = 10$ while the purple line represents $\tau = 1$.

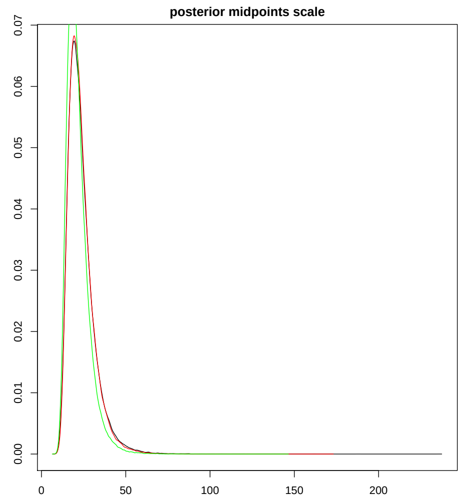

Figure 3.7 – Posterior distribution for $\sigma_{m_{global}}$ fitted with 3 values of scale parameter (A) in the Half-Cauchy distribution. The black line represents $A = 5$, the red line represents $A = 2.5$ and the green line represents $A = 1.25$.

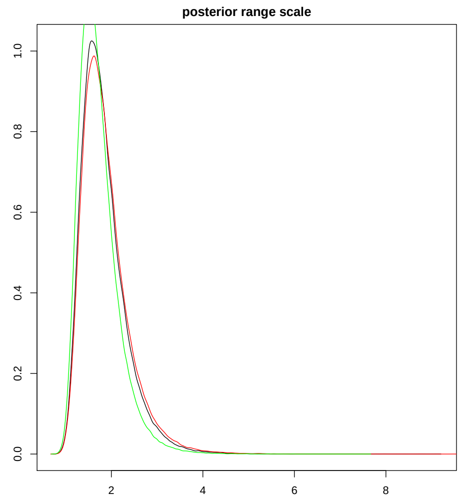

Figure 3.8 – Posterior distribution for $\sigma_{r_{global}}$ fitted with 3 values of scale parameter (A) in the Half-Cauchy distribution. The black line represents $A = 5$, the red line represents $A = 2.5$ and the green line represents $A = 1.25$.

the limit as $A \to \infty$). Figures 3.4 and 3.5 illustrate the posterior distributions of $\sigma_{m_{global}}$ and $\sigma_{rglobal}$ respectively. It is clear that the resulting posterior marginal distributions are largely invariant to the choice of A. Similar outcomes can be observed for the other σ_{mj} and σ_{rj} parameters (Figures A.4 and Figure A.5).

3.4 Discussion

Estimates of species richness are typically, although not exclusively, constructed independently for individual realms or taxa. This means that not only are these estimates potentially inconsistent with estimates for other species categories (Caley et al. 2014), but there is also an estimation inefficiency in that available information for related taxa is not accounted for. Here, we have implemented a meta-analysis that addresses both of these issues, while also presenting a technique for bringing both point estimate and interval estimate data forms into the same analysis. In addition, by adopting a hierarchical model, we have also been able to estimate the number of unobserved species in the 'other' categories, simply as a result of requiring consistency within the hierarchical model. We also provide computer code (see Supporting Information) that enables the addition of new species richness estimates into the analyses presented here so that these global estimates can be updated and improved as new information becomes available and these new results can be used to prioritize future research effort to taxa or regions for which new information would improve estimates the most.

The outputs of this analysis can be evaluated either in terms of predicted intervals (a, b) or interval midpoints $(a + b)/2$, if one is prepared to interpret the interval midpoint as a proxy for a point estimate. Under this interpretation, it is not assumed that the distribution within an interval is symmetric, but merely that the midpoint is a convenient parameterisation of the location of the interval, and it must accordingly be interpreted as such. In terms of providing a single point estimate of species diversity, this interpretation must hold at least until more data on asymmetrical bounds become available to permit asymmetric modelling on $[a, x, b]$, the complete interval plus point estimate. If the asymmetrical bounds around estimates of global species richness reported to date (Table 1) are representative of species categories, the midpoint estimates we provide here are more likely to be over- rather than under-estimates as modal values in asymmetrical distributions so far tend to be located toward the lower end of estimated ranges (but not exclusively) for a range of taxa (e.g. Fisher et al. (2015)). Regardless of the extent to which such a bias over-estimates global species richness, it would appear to be considerably less of a problem than current estimates that violate the necessity of species richness categories being a finite partition of global species richness. That is, correcting logical inconsistencies whereby an estimate for a habitat (e.g. coral reefs) is greater than an estimate for a realm (e.g. all marine species) will likely have a much more profound effect on total global estimates than accounting for the precise locations of modal values with the uncertainty bounds around a point estimate for a taxon. Nonetheless, it is still important to continue to refine our estimation of these uncertainty bounds to better understand the nature of our current uncertainty of these estimates and how to improve them. For example, species richness estimates for taxa on coral reefs with the greatest uncertainty also have the most skewed uncertainties (Fisher et al. (2015)). These greater uncertainties can arise for a variety of reasons including lack of or biased research effort to date, and or lack of characteristics available to distinguish among species (Caley, pers obs). These highly skewed uncertainty bounds, therefore, indicate that species richness estimates might be disproportionately improved by allocating research effort to these taxa.

Each of our global species richness estimates (Table 2) are broadly consistent with previous estimates in the literature, but with some important differences that respect the hierarchical structure of the model, and thereby, provide a set of estimates that are logically consistent with a finite partition of global species richness. For example, in terrestrial arthropods, for which species richness estimates have begun to narrow (Stork et al. 2015), our posterior mean interval of $(8.51, 12.87)$ (million) is wholly above the most recent interval estimate of $(5.9, 7.8)$ (million) by Stork et al. (2015), although our estimate is not inconsistent with the other observed estimates in Table 1. However it is consistent with being the sum of the insects interval estimate of $(3.87, 6.46)$(million) and the other-arthropod estimate of $(4.64, 6.41)$(million). This is not the case for e.g. Stork et al. (2015) who estimate the number of insect species at $(2.6, 7.8)$(million)which has the same upper endpoint as their estimate for arthropods (7.8)(million), implying that there is a positive probability that there are no other arthropod species apart from insects, or that as a proportion the number of these species is too small to be noticed given current levels of uncertainty. Similarly, the global species richness interval estimate of $(7.4, 10)$(million) by Mora et al. (2011) (and indeed, most others in Table 1) does not overlap our global posterior mean interval of $(20.25, 25.61)$(million), determined using currently available species diversity estimates. Again, most reported estimates of global species richness are inconsistent with the estimates in other species categories; they are too low once the hierarchical nature of taxonomy is accounted for.

However, our approach does have some caveats. Primarily, we have assumed that our observed data, published estimates of species richness are independent of each other, and that within any species category the data represent unbiased estimates of the same quantity over time. The first of these assumptions is unlikely to be true as several of the estimates in Table 1 come from the published studies that rely to varying extents on the same data analyses in different ways (e.g. Raven et al. 2000; Novotny et al. 2002; Hamilton et al. 2010; Costello et al. 2011; Stork et al. 2015 and each provide multiple estimates). The second assumption is also unlikely to be true as knowledge has increased since the global species richness estimate of $(3, 4)$ by Raven (1983), definitions of which species are in which category have changed, and the number of actual species has itself changed through the wide spread application of molecular genetic analyses with the power to split cryptic species and synonymies morphospecies. Together this implies that there are potential quality issues with some of these data and the methods used to analyse them that need to be acknowledged. In our opinion, it should no longer be acceptable to simply claim primacy of some method of estimation of global species richness over all others without first ensuring that the method does not violate the finite partition of global species richness within and beyond the taxa of interest, nor without presenting some form of validation of the degree to which the method provides more accurate and precise estimates. Opportunities, however, for such validation will be limited but may in some circumstances be supported by the application of expert knowledge (Fisher et al. (2015)). Superior logic could be argued, for example, if an estimation method incorporates more realistic assumptions and knowledge about how species are distributed in space that were absent from previous methods. Such knowledge of these spatial distributions could also make available information from numerous estimates of species richness at sub-global scales, but doing so will require additional model complexity, associated with increased estimation uncertainty, to accommodate estimates of beta-diversity. Whatever the case, the implications of these choices on the results obtained should be clearly presented. Where such arguments can be successfully mounted, it may be justified to weight the contributions to the likelihood of each observed data point or interval, according to the explicit justification of its reliability. Fully reliable observations would be weighted 1, completely unreliable observations 0, and so would be effectively removed from the analysis, as we implemented here with May's 1992 estimate of the number of terrestrial species. Multiple dependent observations from the same study would receive a weight between these extremes. Ultimately, however, true validation and confidence in these estimates will only be available over the longer-term as the discovery of new species and their taxonomy and systematics proceeds. This progress will provide the opportunity to adaptively learn from testing new estimates against old to assess progress toward convergence as these new estimates propagate up through the hierarchy, and thereby, facilitate the exploration of the consequences of these new estimates on estimates of global species richness.

Chapter 4

New likelihood-based methods for symbolic data analysis

4.1 Introduction

Symbolic data analysis (SDA) is an emerging area of statistics that has immense potential to become a standard inferential technique in the near future (Billard and Diday, 2003a). At its core, it builds on the notion that statistical inferences are commonly required at a group level rather than at an individual level (Billard, 2011; Billard and Diday, 2006). This is the familiar notion behind hierarchical modelling (e.g. Gelman et al., 2013, Chapter 5). For example, the performance of school and higher level units in standardised testing exams is usually of interest rather than the performance of the individual students (Rodrigues et al., 2016; Rubin, 1981).

SDA explicitly embraces this idea by aggregating individual level data (the *micro-data*) into group level distributional summaries (i.e. the *symbols*), and then building models for inference directly at the group-level based on these summaries (Billard, 2011; Billard and Diday, 2006). The most common choice of these summaries is the random interval (or the d-dimensional equivalent, the random rectangle), whereby for individual-level observations $X_1, \dots, X_n \in \mathbb{R}$ the random interval is typically constructed as $S = (\min_i X_i, \max_i X_i) \subseteq \mathbb{R}$. Other common symbol types include random histograms (Dias and Brito, 2015; Le-Rademacher and Billard, 2013) and categorical multi-valued variables (Billard and Diday, 2006). Under the SDA framework, the collection of group-level data summaries $S_1, \dots, S_m \in \mathcal{S}$ are considered the new data "points", whereby each datum is a distribution of some kind with an internal distributional structure. Statistical inference is then performed at the level of the symbols directly, with reference to their distributional forms, and without any further reference to the underlying micro-data. See e.g. Noirhomme-Fraiture and Brito (2011); Billard (2011) and Billard and Diday (2003a) for a comprehensive overview of symbolic data types and their analysis.

This approach is potentially extremely attractive given present technological trends requiring the analysis of increasingly large and complex datasets. Common approaches, such as divide-and-recombine techniques (e.g. Guha et al., 2012; Jordan et al., 2018; Vono et al., 2018; Rendell et al., 2018) or subsampling-based techniques (Quiroz et al., 2018b;

Bardenet et al., 2014; Quiroz et al., 2018a), tend to focus on analysing the full dataset as efficiently as possible by clever use of computing power. In contrast, SDA effectively states that for many analyses this high level of computation is not necessary to make inference at the group level, and by aggregating the micro-data to a much smaller number of group level symbols (where $m \ll n$), 'big data' analyses can be performed cheaply and effectively on low-end computing devices. Beyond data aggregation, distributional-valued observations can arise naturally through the data recording process. This can include observational rounding or truncation, which results in data known to lie within some interval (Heitjan and Rubin, 1991; Vardeman and Lee, 2005), the elicitation of distributions from experts thought to contain quantities of interest (Fisher et al., 2015; Lin et al., 2017), and the construction of particle size distributions of particulate matter, typically in histogram form (Wraith et al., 2014). In this sense, Schweizer (1984)'s often-quoted statement that "distributions are the numbers of the future" seems remarkably prescient.

Many SDA techniques for analysing distributional-valued random variables have been developed, including regression models (Irpino and Verde, 2015; Dias and Brito, 2015; Giordani, 2015), principal component analysis (Kosmelj et al., 2014; Le-Rademacher and Billard, 2013; Ichino, 2011), time series (Lin and González-Rivera, 2016; Wang et al., 2016; Arroyo et al., 2011), clustering (Brito et al., 2015), discriminant analysis (Silva and Brito, 2015) and Bayesian hierarchical modelling (Lin et al., 2017), among others. Likelihood-based inference was introduced by Le-Rademacher and Billard (2011) with further development and application by Brito and Duarte Silva (2012); Zhang and Sisson (2016) and Lin et al. (2017).

However, while there have been many successes in the analysis of symbolic data, from the perspective of the statistical analyst there are several methodological weaknesses in the current SDA framework that prevent SDA methods from realising their potential in the modern statistician's toolkit. One issue is that the large majority of SDA techniques are descriptive and do not permit statistical inference on model parameters. For example, regression models tend to be fitted by symbolic variants of least squares. Le-Rademacher and Billard (2011)'s likelihood-based framework is one clear and welcome exception, however even here specifying credible models can be problematic: while the statistician can readily and intuitively specify a wide range of models for the underlying micro-data, it is not really clear how equivalent or similarly meaningful models can be specified for distribution-valued random variables.

The likelihood approach of Le-Rademacher and Billard (2011) maps each symbol to a random vector that uniquely defines the symbol, and then specifies a standard statistical likelihood model for each of the observed symbols. For example, suppose that $X_{ij} \in \mathbb{R}$ is the value of some process recorded on the i-th second, $i = 1, \ldots, n = 86400$, of the j-th day, $j = 1, \ldots, m$. If interest is in modelling these data as, say, i.i.d draws from a skew-normal distribution $X_{ij} \sim SN(\mu_0, \sigma_0, \alpha_0)$, the likelihood function $L(x|\theta)$, $\theta \in \Theta$, may then be constructed in the usual way. However, suppose that interval symbols are now constructed so that $S_j = (\min_i X_{ij}, \max_i X_{ij}) \subseteq \mathbb{R}$ is the random interval describing the observed range of the process on day j. Due to the equivalence of representing continuous subsets of $\mathbb{R}$ by the associated bivariate vector in this setting (Zhang and Sisson, 2016), the approach of Le-Rademacher and Billard (2011) constructs a model for the vectorised symbols $S_1, \ldots, S_m$, perhaps after a reparameterisation. For example,

$$S_j \sim SN_2(\mu, \Sigma, \alpha) \qquad \text{or} \qquad \tilde{S}_j \sim SN_2(\mu, \Sigma, \alpha),$$

where $\tilde{S}_j = ((a+b)/2, \log(b-a))$ is the typical reparameterisation of $S_j = (a, b)$ into a function of interval mid-point and log range (Brito and Duarte Silva, 2012). While there is inferential value in models of these kind (e.g. Brito and Duarte Silva, 2012; Lin et al., 2017), it is clear that if there is interest in modelling the underlying X_{ij} as skew-normal, it is difficult to construct even a loosely equivalent model at the level of the symbol S_j (or $\tilde{S}_j$). That is, while the statistician may intuitively construct complex statistical models at the level of the micro-data, it is less obvious how to construct models at the symbolic level and for different symbolic forms.

By design, modelling symbols directly, without specifying a probabilistic model for the underlying micro-data, only permits inference and predictions at the symbol level. This is unsatisfactory for two reasons. Firstly, predictive inference for the underlying micro-data is often of interest, even if primary focus is on group-level analyses. Secondly, and as we will demonstrate in Section 4.3.3, ignoring the structure of the micro-data can result in symbolic-level analyses producing the "wrong" inferential outcome.

One clear and acknowledged problem (Kosmelj et al., 2014; Cariou and Billard, 2015) is that existing SDA techniques almost exclusively assume that the distribution of the micro-data within random intervals and rectangles and within histogram bins is uniform. When one considers that random intervals are typically constructed from the X_{ij} by specifying $S_j = (\min_i X_{ij}, \max_i X_{ij})$, it is almost certain that the distribution of the underlying data within S_j is non-uniform. This implies that any inferential procedure built on the uniformity assumption (which is the case with almost all current SDA methods) is highly likely to produce questionable results.

Finally, a major and surprising failing within the current SDA literature is that of symbol design. One principled difference between SDA and regular statistical analyses is that the analysed data are first constructed by the analyst. This raises the question of how this should be undertaken. Intuitively, if the statistician is looking to design, say, a random interval S_j to maximise the information about a location parameter, they would be unlikely to use the sample maximum and minimum to achieve this, as these statistics are highly variable. A more useful alternative could use e.g. sample quantiles to define the interval. While sample quantiles have been considered in SDA methods, they have only been used as a robust method to avoid outliers that would otherwise dominate the size of a random interval (Hron et al., 2017). In general, little consideration has been given to the design of informative symbols.

In this paper we introduce a new general method for constructing likelihood functions for symbolic data based on specifying a standard statistical model $L(x|\theta)$ for the underlying micro-data and then deriving the implied model $L(S|\theta)$ at the symbolic level by considering how S is constructed from x. This means that it is possible to fit the micro-level data model $L(x|\theta)$ while only observing the symbol level data, S. It provides both a natural way of specifying models for symbolic data, while also opening up SDA methods to become a mainstream statistical technique for the fast analysis of large and complex datasets. This approach naturally avoids the incorrect assumptions of within-symbol uniformity, allows inference and predictions at both the micro-data and symbolic data levels, permits symbolic inferences using multivariate symbols (approximately 99% of all symbolic data analyses are based on vectors of univariate symbols), and provides a much higher quality of inference than is available using standard SDA techniques. The method recovers some known models in the statistical literature, as well as introducing several new ones, encapsulates the current symbolic likelihood approach of Le-Rademacher and Billard (2011) as a special case, and reduces to standard likelihood-based inference for the micro data (so that $L(S|\theta) \to L(x|\theta)$) when the symbols are reduced to standard micro-data.

By examining the performance of this new approach we take the opportunity to demonstrate the weaknesses of current symbol construction techniques, and that more powerful inferences can be obtained by redesigning how symbols are constructed. In particular we introduce a new class of univariate and multivariate random rectangles. These new symbol variations produce more efficient analyses than existing symbol constructions, and permit the estimation of within-symbol multivariate dependencies that were not previously estimable (or were only weakly estimable).

The construction of the new symbolic likelihood function is presented in Section 4.2 along with specific results for random intervals (rectangles) and histograms. For clarity of exposition all derivations are relegated to the Appendix. The performance of these models is demonstrated in Section 4.3 through a meta-analysis of univariate histograms, a simulation study of the inferential performance of the new class of multivariate random rectangle constructions, and an analysis of a large loan dataset. In all cases, the existing state-of-the-art models and symbolic constructions are outperformed by the new symbolic model. We conclude with a discussion in Section 4.4.

4.2 A general construction tool for symbolic likelihood functions

We introduce the new symbolic likelihood function in Section 4.2.1 before considering the specific cases of models for random rectangles and random histograms in Sections 4.2.2–4.2.4.

4.2.1 Symbolic likelihood functions

Suppose that Ω is a population of interest defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and that each individual in Ω is described by a measurable random variable X, defined by $X : \Omega \to \mathcal{X}$, and $\mathbb{P}(X \in \mathcal{Y}) = \mathbb{P}(\omega \in \Omega | X(\omega) \in \mathcal{Y})$, for $\mathcal{Y} \subset \mathcal{X}$. We follow the standard SDA construction of a *class* (Billard and Diday, 2003a) and let the random variable $C : \Omega \to \mathcal{C}$ denote the class to which an individual belongs. For simplicity we assume that

$\mathcal{C} = \{1, \ldots, m\}$ is finite. Consequently let $\Omega_c = \{\omega \in \Omega \text{ s.t. } C(\omega) = c\}$ be the set of individuals in the population that belong to class $c \in \mathcal{C}$ such that $\cup_{c\in\mathcal{C}}\Omega_c = \Omega$, and define $X_c : \Omega_c \to \mathcal{X}_c \subseteq \mathcal{X}$ as the random variable that describes them. We denote $\mathrm{Card}(\Omega) = N$ and $\mathrm{Card}(\Omega_c) = N_c$ where $\sum_{c\in\mathcal{C}} N_c = N$.

A symbolic random variable S_c for class $c \in \mathcal{C}$ is the result of the aggregation of the random variables $\boldsymbol{X}_c = (X_{1,c}, \ldots, X_{N_c,c})$ via some aggregation function π_c, so that $S_c = \pi_c(\boldsymbol{X}_c) : [\mathcal{X}_c]^{N_c} \to \mathcal{S}_c$ and $\boldsymbol{x}_c \mapsto \pi(\boldsymbol{x}_c)$. That is, a symbolic random variable represents a summary of the information brought by measurement over individuals. The choice of this summary (and thus of the aggregation function) is critical and we explore this in later Sections. In the following we refer to random variables of the micro-data X as *classical* random variables. By construction symbolic random variables require knowledge of the underlying classical random variables. Accordingly, this should also be true when dealing with likelihood functions, particularly if inference is required at both classical and symbolic levels, but when only information at the symbolic level is observed.

To construct a symbolic likelihood function, suppose that the classical random variable X has probability density and distribution functions $g_X(\cdot;\theta)$ and $G_X(\cdot;\theta)$ respectively, where $\theta \in \Theta$. Consider a random classical data sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ of size $n < N$ from the population, and denote by $\boldsymbol{x}_c = (x_{1,c}, \ldots, x_{n_c,c})$, the collection of those in class c, where $\sum_{c\in\mathcal{C}} n_c = n$. Similarly let $s_c = \pi_c(\boldsymbol{x}_c)$ be the resulting observed symbol obtained through the aggregate function π_c and define the symbolic dataset to be the collection of symbols $\boldsymbol{s} = (s_c; c \in \mathcal{C})$.

Proposition 4.1. *For the subset $\boldsymbol{x}_c$ of $\boldsymbol{x}$ associated with class $c \in \mathcal{C}$, the likelihood function of the corresponding symbolic observation $s_c = \pi_c(\boldsymbol{x}_c)$ is given by*

$$L(s_c; \vartheta, \theta) \propto \int_{\mathcal{X}^n} f_{S_c|\boldsymbol{X}_c=\boldsymbol{z}_c}(s_c; \vartheta) g_{\boldsymbol{X}}(\boldsymbol{z}; \theta)\, d\boldsymbol{z}, \quad \forall c \in \mathcal{C}, \tag{4.1}$$

where $\boldsymbol{z}_c \in \mathcal{X}_c^{n_c}$ is a subset of $\boldsymbol{z} \in \mathcal{X}^n$, $f_{S_c|\boldsymbol{X}_c}(\cdot;\vartheta)$ is the conditional density of S_c given $\boldsymbol{X}_c$ and $g_{\boldsymbol{X}}(\cdot;\theta)$ is the joint density of $\boldsymbol{X}$.

We refer to $L(s_c; \vartheta, \theta)$ given in (4.1) as the symbolic likelihood function. A discrete version of (4.1) is easily constructed. Note that by writing the joint density $g_{\boldsymbol{X}}(\cdot;\theta) = g_{\boldsymbol{X}_c}(\cdot;\theta) g_{\boldsymbol{X}_{-c}|\boldsymbol{X}_c}(\cdot;\theta)$, where $\boldsymbol{X}_{-c} = \boldsymbol{X}\backslash\boldsymbol{X}_c$, then after integration with respect to $\boldsymbol{x}_{-c} = \boldsymbol{x}\backslash\boldsymbol{x}_c$, equation (4.1) becomes

$$L(s_c; \vartheta, \theta) \propto \int_{\mathcal{X}_c^{n_c}} f_{S_c|\boldsymbol{X}_c=\boldsymbol{z}_c}(s_c; \vartheta) g_{\boldsymbol{X}_c}(\boldsymbol{z}_c; \theta) \mathrm{d}\boldsymbol{z}_c.$$

This construction method can easily be interpreted: the probability of observing a symbol s_c is equal to the probability of generating a classical dataset under the classical data model that produces the observed symbol under the aggregation function π_c. That is, we have established a direct link between the user-specified classical likelihood function $L(\boldsymbol{x}|\theta) \propto g_{\boldsymbol{X}}(\boldsymbol{x};\theta)$ and the resulting probabilistic model on the derived symbolic data. As a result we may directly estimate the parameters θ of the underlying classical data model, based only on observing the symbols $\boldsymbol{s}$.

In the case where there is no aggregation of $\boldsymbol{x}_c$ into a symbol, so that $\pi(\boldsymbol{x}_c) = \boldsymbol{x}_c$ and $\mathcal{S}_c = [\mathcal{X}_c]^{N_c}$, then $f_{S_c|\boldsymbol{X}_c=\boldsymbol{z}_c}(s_c) \equiv f_{\pi(\boldsymbol{X}_c)|\boldsymbol{X}_c=\boldsymbol{z}_c}(\pi(\boldsymbol{x}_c)) = f_{\boldsymbol{X}_c|\boldsymbol{X}_c=\boldsymbol{z}_c}(\boldsymbol{x}_c) = \delta_{\boldsymbol{z}_c}(\boldsymbol{x}_c)$, where $\delta_{\boldsymbol{z}_c}(\boldsymbol{x}_c)$ is the Dirac delta function, taking the value 1 if $\boldsymbol{z}_c = \boldsymbol{x}_c$ and 0 otherwise. As a result the symbolic likelihood function reduces to $g_{\boldsymbol{X}_c}(\boldsymbol{x}_c;\theta)$, the classical likelihood contribution of class c. Under the assumption that the classical data are independently distributed between classes, so that $g_{\boldsymbol{X}}(\,\cdot\,;\theta) = \prod_{c\in\mathcal{C}} g_{\boldsymbol{X}_c}(\,\cdot\,;\theta)$, the associated symbols are also independent and the likelihood of the symbolic dataset $\boldsymbol{s}$ is given by

$$L(\boldsymbol{s};\vartheta,\theta) = \prod_{c\in\mathcal{C}} L(s_c;\vartheta,\theta) \propto \prod_{c\in\mathcal{C}} \int_{\mathcal{X}_c^{n_c}} f_{S_c|\boldsymbol{X}_c=\boldsymbol{z}_c}(s_c;\vartheta) g_{\boldsymbol{X}_c}(\boldsymbol{z}_c;\theta)\mathrm{d}\boldsymbol{z}_c.$$

If, further, the observations within a class $c \in \mathcal{C}$ are independent and identically distributed, then in the scenario where $\pi(\boldsymbol{x}_c) = \boldsymbol{x}_c$ we have $L(\theta) = \prod_{i=1}^{n} g_X(x_i;\theta)$. Because $\mathrm{Card}(\mathcal{C}) = m$ and typically $m \ll n$, this implies that large computational savings can be made through the analysis of symbolic rather than classical data, depending on the complexity of the classical data likelihood function. The methodology established in Proposition 4.1 specifies a probability model for the micro-data which, combined with the knowledge about the aggregation process induces a likelihood function at the aggregates level. Comparatively the likelihood function defined by Le-Rademacher and Billard (2011) specifies a distributional assumption directly on the symbols.

In the following Subsections, we establish analytical expressions of the symbolic likelihood function based on various choices of the aggregation function π, which leads to different symbol types. The performance of each of these models will be examined in Section 4.3. For clarity of presentation the class index c is omitted in the remainder of this Section as the results presented are class specific.

4.2.2 Modelling random intervals

The univariate random interval is the most common symbolic form, and is typically constructed as the range of the underlying classical data e.g. $S = (\min_i X_i, \max_i X_i)$. Here we generalise this to the context of order statistics $S = (X_{(l)}, X_{(u)})$ for indices $l \leq u$ given their higher information content. We define an interval-valued symbolic random variable to be constructed by the aggregation function π where

$$S = \pi(\boldsymbol{X}) : \mathbb{R}^N \to \mathcal{S} = \{(a_1, a_2) \in \mathbb{R}^2 : a_1 \leq a_2\} \times \mathbb{N} \tag{4.2}$$

so that $\boldsymbol{x} \mapsto (x_{(l)}, x_{(u)}, N)$, where $x_{(k)}$ is the k-th order statistic of $\boldsymbol{x}$ and $l, u \in \{1, \ldots, N\}, l \leq u$ are fixed. Taking $l = 1, u = N$ corresponds to determining the range of the data. Note that this construction explicitly includes the number of underlying datapoints N in the interval as part of the symbol, in direct contrast to almost all existing SDA techniques. This allows random intervals constructed using different numbers of underlying classical datapoints to contribute to the likelihood function in relation to the size of the data that they represent. This is currently not available in the construction of Le-Rademacher and Billard (2011) – see below.

Lemma 4.2. *Consider a univariate interval-valued random variable* $S = (s_l, s_u, n) \in \mathcal{S}$*, obtained through* (4.2) *and assume that* $g_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^{n} g_X(x_i;\theta), \boldsymbol{x} \in \mathbb{R}^n$*. The corresponding symbolic likelihood function is then given by*

$$L(s_l, s_u, n; \theta) = \frac{n!}{(l-1)!(u-l-1)!(n-u)!} \left[G_X(s_l;\theta)\right]^{l-1} \left[G_X(s_u;\theta) - G_X(s_l;\theta)\right]^{u-l-1} \times \left[1 - G_X(s_u;\theta)\right]^{n-u} g_X(s_l;\theta) g_X(s_u;\theta).$$

It is worth noting that this expression can also be obtained by evaluating $\mathbb{P}(S_l \leq s_l, S_u \leq s_l) = \mathbb{P}(X_{(l)} \leq s_l, X_{(u)} \leq s_u)$ and then taking derivatives with respect to s_l and s_u, and corresponds to the joint distribution of order two statistics. This model was previously established by Zhang and Sisson (2016) as a generative model for random intervals built from i.i.d. random variables. Additionally, when univariate intervals are described by their midpoint (M) and (half)-range (R) a simple change of variable $M = (S_1 + S_n)/2$ and $R = (S_n - S_1)/2$ can be applied (e.g. Brito and Duarte Silva (2012)).

4.2.3 Modelling random rectangles

The typical current method of constructing multivariate random rectangles from underlying d-dimensional data $X \in \mathbb{R}^d, d \in \mathbb{N}$ is by taking the cross product of each of the d univariate random intervals described by their marginal minima and maxima (e.g. Neto et al., 2011; Ichino, 2011). The number of underlying datapoints in this rectangle is often not ulitised. Here we improve on this scheme by firstly making use of additional information available at the time of rectangle construction (Section 4.2.3.1), and then by developing several alternative constructions for random rectangles based on marginal order statistics (Section 4.2.3.2).

4.2.3.1 Using marginal maxima and minima

While it is in principle possible to identify a small amount of information about the dependence between two variables summarised by a marginally constructed bounding box, this information content is very weak, and the direction of dependence is not identifiable (Zhang and Sisson, 2016). E.g. if n datapoints are generated from a bivariate normal distribution and the marginal minimum and maximum values are presented, what can be said about the correlation strength and direction? If the correlation is strong relative to the number of observations n, then dependence information can be obtained if the locations of those datapoints involved in construction of the bounding rectangle, and the total number of points are known. For the bivariate examples illustrated in Figure B.1 (top), if the rectangle is generated from only two points (left panel) one can surmise stronger dependence than if three points are used (centre panel), with rectangle construction based on four points (right panel) demonstrating the weakest dependence of all. The exact locations of these bounding points is informative of dependence direction.

As such, we define the aggregation function π to incorporate these construction points (where available) into the definition of the random rectangle as

$$S = \pi(\boldsymbol{X}) : \mathbb{R}^{d\times N} \to \mathcal{S} = \{(a_1, a_2) \in \mathbb{R}^2 : a_1 \le a_2\}^d \times \{2, \dots, \max(2d, n)\} \times \mathcal{T} \times \mathbb{N} \qquad (4.3)$$

so that $\boldsymbol{x} \mapsto \big((x_{(1),i}, x_{(n),i})_{i=1,\dots,d}, p, I(p), N\big)$, where $\boldsymbol{x} = (x_1, \dots, x_n)$, $x_j = (x_{j,1}, \dots, x_{j,d})^\top$ and $x_{(k),i}$ corresponds to the k-th order statistic of the i-th marginal component of $\boldsymbol{x}$. The quantities p and $I(p)$ represent the number of points involved in constructing the random rectangle, and the information about their locations (taking values in $\mathcal{T}$), respectively. In this context a symbol is written as $S = (S_{\min}, S_{\max}, S_p, S_{I_p}, N)$, where $S_{\min}$ and $S_{\max}$ are respectively the d-vectors corresponding to the marginal minima and maxima.

Lemma 4.3. *Consider a multivariate random rectangle $S \in \mathcal{S}$, obtained through* (4.3) *and assume that $g_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^n g_X(x_i;\theta), \boldsymbol{x} \in \mathbb{R}^{n\times d}$. Then the symbolic likelihood function is given by*

$$L(s;\theta) = \frac{n!}{(n-s_p)!}\left[\int_{s_{\min}}^{s_{\max}} g_X(z;\theta)\,dz\right]^{n-s_p} \times \ell_{s_p}, \qquad (4.4)$$

where the multivariate integral is taken over the rectangular region defined by $s_{\min}$ and $s_{\max}$, and where ℓ_{s_p} is defined as follows. If $s_p = 2$, then $s_{I_p} = (s_a, s_b)$ indicates the two co-ordinates of d-dimensional space which define the bounding rectangle. In this case $\ell_2 = g_X(s_a;\theta)g_X(s_b;\theta)$. If $s_p = 2d$, then $s_{I_p} = \varnothing$ (the empty set) and

$$\begin{aligned}\ell_{2d} = &\prod_{i=1}^d \left[G_{X_{-i}|X_i=s_{\min,i}}(s_{\max,-i};\theta) - G_{X_{-i}|X_i=s_{\min,i}}(s_{\min,-i};\theta)\right] g_{X_i}(s_{\min,i}) \\ &\times \prod_{i=1}^d \left[G_{X_{-i}|X_i=s_{\max,i}}(s_{\max,-i};\theta) - G_{X_{-i}|X_i=s_{\max,i}}(s_{\min,-i};\theta)\right] g_{X_i}(s_{\max,i}), \end{aligned} \qquad (4.5)$$

where X_i is the i-th component of X, $X_{-i} = X\backslash X_i$ and similarly for $s_{\min,-i}, s_{\max,-i}, s_{\min,i}$ and $s_{\max,i}$, and $G_{X_{-i}|X_i}$ is the conditional distribution function of X_{-i} given X_i.

In (4.5) the product terms represent the joint distributions functions of X_{-i} being between $s_{\min,-i}$ and $s_{\max,-i}$ given that X_i is equal to $s_{\min,i}$ or $s_{\max,i}$. When $s_p = 2$, (4.5) reduces to $\ell_{2d} = \ell_2$. General expressions for ℓ_{s_p} for $p \ne 2$ or $2d$ can be complex. A simple expression is available in the bivariate case ($d = 2$) for $s_p = 3$.

Corollary 4.4. *For a bivariate random rectangle, if $s_p = 3$ then $S_{I_p} = s_c \in \mathbb{R}^2$ is the co-ordinate of the point defining the bottom-left, top-left, top-right or bottom-right corner of the rectangle. In this case, if $\bar{s}_c$ is the element-wise complement of s_c, then*

$$\ell_3 = g_X(s_c;\theta) \times \prod_{i=1}^2 \left[G_{X_{-i}|X_i=\bar{s}_{c,i}}(s_{\max,-i};\theta) - G_{X_{-i}|X_i=\bar{s}_{c,i}}(s_{\min,-i};\theta)\right] g_{X_i}(\bar{s}_{c,i};\theta). \qquad (4.6)$$

E.g. if $s_c = (s_{\min,1}, s_{\min,2})$ is in the bottom-left corner, then $\bar{s}_c = (s_{\max,1}, s_{\max,2})$.

The first term in (4.6) corresponds to the density of the point in the corner of the rectangle, and the other terms are the probabilities of the two points on the edges being between two interval values given that the other component is fixed. Qualitatively similar expressions can be derived for d-dimensional random rectangles in the cases where $s_p \neq 2$ or $2d$, although there is no simple general expression.

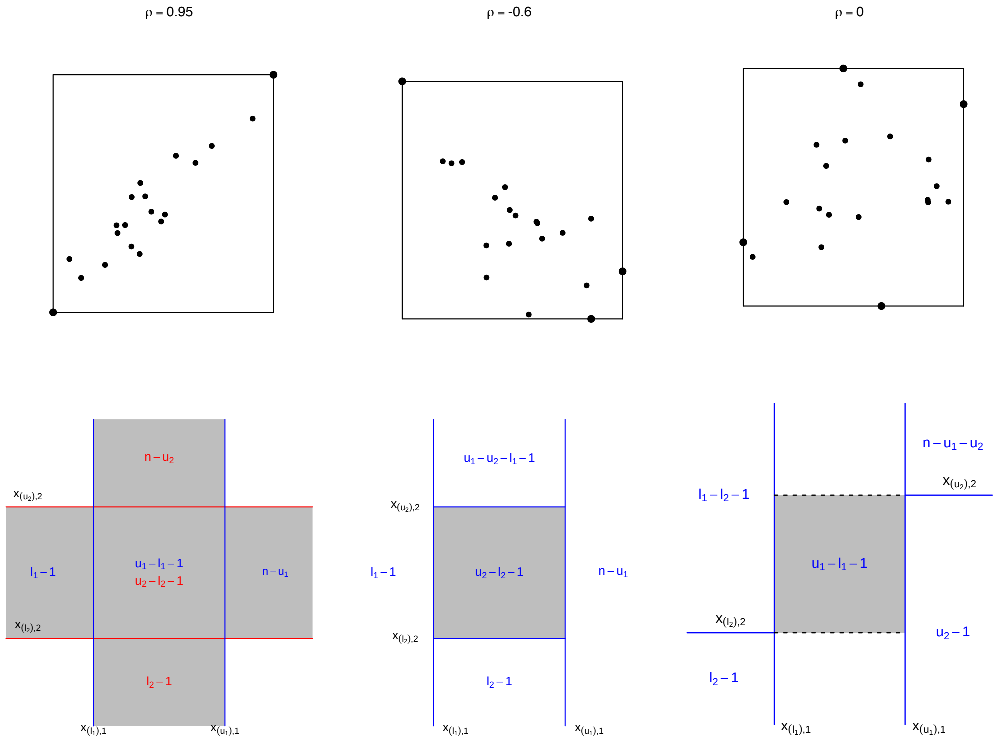

Figure 4.1 – Construction methods for bivariate intervals using marginal minima/maxima (top panels) or marginal order statistics (bottom). Top panels: Illustrative random rectangles constructed from 2 points (high correlation), 3 points (moderate correlation) and 4 points (low/no correlation). Bottom panels: Three alternative construction methods: marginal only (left panel), sequential nesting (centre; equation (4.9)) and iterative segmentation (right; equation (4.11)). Values in blue (red) denote the number of observations in the area bounded by blue (red) lines.

4.2.3.2 Using marginal order statistics

As order statistics are defined in the univariate setting, there are a number of methods to use fixed vectors of lower $l = (l_1, \ldots, l_d)^\top$ and upper $u = (u_1, \ldots, u_d)^\top$ order statistic values, with $1 \leq l_i < u_i \leq N$, to define a d-dimensional random rectangle. The simplest takes the cross product of the d-univariate marginal quantiles (e.g. Neto et al., 2011). Here the aggregation function π is defined as

$$S = \pi(\boldsymbol{X}) : \mathbb{R}^{d\times N} \rightarrow \mathcal{S} = \{(a_1, a_2) \in \mathbb{R}^2 : a_1 \leq a_2\}^d \times \mathbb{N} \tag{4.7}$$

$$\boldsymbol{x} \mapsto \left(\left(x_{(l_i),i}, x_{(u_i),i} \right)_{i=1,\ldots,d}, N \right). \tag{4.8}$$

In this context the symbol is written as $S = (S_l, S_u, N)$, where S_l and S_u are respectively the d-vectors corresponding to the marginal lower and upper order statistics. This process is illustrated in Figure B.1 (bottom left panel) in the $d = 2$ setting. For fixed l and u, the observed counts in each region are then known as a function of the construction (4.8).

The resulting symbolic likelihood function is then given by

$$L(s;\theta) = \prod_{i=1}^{d} L(s_{l_i}, s_{u_i}, n; \theta_i)$$

where $L(s_{l_i}, s_{u_i}, n; \theta_i)$ is as obtained in Lemma 4.2 using the i-th marginal distribution with parameter $\theta_i \in \Theta$. Indeed, the symbol S can also be written as $(S_1, \dots, S_d)$, a d-vector of independent random intervals obtained through (4.2). However, as the construction (4.8) only contains marginal information, such a symbol will fail to adequately capture dependence between variables. As an alternative, we introduce two new order-statistic based representations of random rectangles that do account for such dependence.

The first, sequential nesting (Figure B.1, bottom centre panel), constructs the order statistics within dimension i conditionally on already being within the random rectangle in dimensions $j < i$. The aggregation function π is given by (4.7) as before, but where now

$$\boldsymbol{x} \mapsto \Big(\big(\big(x_{(l_i),i}, x_{(u_i),i} \big) | \{ x_{(l_j),j} < x_j < x_{(u_j),j}; j < i \} \big)_{i=1,\dots,d}, N \Big). \tag{4.9}$$

As before, $S = (S_l, S_u, N)$, but where the known observed counts now lie in different regions (Figure B.1), and with the additional constraints of $2 \le u_{i+1} \le u_i - l_i - 1$.

Lemma 4.5. *Consider a multivariate random rectangle $S \in \mathcal{S}$, constructed via* (4.9) *and suppose that $g_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^{n} g_X(x_i;\theta), \boldsymbol{x} \in \mathbb{R}^{n\times d}$. The symbolic likelihood function is then given by*

$$L(s;\theta) \propto \mathbb{P}(s_{\boldsymbol{l}} < X < s_{\boldsymbol{u}})^{u_d - l_d - 1} d\mathbb{P}(X_1 < s_{l,1}) d\mathbb{P}(X_1 < s_{u,1}) \prod_{i=1}^{d} p_i(s_{\boldsymbol{l}}) q_i(s_{\boldsymbol{u}}), \tag{4.10}$$

where $p_1(s_{\boldsymbol{l}}) = \mathbb{P}(X_1 < s_{l,1})^{l_1 - 1}$, $q_1(s_{\boldsymbol{u}}) = \mathbb{P}(X_1 > s_{u,1})^{n-u_1}$ and for $i = 2, \dots, d$,

$$\begin{aligned} p_i(s_{\boldsymbol{l}}) =& \mathbb{P}(s_{l,j} < X_j < s_{u,j}; j < i | X_i = s_{l,i}) d\mathbb{P}(X_i < s_{l,i}) \\ & \times \mathbb{P}(X_i < s_{l,i} | s_{l,j} < X_j < s_{u,j}; j < i)^{l_i - 1} \\ q_i(s_{\boldsymbol{u}}) =& \mathbb{P}(s_{l,j} < X_j < s_{u,j}; j < i | X_i = s_{u,i}) d\mathbb{P}(X_i < s_{u,i}) \\ & \times \mathbb{P}(X_i > s_{u,i} | s_{l,j} < X_j < s_{u,j}; j < i)^{u_{i-1} - u_i - l_{i-1} - 1}. \end{aligned}$$

Corollary 4.6. *With $d = 2$, the symbolic likelihood function in Lemma 4.5 is given by*

$$\begin{aligned} L(s;\theta) \propto & \left(G_X(s_{\boldsymbol{u}}) - G_X(s_{\boldsymbol{l}}) \right)^{u_2 - l_2 - 1} g_{X_1}(s_{l,1}) g_{X_1}(s_{u,1}) g_{X_2}(s_{l,2}) g_{X_2}(s_{u,2}) \\ & \times G_{X_1}(s_{l,1})^{l_1 - 1} \left[1 - G_{X_1}(s_{u,1}) \right]^{n - u_1} \left[G_{X_1|X_2 = s_{l,2}}(s_{u,1}) - G_{X_1|X_2 = s_{l,2}}(s_{l,1}) \right] \\ & \times \left[G_{X_1|X_2 = s_{u,2}}(s_{u,1}) - G_{X_1|X_2 = s_{u,2}}(s_{l,1}) \right] \left[G_X((s_{u,1}, s_{l,2})) - G_X(s_{\boldsymbol{l}}) \right]^{l_2 - 1} \\ & \times \left[G_{X_1}(s_{u,1}) - G_X(s_{\boldsymbol{u}}) - G_{X_1}(s_{l,1}) + G_X((s_{l,1}, s_{u,2})) \right]^{u_1 - u_2 - l_1 - 1}, \end{aligned}$$

where $G_{X_i}(\cdot) \equiv G_{X_i}(\cdot;\theta)$ and $G_{X_i|X_j}(\cdot) \equiv G_{X_i|X_j}(\cdot;\theta); i \neq j$ respectively denote the marginal and conditional distribution functions of $g_{\boldsymbol{X}}(\boldsymbol{x};\theta)$.

An alternative to sequential nesting is an iterative segmentation construction (Figure B.1, bottom right). As before, for fixed vectors l and u, the aggregation function π is given by (4.7) but where

$$\boldsymbol{x} \mapsto \left(\left(x_{(l_i),i} | \{ x_j < x_{(l_j),j}; j < i \}, x_{(u_i),i} | \{ x_j > x_{(u_j),j}; j < i \} \right)_{i=1,\dots,d}, N \right). \tag{4.11}$$

Again $S = (S_l, S_u, N)$, but now where $S_{l,i}$, the l_i-th order statistic of the i-th margin, is restricted to the area where the previous margins $j < i$ are all below their respective lower (l_j-th) order statistic. Similarly, $S_{u,i}$ is restricted to the area where the previous margins $j < i$ are all above their respective upper order statistic. For fixed l and u the observed counts are then known (Figure B.1, bottom right) but are attributed to different regions than for sequential nesting. Iterative segmentation implies the additional constraints $l_{i+1} < l_i - 1$ and $u_{i+1} < N - \sum_{j=1}^{i} u_j$ for $i = 1, \dots, d-1$.

Lemma 4.7. *Consider a multivariate random rectange $S \in \mathcal{S}$, constructed via (4.11) and suppose that $g_{\boldsymbol{X}}(\boldsymbol{x}; \theta) = \prod_{i=1}^{n} g_X(x_i; \theta), \boldsymbol{x} \in \mathbb{R}^{n \times d}$. The symbolic likelihood function is then given by*

$$L(s;\theta) \propto \mathbb{P}(s_{l,1} < X_1 < s_{u,1})^{u_1 - l_1 - 1} d\mathbb{P}(X_1 < s_{l,1}) d\mathbb{P}(X_1 < s_{u,1}) \prod_{i=2}^{d+1} p_i(s_{\boldsymbol{l}}) q_i(s_{\boldsymbol{u}}), \tag{4.12}$$

where $p_{d+1}(s_{\boldsymbol{l}}) = \mathbb{P}(X_1 < s_{l,1}, \dots, X_d < s_{l,d})^{l_d - 1}$, $q_{d+1}(s_{\boldsymbol{u}}) = \mathbb{P}(X_1 > s_{u,1}, \dots, X_d > s_{u,d})^{n - \sum_{i=1}^{d} u_i}$ and for $i = 2, \dots, d$

$$\begin{aligned} p_i(s_{\boldsymbol{l}}) =& \mathbb{P}(X_j < s_{l,j}; j < i | X_i = s_{l,i}) d\mathbb{P}(X_i < s_{l,i}) \\ & \qquad \times \left[\mathbb{P}(X_j < s_{l,j}; j < i) - \mathbb{P}(X_j < s_{l,j}; j \le i) \right]^{l_i - l_{i-1} - 1} \\ q_i(s_{\boldsymbol{u}}) =& \mathbb{P}(X_j > s_{u,j}; j < i | X_i = s_{u,i}) d\mathbb{P}(X_i < s_{u,i}) \\ & \qquad \times \left[\mathbb{P}(X_j > s_{u,j}; j < i) - \mathbb{P}(X_j > s_{u,j}; j \le i) \right]^{u_i - 1}. \end{aligned}$$

Corollary 4.8. *With $d = 2$, the symbolic likelihood function in Lemma 4.7 is given by*

$$\begin{aligned} L(s;\theta) \propto & \left(G_{X_1}(s_{u,1}) - G_{X_1}(s_{l,1}) \right)^{u_1 - l_1 - 1} g_{X_1}(s_{l,1}) g_{X_1}(s_{u,1}) g_{X_2}(s_{l,2}) g_{X_2}(s_{u,2}) \\ & \times G_{X_1 | X_2 = s_{l,2}}(s_{l,1}) (1 - G_{X_1 | X_2 = s_{u,2}}(s_{u,1})) \left[G_{X_1}(s_{l,1}) - G_X(s_{\boldsymbol{l}}) \right]^{l_2 - l_1 - 1} \\ & \times \left[G_{X_2}(s_{u,2}) - G_X(s_{\boldsymbol{u}}) \right]^{u_2 - 1} G_X(s_{\boldsymbol{l}})^{l_2 - 1} (1 - G_{X_1}(s_{u,1}) - G_{X_2}(s_{u,2}) - G_X(s_{\boldsymbol{u}}))^{n - u_1 - u_2}, \end{aligned}$$

where $G_{X_i}(\cdot) \equiv G_{X_i}(\cdot; \theta)$ and $G_{X_i|X_j}(\cdot) \equiv G_{X_i|X_j}(\cdot; \theta); i \neq j$ respectively denote the marginal and conditional distribution functions of $g_{\boldsymbol{X}}(\boldsymbol{x}; \theta)$.

When $l_1 = \dots = l_d = 1$ and $u_i = n - 2(i-1)$, the sequential nesting random interval construction (4.9) approximately reduces to the random rectangle construction (4.3) based on univariate marginal maxima and minima, indicating some degree of construction consistency. That is, $S = (S_l, S_u, N)$ contains almost exactly the same information as the symbol $S = (S_{\min}, S_{\max}, S_p, S_{I_p}, N)$ when $S_p = 2d$, and so the symbolic likelihood function (4.10) approximately reduces to (4.4). For highly correlated data $S = (S_l, S_u, N)$ is slight more informative as the lower and upper bounds of each dimension i are calculated on a set from which the $(i-1)$ lowest and largest observations are been removed. The approximation improves as the correlation decreases until both symbols are identical when the data are completely independent. A similar reduction cannot be obtained for the iterative segmentation construction. Also note that both sequential nesting and iterative segmentation are dependent on the ordering of the variables, and that varying the ordering will produce different representations of the same underlying dataset.

4.2.4 Modelling histograms with random counts

Histograms are a very popular SDA tool to represent the distribution of continuous data, with a typical focus on univariate histograms. They are most commonly constructed as a set of fixed consecutive intervals for which the random relative frequencies (or counts) are reported (e.g. Bock and Diday, 2000; Billard and Diday, 2006) Let $\mathcal{X} = \mathbb{R}^d, d \in \mathbb{N}$. Following Le-Rademacher and Billard (2011), a histogram-valued random variable may be defined as a set of counts associated with a deterministic partition of the domain $\mathcal{X}$. Suppose that the i-th margin of $\mathcal{X}$ is partitioned into B^i bins, so that $B^1 \times \cdots \times B^d$ bins are created in $\mathcal{X}$ through the d-dimensional intersections of each marginal bin. Index each bin by $\boldsymbol{b} = (b_1, \cdots, b_d)$, $b_j = 1, \ldots, B^j$ as the vector of co-ordinates of each bin in the histogram. Each bin $\boldsymbol{b}$ may then be constructed as

$$\mathcal{B}_{\boldsymbol{b}} = \mathcal{B}^1_{b_1} \times \cdots \times \mathcal{B}^d_{b_d} \quad \text{where} \quad \mathcal{B}^j_{b_j} = (y^j_{b_j-1}, y^j_{b_j}],\ j = 1, \ldots, d,$$

where for each j, the marginal sequences $-\infty < y^j_0 < y^j_1 < \ldots < y^j_{B^j} < \infty$ are fixed. We assume that all data counts outside of the constructed histogram are zero. A d-dimensional histogram-valued random variable is constructed through the aggregation function π where

$$\begin{array}{lll} S = \pi(\boldsymbol{X}): & \mathbb{R}^{d\times N} & \to \mathcal{S} = \{0, \ldots, N\}^{B^1 \times \cdots \times B^d} \\ & \boldsymbol{x} & \mapsto \left(\sum_{i=1}^n \mathbb{I}\{x_i \in \mathcal{B}_{\mathbf{1}}\}, \ldots, \sum_{i=1}^n \mathbb{I}\{x_i \in \mathcal{B}_{\boldsymbol{B}}\}\right), \end{array} \tag{4.13}$$

where $\mathbf{1} = (1, \ldots, 1)$ and $\boldsymbol{B} = (B^1, \ldots, B^d)$, and $\mathbb{I}$ is the indicator function. The resulting symbol $S = (S_{\mathbf{1}}, \ldots, S_{\boldsymbol{B}})$ is a vector of counts, where $S_{\boldsymbol{b}}$ denotes the frequency of data in bin $\mathcal{B}_{\boldsymbol{b}}$, such that $\sum_{\boldsymbol{b}} S_{\boldsymbol{b}} = N$.

Lemma 4.9. *Consider a multivariate histogram-valued random variable $S \in \mathcal{S}$, constructed via* (4.13) *and suppose that $g_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^n g_X(x_i;\theta), \boldsymbol{x} \in \mathbb{R}^{n\times d}$. The symbolic likelihood function is then given by*

$$L(s;\theta) = \frac{n!}{s_{\mathbf{1}}! \cdots s_{\boldsymbol{B}}!} \prod_{\boldsymbol{b}} \left(\int_{\mathcal{B}_{\boldsymbol{b}}} g_X(z;\theta)\, dz \right)^{s_{\boldsymbol{b}}}, \tag{4.14}$$

where the integral denotes the probability that data $x \in \mathcal{X}$ falls in bin $\mathcal{B}_{\boldsymbol{b}}$ under the model.

In the univariate setting, the resulting multinomial likelihood coincides with the likelihood function for binned and truncated data introduced by McLachlan and Jones (1988). The resulting likelihood also agrees with and extends the methodology of Heitjan and Rubin (1991) who construct corrected likelihood functions for coarsened data, where the authors highlight the necessity to account for both the grouping and the stochastic nature of the coarsening. In our construction, this latter point is achieved in (4.1) by the conditional density $f_{S|\boldsymbol{X}}$.

In the limit as the histogram is reduced to its underlying classical data, the likelihood (4.14) reduces to the classical data likelihood. In this case, as the number of bins becomes large each bin of the histogram reduces in size and approaches a single point $\mathcal{B}_{\boldsymbol{b}} \to x_b \in \mathbb{R}^d$. In the limit as the number of bins $\to \infty$, only those n coinciding with the underlying data points will have a count of 1, while the others will have a count of 0. The likelihood contribution of the non-empty bins $\mathcal{B}_{\boldsymbol{b}}$ is then $g_X(x_b;\theta)$. This is equivalent to specifying $f_{S|\boldsymbol{X}=\boldsymbol{z}}(s;\vartheta) = \prod_{i=1}^n \delta_{z_i}(x_i)$ in (4.1). Consequently the symbolic likelihood function reduces to $L(\boldsymbol{x};\theta) \propto \prod_{i=1}^n g_X(x_i;\theta)$.

Finally, note that while we assumed that the histogram covers the domain $\mathcal{X}$ of the classical random variable, this will not be the case when e.g. the classical data is only observed on a subset of the domain. In this scenario the distribution of the classical variable $g_{\boldsymbol{X}}(\boldsymbol{x};\theta)$ should be truncated and rescaled over the same subdomain.

4.2.5 Modelling histograms with random bins

A common alternative to histograms with random counts over fixed bins is constructing histograms with fixed counts within random bins. (e.g. Mousavi and Zaniolo, 2011; Ioannidis, 2003). Such random histograms can be seen as a generalisation of the interval-valued random variables from Sections 4.2.2–4.2.3. In particular, random intervals can be viewed as histograms with the number of bins ranging from 1 (when all margins are intervals calculated from sample minima and maxima; Figure B.1 top panels) to $3d$ (where all margins are intervals calculated from order statistics $l > 1$ and $u < n$; Figure B.1 bottom left). In the following we focus on the univariate setting ($\mathcal{X} = \mathbb{R}$) since extension to d-dimensions is challenging. E.g. given a matrix of counts, then a simply constructed grid matching these counts does not necessarily exist.

We construct a univariate random histogram using order statistics. For a vector of orders $k = (k_1, \ldots, k_B)$, such that $1 \le k_1 \le \cdots \le k_B \le N$, a univariate random histogram is constructed through the aggregation function π where

$$\begin{aligned} S = \pi(\boldsymbol{X}): \quad \mathbb{R}^N \quad &\to \mathcal{S} = \{(a_1, \ldots, a_B) \in \mathbb{R}^B : a_1 \le \cdots \le a_B\} \times \mathbb{N} \\ \boldsymbol{x} \quad &\mapsto (x_{(k_1)}, \ldots, x_{(k_B)}, N). \end{aligned} \tag{4.15}$$

This defines a histogram with bin b located at $(s_{b-1}, s_b]$ with fixed count $k_b - k_{b-1}$, for $b = 1, \ldots, B+1$, where $s_0 = -\infty, s_{B+1} = +\infty, k_0 = 0$ and $k_{B+1} = N+1$, and knowledge that there is a point located at each s_b, $b = 1, \ldots, B$. The resulting symbol $S = ((S_1, \ldots, S_B), N)$ is a B-vector of order statistics plus the total number of datapoints.

Lemma 4.10. *Consider a univariate random histogram $S \in \mathcal{S}$, obtained through* (4.15) *and assume that $g_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^n g_X(x_i), \boldsymbol{x} \in \mathbb{R}^{n\times d}$. Then the symbolic likelihood function* *is given by*

$$L(s;\theta) = n! \prod_{b=1}^{B} g_X(s_b;\theta) \prod_{b=1}^{B+1} \frac{(G_X(s_b;\theta) - G_X(s_{b-1};\theta))^{k_b - k_{b-1} - 1}}{(k_b - k_{b-1} - 1)!}. \tag{4.16}$$

When $B = 2$, $k_1 = l$ and $k_2 = u$ with $l, u = 1, \ldots, n; l \neq u$, then (4.16) reduces to the likelihood function in Lemma 4.2 (see Appendix B.1.4). Further, under this construction it is straightforward to show that if $B = N$ then the symbolic likelihood (4.16) recovers the classical data likelihood. Specifically this implies $k_b = b$ for all $b = 1, \ldots, B$ so that the aggregation function (4.15) is $S = \pi(\boldsymbol{X}) = ((X_{(1)}, \ldots, X_{(n)}), N)$, $k_b - k_{b-1} = 1$ for all b and so $L(s;\theta) \propto \prod_{b=1}^{N} g_X(x_b;\theta)$.

4.3 Illustrative analyses

The symbolic likelihood function introduced in Section 4.2 not only resolves many of the conceptual and practical issues with current SDA methods, but it also opens the door for new classes of symbol design and construction, in addition to opening up SDA as a viable tool to enable and improve upon classical data analyses. We explore each of these benefits below.

4.3.1 Data reconstruction for meta-analyses

In medical research, meta-analyses of results from multiple trials are often implemented to systematically examine the clinical effects of certain treatments. These meta-analyses typically require the effect sample mean and standard deviation from each individual study. However it is common practice for such studies to only report various quantile summary statistics, namely the sample minimum (q_0), maximum (q_4) and the sample quartiles (q_1, q_2, q_3). This introduces the problem of accurately estimating a sample mean and standard deviation from these quantiles.

The most sophisticated practiced method to estimate the sample mean was developed by Luo et al. (2018) based on previous work by Hozo et al. (2005) and Wan et al. (2014), whereby

$$\hat{\bar{x}}_L = w_1 \left(\frac{q_0 + q_4}{2} \right) + w_2 \left(\frac{q_1 + q_3}{2} \right) + (1 - w_1 - w_2) q_2, \tag{4.17}$$

with $w_1 = 2.2/(2.2 + n^{0.75})$ and $w_2 = 0.7 - 0.72/n^{0.55}$. Based on previous work by Hozo et al. (2005) and Bland (2015) the best performing estimators of the sample standard deviation are due to Wan et al. (2014) and Shi et al. (2018), which are respectively given by

$$\hat{s}_W = \frac{1}{2} \left(\frac{q_4 - q_0}{\zeta(n)} + \frac{q_3 - q_1}{\eta(n)} \right) \quad \text{and} \quad \hat{s}_S = \frac{q_4 - q_0}{\theta_1(n)} + \frac{q_3 - q_1}{\theta_2(n)}, \tag{4.18}$$

where $\zeta(n) = 2\Phi^{-1}\left(\frac{n-0.375}{n+0.25}\right)$, $\eta(n) = 2\Phi^{-1}\left(\frac{0.75n-0.125}{n+0.25}\right)$, $\theta_1(n) = (2 + 0.14n^{0.6})\Phi^{-1}(\frac{n-0.375}{n+0.25})$, $\theta_2(n) = (2 + \frac{2}{0.07n^{0.6}})\Phi^{-1}(\frac{0.75n-0.125}{n+0.25})$, and $\Phi^{-1}(\cdot)$ is the inverse of the standard normal c.d.f. Each estimator in (4.17) and (4.18) assumes the underlying data are normally distributed.

In the context of the symbolic random variables developed in Section 4.2, this setting corresponds to constructing the symbolic variable S defined through (4.15) with $n = 4Q + 1, Q \in \mathbb{N}$ where $k = (1, Q+1, 2Q+1, 3Q+1, n)$ i.e. a histogram with $B = 4$ random bins and equal counts.

If we make the same assumption of i.i.d. normality of the underlying data, then maximising the symbolic likelihood (4.16) with $g_X(x;\theta) = \phi(x;\mu,\sigma)$ will yield maximum likelihood estimators $\hat{\theta} = (\hat{\mu}, \hat{\sigma}) \approx (\bar{x}, \sqrt{(n-1)/ns})$ which provide direct estimates $(\hat{\bar{x}}_*, \hat{s}_*) = (\hat{\mu}, \sqrt{n/(n-1)}\hat{\sigma})$ of the sample mean $\bar{x}$ and standard deviation s of the underlying data. Of course, the symbolic likelihood (4.16) can make any alternative distributional assumption on the underlying data.

Figure 4.2 illustrates the performance of each estimator when compared to the true sample values (i.e. $(\hat{\bar{x}} - \bar{x}_0)$ and $(\hat{s} - s_0)$) based on data generated from normal (top panels) and lognormal (bottom) distributions, averaged over 10,000 replicates, and for a range of sample sizes n.

For normally distributed data, the estimator of the sample mean $\hat{\bar{x}}_L$ by Luo et al. (2018) (red squares) and the symbolic likelihood-based estimator (green circles) perform comparably (top left panel). Identifying performance differences when estimating the sample standard deviation is much clearer however (top right panel), with the symbolic estimator strongly outperforming the discipline-standard estimators of Wan et al. (2014) and Shi et al. (2018) (blue triangles and purple diamonds, respectively). The differences are particularly stark for low sample sizes. Because $\hat{s}_W$ and $\hat{s}_S$ are substantially overestimating the true standard deviation, this means that their usage in medical meta-analyses will systematically undervalue the contribution of each study in the larger analysis, potentially weakening the power of the study to detect significant clinical effects.

Note that for $n = 5$, the symbolic estimator of the sample standard deviation is exact (i.e. zero error) as the symbolic likelihood (4.16) reduces to the classical likelihood in this case.

When the sample data are lognormal (bottom panels), both symbolic (light green) and the industry-standard estimators perform poorly. This is not surprising as they are each based on a normality assumption. While estimators equivalent to those in (4.17) and (4.18) but for lognormally distributed data could in principle be derived, it is trivial to achieve this for the symbolic estimator by substituting the lognormal density (or any other desired distribution) for $g_X(\cdot;\theta)$ in (4.16). The resulting sample mean and standard deviation estimators assuming the lognormal distribution are illustrated by dark green circles. The lognormal-based symbolic likelihood estimator performance is clearly excellent in comparison.

One factor that influences the efficiency of the symbolic mle is the form and specification of the symbol as a summary representation of the underlying data. While a random histogram with more bins should be more informative than one with less, for a fixed number of bins, sensible choice of their location can result in increased precision of the symbolic mle. This idea of *symbol design* has been largely ignored in the SDA literature, for example, with random intervals being routinely constructed using sample maxima and minima.

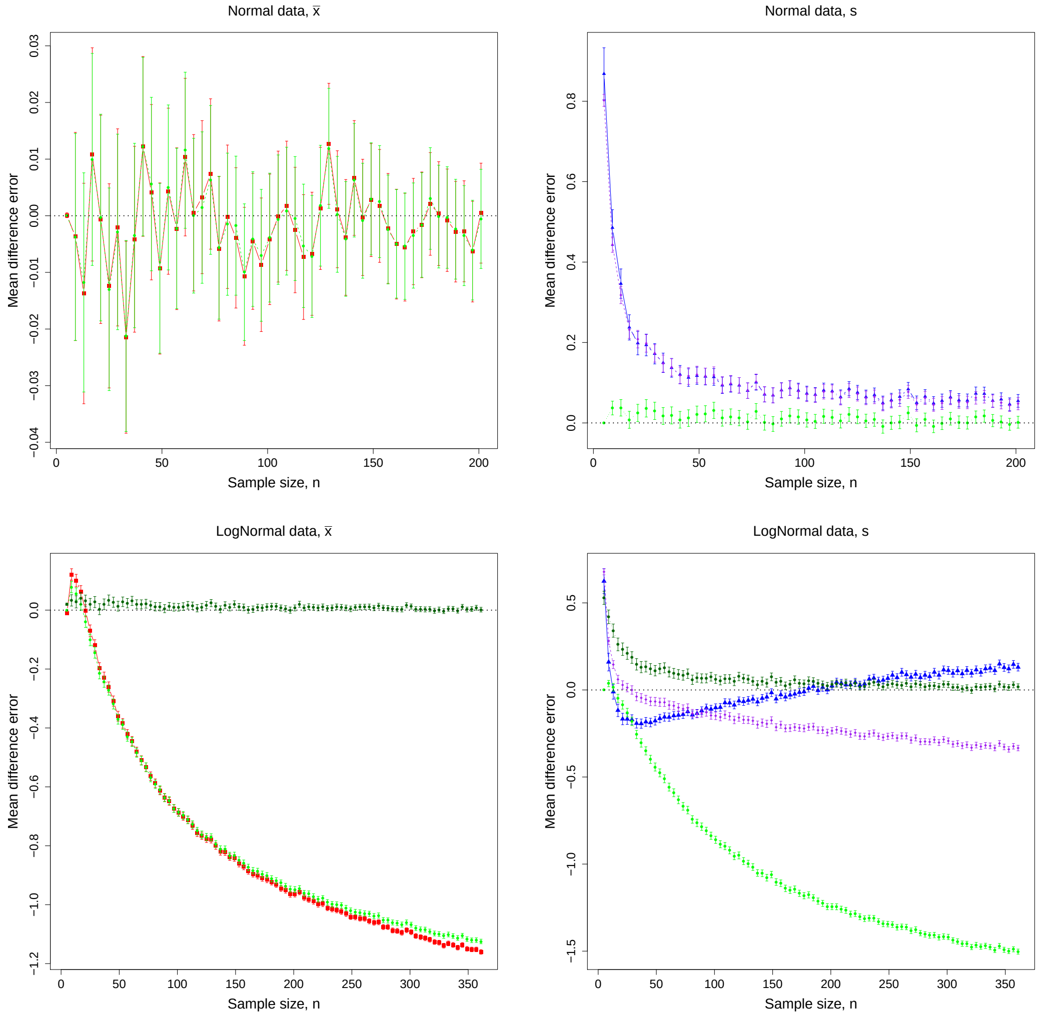

Figure 4.2 – Mean difference errors, $(\hat{\hat{x}} - \bar{x}_0)$ and $(\hat{s} - s_0)$, of various estimates of the sample mean (left panels) and standard deviation (right panels) as a function of sample size $n = 4Q + 1, Q = 1, \ldots, 50$ or 90, and for both normally (top panels) and log-normally (bottom-panels) distributed data. $\bar{x}_0$ and s_0 denote the true sample mean and standard deviation for each dataset. Errors are averaged over 10,000 dataset replicates generated from $\theta_0 = (\mu_0, \sigma_0) = (50, 17)$ (normal data) and $\theta_0 = (\mu_0, \sigma_0) = (4, 0.3)$ following Hozo et al. (2005) and Luo et al. (2018). Colouring indicates the SDA estimates (light and dark green circles), $\hat{\hat{x}}_L$ (red squares), $\hat{s}_W$ (blue triangles) and $\hat{s}_S$ (purple diamonds). Confidence intervals indicate ±1.96 standard errors.

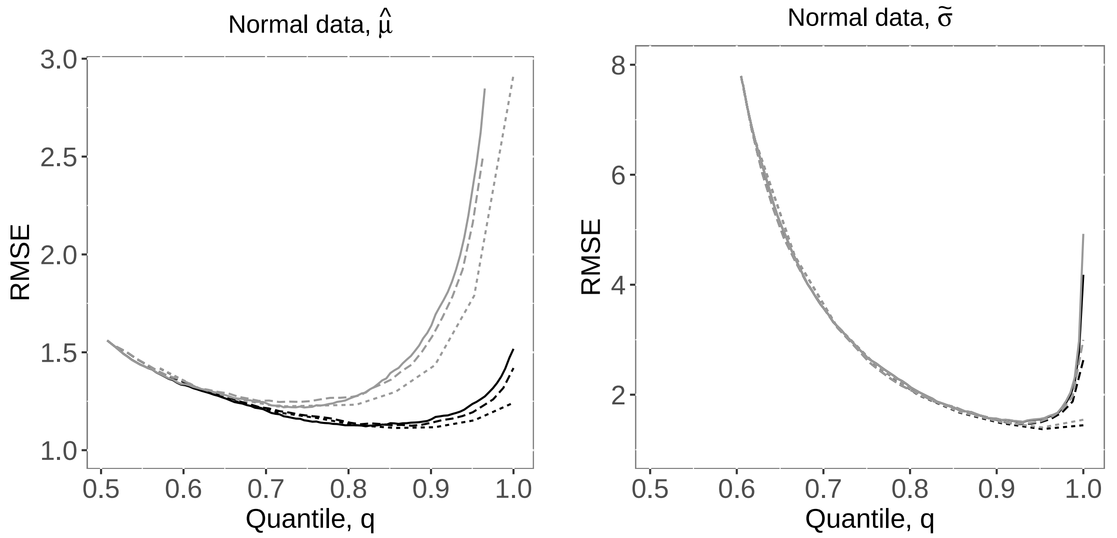

Figure 4.3 – $\text{RMSE}_{\hat{\mu}}$ (left) and $\text{RMSE}_{\tilde{\sigma}}$ (right) as a function of quantile $q = (n+1-i)/n$ for $i = 1, \ldots, (n+1)/2$. Grey and black lines respectively denote random intervals and histograms. Solid, long-dashed and short-dashed lines indicate samples of size $n = 21, 81$ and 201 respectively.

We consider the simplified setting of the univariate random interval $S = (s_l, s_u, n)$ defined in Lemma 4.2 constructed using symmetric upper and lower order statistics, and the associated 2-bin random histogram (4.15) that results by additionally including the sample median, q_2.

That is, for sample sizes $n = 4Q+1, Q \in \mathbb{N}$ we have $l = i, u = n+1-i$ for the interval and $k = (i, 2Q+1, n+1-i)$ for the histogram, where we examine the efficiency of the resulting symbolic mle for the symbols defined by $i = 1, \ldots, 2Q$. For each of $t = 1, \ldots, T = 10,000$ replicate datasets of size $n = 21, 81$ and 201 (i.e. $Q = 5, 20$ and 50) drawn from a $N(\mu_0, \sigma_0)$ distribution with $(\mu_0, \sigma_0) = (50, 17)$, we compute the rescaled symbolic mle $(\hat{\mu}_t, \tilde{\sigma}_t)$ where $\tilde{\sigma}_t = \sqrt{n/(n-1)}\hat{\sigma}_t$, and then calculate the relative mean square errors (RMSE) defined by

$$\text{RMSE}_{\hat{\mu}} = \frac{\sum_{t=1}^{T}(\hat{\mu}_t - \mu_0)^2}{\sum_{t=1}^{T}(\bar{x}_t - \mu_0)^2} \quad \text{and} \quad \text{RMSE}_{\tilde{\sigma}} = \frac{\sum_{t=1}^{T}(\tilde{\sigma}_t - \sigma_0)^2}{\sum_{t=1}^{T}(s_t - \sigma_0)^2},$$

where $\bar{x}_t$ and s_t denote the sample mean and standard deviation of the t-th replicate.

Figure 4.3 displays the RMSEs as function of the quantile $q = (n+1-i)/n$ used to construct the symbol. As might be expected, using a histogram (dark lines) provides more information about μ than the associated random interval (grey lines), as the extra information contained in the median is informative for this parameter. In contrast, the median does not provide any information about σ in addition to the two bounding quantiles, given the symmetric underlying distribution. The inclusion of alternative quantiles would be informative, however.

The convex shapes of each RMSE curve indicates that the current prevailing SDA practice of constructing intervals from sample minima and maxima ($i = 1, q = n$) is highly inefficient for parameter estimation. Greater precision for both location and scale parameters is achieved by using less extreme quantiles, in this setting around the $q = 0.85$-0.90 range (trading off the optimal minimum RMSE values between the two parameters). There is additionally a severe penalty for using too low quantiles when estimating σ, as the scale of the data can not easily be estimated using overly central quantities. Estimating μ is less sensitive in this regard. These conclusions are robust to the sample size, n. Overall this analysis indicates that substantial efficiency gains should be possible in standard SDA analyses with more informed symbol design.

4.3.2 Information content in multivariate random rectangles

In Sections 4.2.3.1 and 4.2.3.2 we introduced two new symbolic constructions to increase the information content within multivariate random rectangles. We now examine the performance of each of these representations and contrast them with standard SDA constructions. We focus on bivariate intervals for clarity, where extension of the results to higher dimensions is immediate.

When constructing random rectangles from marginal minima and maxima, Lemma 4.3 and Corollary 4.4 provide an expression for the symbolic likelihood that incorporates full knowledge of the number and location of the unique points from which the interval is constructed (e.g. Figure B.1). We denote the resulting likelihood function (4.4) by $L_{\text{full}}(s;\theta)$. Existing SDA definitions of random rectangles do not use this information. In its absence, the best likelihood model that can be constructed is by averaging the likelihood L_{full} over all possible combinations of the unique point constructions, weighted according to the probability of that configuration arising under the classical data model. That is,

$$L_{\varnothing}(s;\theta) = \sum_{t_p}\sum_{t_{I_p}} L_{\text{full}}((s_{\min}, s_{\max}, t_p, t_{I_p}, n);\theta)\mathbb{P}(S_p = t_p, S_{I_p} = t_{I_p};\theta),$$

where

$$\mathbb{P}(S_p = t_p, S_{I_p} = t_{I_p};\theta) = \int\int L_{\text{full}}((a, b, t_p, t_{I_p}, n);\theta)I(a \le b)\mathrm{d}a\mathrm{d}b, \tag{4.19}$$

where $a = (a_1, \ldots, a_d)$, $b = (b_1, \ldots, b_d)$, and where $I(a \le b) = \prod_{i=1}^{d} I(a_i \le b_i)$. In the following analyses we estimate the probabilities (4.19) to a high accuracy using Monte Carlo with a large number of samples, each time $L_{\varnothing}$ is evaluated. Clearly this is not viable in practice. One alternative is to assume that each random rectangle has been constructed by the maximum number of unique points ($2d$), which is perhaps not completely unrealistic when the number of points n underlying a symbol is large compared to the dependence between the variables. We denote the likelihood $L_{2d}(s;\theta)$ as the particular case of L_{full} with $S_p = 2d$. In the following, L_{2d} effectively represents the current state-of-the-art in SDA methods, $L_{\varnothing}$ represents the best that can likely be done with the existing constructions of random rectangles in the SDA literature (although it is challenging to implement in practice), and L_{full} is our proposed construction.

Following the notation of Section 4.2, we assume $m = 20, 50$ classes ($\mathcal{C} = \{1, \ldots, m\}$) for each of which a random sample of size $n_c = 5, 10, 50, 100$ is drawn from a $N_2(\mu_0, \Sigma_0)$ distribution ($d = 2$) with $\mu_0 = (2, 5)^\top$, $\mathrm{diag}(\Sigma_0) = (\sigma^2_{0,1}, \sigma^2_{0,2}) = (0.5, 0.5)$ and correlation

$\rho_0 = 0, 0.3, 0.5, 0.7, 0.9$. The m random rectangles are then constructed, retaining the information (s_p, s_{I_p}) required to maximise L_{full} but which is ignored when maximising $L_{\varnothing}$ and L_4. For each of $T = 100$ replicate datasets, the symbolic mle $\hat{\theta} = (\hat{\mu}, \hat{\Sigma})$ is computed.

		$m = 20$				$m = 50$			
n_c		5	10	50	100	5	10	50	100
$\rho_0 = 0.0$	L_4	0.0043	−0.0011	−0.0004	0.0022	−0.0036	−0.0025	0.0008	0.0009
		(0.0711)	(0.0540)	(0.0262)	(0.0197)	(0.0560)	(0.0321)	(0.0164)	(0.0152)
	$L_{\varnothing}$	−0.0178	−0.0170	0.0164	−0.0056	−0.0545	−0.0179	0.0163	−0.0093
		(0.4763)	(0.0592)	(0.0138)	(0.0173)	(0.3989)	(0.0289)	(0.0106)	(0.0159)
	L_{full}	−0.0006	0.0145	−0.0008	0.0061	−0.0090	0.0011	−0.0082	−0.0009
		(0.1262)	(0.1233)	(0.1309)	(0.1461)	(0.0871)	(0.0823)	(0.0897)	(0.1005)
0.3	L_4	0.0821	0.0342	0.0104	0.0060	0.0888	0.0406	0.0102	0.0072
		(0.0802)	(0.0550)	(0.0296)	(0.0251)	(0.0457)	(0.0351)	(0.0177)	(0.0150)
	$L_{\varnothing}$	0.4988	0.0145	0.0217	−0.0051	0.5231	0.0336	0.0225	−0.0043
		(0.2813)	(0.0588)	(0.0110)	(0.0188)	(0.1150)	(0.0435)	(0.0094)	(0.0175)
	L_{full}	0.3036	0.2974	0.2957	0.2732	0.3063	0.3031	0.3037	0.2892
		(0.1123)	(0.1288)	(0.1222)	(0.1600)	(0.0669)	(0.0662)	(0.0721)	(0.1002)
0.5	L_4	0.1468	0.0726	0.0198	0.0142	0.1566	0.0820	0.0221	0.0151
		(0.0806)	(0.0596)	(0.0329)	(0.0255)	(0.0481)	(0.0375)	(0.0207)	(0.0159)
	$L_{\varnothing}$	0.6865	0.0527	0.0227	−0.0004	0.6773	0.0767	0.0263	0.0026
		(0.1169)	(0.0686)	(0.0129)	(0.0175)	(0.0666)	(0.0488)	(0.0079)	(0.0167)
	L_{full}	0.5052	0.4993	0.5016	0.4895	0.5085	0.5030	0.5083	0.4943
		(0.0938)	(0.1045)	(0.1134)	(0.1337)	(0.0584)	(0.0553)	(0.0537)	(0.0828)
0.7	L_4	0.2390	0.1340	0.0431	0.0335	0.2519	0.1485	0.0491	0.0336
		(0.0829)	(0.0708)	(0.0390)	(0.0308)	(0.0506)	(0.0441)	(0.0254)	(0.0188)
	$L_{\varnothing}$	0.8213	0.2021	0.0294	0.0074	0.8189	0.1545	0.0303	0.0110
		(0.0562)	(0.2624)	(0.0138)	(0.0194)	(0.0349)	(0.1128)	(0.0061)	(0.0131)
	L_{full}	0.7008	0.6998	0.7045	0.6960	0.7064	0.7022	0.7047	0.7013
		(0.0770)	(0.0743)	(0.0720)	(0.0791)	(0.0444)	(0.0392)	(0.0365)	(0.0475)
0.9	L_4	0.4077	0.2668	0.1172	0.0981	0.4253	0.2900	0.1286	0.0949
		(0.0928)	(0.0959)	(0.0686)	(0.0597)	(0.0553)	(0.0599)	(0.0417)	(0.0340)
	$L_{\varnothing}$	0.9363	0.9331	0.4578	0.2821	0.9351	0.9365	0.5099	0.1879
		(0.0173)	(0.0363)	(0.4545)	(0.4202)	(0.0099)	(0.0099)	(0.4465)	(0.3553)
	L_{full}	0.9007	0.8994	0.9036	0.9009	0.9020	0.9006	0.9014	0.9005
		(0.0294)	(0.0258)	(0.0228)	(0.0247)	(0.0171)	(0.0144)	(0.0127)	(0.0159)

Table 4.1 – Mean (and standard deviation) of the symbolic maximum likelihood estimate of the correlation, ρ, over $T = 100$ replicate bivariate random rectangle datasets. The symbolic datasets vary in the number of symbols (m), the number of classical datapoints per symbol (n_c), and the strength of the correlation between the two variables (ρ_0). Estimates maximise the three symbolic likelihoods L_{full}, $L_{\varnothing}$ and L_4.

Table 4.1 reports the mean and standard deviation of $\hat{\rho}$ over the replicate datasets under each symbolic likelihood. The marginal parameters (μ, σ_1 and σ_2) are well estimated in each case (see Supplementary Information B.2.1). The main conclusion to be drawn from Table 4.1 is that only the likelihood, L_{full}, that incorporates full information of the number and location of the unique points that define the random rectangle, is able to accurately estimate the dependence ρ between the variables. For L_4 and $L_{\varnothing}$ the mle's are either zero (no dependence can be estimated) or they are biased upwards. Note that for L_{full} the variability of the mle increases slightly as n_c increases, and is also more variable for lower correlation values. This can be explained as the dependence information is contained in the proportion of rectangles which are constructed from 2 and 3 unique points (and their locations). For a fixed correlation, as n_c gets larger it is increasingly likely that the random rectangles will be generated by 4 unique points, and thereby weakening the dependence information that the sample of random rectangles can contain. This weakening naturally occurs more slowly for higher correlations, and so the correlation mle has greater precision for stronger dependence. This insight identifies clear limits on the dependence information content that this interval construction can possess.

Given the statistical inefficiency of intervals constructed from minima and maxima (Figure 4.3), and the informational limits of these intervals as discussed above, a sensible alternative is to construct the random rectangles using marginal order statistics (Section 4.2.3.2), which should be robust to these limitations. Given that such intervals constructed from independent marginal quantiles (equations 4.7 and 4.8) will not contain any dependence information, we now examine the performance of the sequential nesting (4.9) and iterative segmentation (4.11) constructions, for which we denote the respective likelihood functions as $L_{\text{sn}}(s;\theta)$ and $L_{\text{is}}(s;\theta)$.

Similarly to before, for each of $T = 100$ replicate datasets, we generate $m = 20$ classes, each constructed from $n_c = 60$, and 300 draws from a bivariate $(d = 2)$ $N_2(\mu_0, \Sigma_0)$ distribution with $\mu_0 = (2,5)^\top$, $[\Sigma_0]_{11} = [\Sigma_0]_{22} = 0.5$ and correlation $\rho_0 = -0.7, 0, 0.7$. The symbols are constructed in four ways: $L_{\text{sn},x}$ using sequential nesting (4.9); $L_{\text{sn},y}$ using sequential nesting but by exchanging the conditioning order of the x and y margins for symbol constuction; $L_{\text{is},x}$ using iterative segmentation (4.11); $L_{\text{is},y}$ using iterative segmentation but by exchanging the conditioning order of the x and y margins for symbol construction.

Table 4.2 reports the mean (and standard deviation) of the elements of $\hat{\Sigma}$ under each experimental setup when $\rho = 0.7$ (results for $\rho = -0.7$ and 0 are in Supplementary Information B.2.3). The standard deviations σ_1 and σ_2 are estimated unbiasedly for any rectangle configuration. However the standard deviations of the estimates are smaller for the components which are conditioned on first in the symbol construction e.g. σ_1 is more precisely estimated by $L_{sn,x}$ and $L_{is,x}$ and σ_2 by $L_{sn,y}$ and $L_{is,y}$. While estimated unbiasedly for all symbol configurations, constructing the intervals using iterative segmentation produces more precise estimates of the correlation ρ than when using sequential nesting. This is likely because iterative segmentation provides more information about the joint upper and lower values of the margins than nested segmentation, which provides stronger information about the centre of the marginal distributions (see Figure B.1). Different axis constructions ($L_{\cdot,x}$ or $L_{\cdot,y}$) have little effect on the estimates in this case, due to the symmetry of the underlying Gaussian distribution. Also, as expected, increasing the amount of data per symbol, n_c, leads to more precise estimates of all parameters.

All of the estimates of ρ are more precise than that obtained using the marginal minima and maxima, which gave a mle standard deviation of 0.0720 (for $n_c = 50, m = 20, \rho_0 = 0.7$ and using L_{full} in Table 4.1).

Similar to the results in Figure 4.3, within any method of symbol construction, the choice of order statistics has an impact on the precision of the mle of all parameters. Clearly there is an interesting optimal symbol design question here to be addressed, that goes beyond the scope of this paper. However, the iterative segmentation approach appears

	Orders (l,u)	$n_c = 60$ σ_1	ρ	σ_2	$n_c = 300$ σ_2	ρ	σ_2
$L_{\text{sn},x}$	$((6,5),(55,35))$	0.4992	0.6933	0.5050	0.4984	0.6772	0.5075
		(0.0019)	(0.0255)	(0.0054)	(0.0004)	(0.0146)	(0.0024)
	$((16,6),(45,24))$	0.4981	0.6402	0.5043	0.4985	0.6739	0.5177
		(0.0021)	(0.0273)	(0.0107)	(0.0005)	(0.0115)	(0.0048)
	$((20,5),(41,16))$	0.4991	0.6396	0.5054	0.4981	0.6451	0.5141
		(0.0027)	(0.0256)	(0.0129)	(0.0006)	(0.0127)	(0.0059)
$L_{\text{sn},y}$	$((5,6),(35,55))$	0.5106	0.6912	0.4974	0.5082	0.6774	0.4998
		(0.0061)	(0.0339)	(0.0016)	(0.0024)	(0.0156)	(0.0004)
	$((6,16),(24,45))$	0.5289	0.6933	0.4986	0.5088	0.6453	0.4994
		(0.0123)	(0.0239)	(0.0021)	(0.0049)	(0.0129)	(0.0004)
	$((5,20),(16,41))$	0.5231	0.6699	0.5004	0.5154	0.6702	0.4992
		(0.0127)	(0.0253)	(0.0024)	(0.0053)	(0.0106)	(0.0005)
$L_{\text{is},x}$	$((6,3),(55,3))$	0.4993	0.7130	0.4900	0.4984	0.7124	0.4932
		(0.0019)	(0.0067)	(0.0037)	(0.0004)	(0.0032)	(0.0019)
	$((16,10),(45,2))$	0.4981	0.7037	0.4806	0.4985	0.7051	0.4866
		(0.0021)	(0.0039)	(0.0064)	(0.0005)	(0.0011)	(0.0025)
	$((20,7),(41,14))$	0.4993	0.7465	0.4871	0.4981	0.7169	0.4979
		(0.0027)	(0.0128)	(0.0037)	(0.0006)	(0.0051)	(0.0013)
$L_{\text{is},y}$	$((3,6),(3,55))$	0.4929	0.7133	0.4975	0.4896	0.7151	0.4998
		(0.0051)	(0.0064)	(0.0016)	(0.0018)	(0.0032)	(0.0004)
	$((10,16),(2,45))$	0.4868	0.7053	0.4986	0.4848	0.7066	0.4993
		(0.0068)	(0.0035)	(0.0021)	(0.0026)	(0.0011)	(0.0004)
	$((7,20),(14,41))$	0.4933	0.7311	0.5004	0.4947	0.7268	0.4993
		(0.0040)	(0.0115)	(0.0023)	(0.0016)	(0.0057)	(0.0005)

Table 4.2 – Mean (and standard deviation) of the symbolic maximum likelihood estimate of σ_1, ρ and σ_2, over $T = 100$ replicate bivariate random rectangle datasets containing $m = 20$ symbols. The symbolic datasets vary in the number of classical datapoints per symbol (n_c), the type of symbol construction (sn = sequential nesting; is = iterative segmentation), which axis is used first in the symbol construction (x or y), and the vectors of lower (l) and upper (u) order statistics used. The true parameter values are $\sigma_{0,1} = \sigma_{0,2} = 0.5$ and $\rho_0 = 0.7$. For $L_{sn,x}$, orders $(l,u) = ((6,5),(55,35))$ mean firstly take the (6,55) lower/upper order statistics on the x-axis, and then take the $(5,35)$ y-order statistic of the remaining $n_c - 12$ observations in the central x range (see Figure B.1, bottom centre panel). For $L_{is,x}$, orders $(l,u) = ((6,3),(55,3))$ mean firstly take the $(6,55)$ lower/upper order statistics on the x-axis, and then take the 3-rd y-order statistic of the remaining 5 observations below the lower x quantile, and the 3-rd y-order statistic of the remaining 5 observations above the upper x order statistic (see Figure B.1, bottom right panel). For $L_{\cdot,y}$ the procedure is the same as for $L_{\cdot,x}$ but starting with the y-quantiles. In this manner, the resulting 3 bivariate intervals for e.g. $L_{sn,x}$ are identical to those for $L_{sn,y}$. The orders shown are for $n_c = 60$. For $n_c = 300$ the utilised orders are multiplied by 5 so that the intervals are directly comparable between sample sizes n_c.

to be more informative for all parameters, for reasons described above. It is likely that there are other random rectangle constructions that would be even more informative.

4.3.3 Analysis of the loan data set

We illustrate the proposed methodology with an analysis of a loan data from the US peer-to-peer lending company LendingClub. The dataset can be retrieved from the Kaggle platform (`https://www.kaggle.com/wendykan/lending-club-loan-data`) and consist, after removing missing values, of 887,373 loans issued through 2007 to 2015. Based on risk and market conditions, each loan has an associated grade ranging from A1 to G5 (35 in total) which defines the interest rate. Grade A1 loans correspond to least risky credits and thus have the lowest interest rate whereas grade G5 are the riskiest ones.

We focus our attention on the borrowers' annual income (in US$) with the intention to develop a deep understanding of its behaviour. In particular the link between loan grade and income is investigated. Performing a statistical analysis on such dataset induces exploding computational challenges as the model complexity increases. We will show the benefits associated with the use of aggregates rather than the entire dataset.

First a logarithmic transformation of the income random variable is applied to be defined on the real line. Then taking advantage of the natural grade grouping, the data are aggregated into 5 bin histograms through (4.13).

Normal and skew-Normal distributions are fitted at the loan grade level using the classical and symbolic observations. Likelihood ratio tests identify the presence of asymmetry in 34 groups at a $\alpha = 0.05$ level of significance, independently of the method considered. Note the sample size of each group ranges from 576 for grade G5 to 56,323 for grade B3 and, for the largest ones, may have had an impact on the p-values of the likelihood ratio test. As a result of this preliminary analysis it is decided to model the log-income of loan borrowers via a hierarchical model. Denote by X the log-income random variable and by $X_i, i = 1, \ldots, 35$ the grade specific variables. When $X_i \sim N(\mu_i, \sigma_i^2)$, the parameters are modelled as

$$\begin{aligned} \mu_i &\sim T(c_0 + c_1 i + c_2 i^2, \tau^2, \nu) \\ \sigma_i^2 &\sim IG(\alpha, \beta), \end{aligned} \tag{4.20}$$

where $T(\mu, \sigma^2, \nu)$ represents the Student-t distribution with mean μ, variance σ^2 and ν degrees of freedom, and $IG(\alpha, \beta)$ the inverse-Gamma distribution with shape α and scale β. A similar model is considered when $X_i \sim SN(\mu_i, \sigma_i^2, \gamma_i)$, for consistency parametrised as mean, variance and coefficient of skewness (see 'cp' parametrisation in Azzalini (2014, Section 3.1.4.)) with the additional $\gamma_i \sim N(\eta, \epsilon)$. First we demonstrate that, using our methodology, symbolic observations can be an efficient surrogate to the full dataset. A comparison with the most popular method in the SDA literature, given in Le-Rademacher and Billard (2011) (denoted LRB), is established to quantify the performance of our approach. There the group mean and variances, μ_i and σ_i^2, correspond to the histogram mean and variances (Le-Rademacher and Billard, 2011, Section 2.3) and modelled through (4.20).

Figure 4.4 presents the fitted group means and variances obtained through the three approaches while assuming the degree of freedom of the Student-t fixed, $\nu = 3$. The solid lines represent the mean of the corresponding parameter with 95% confidence band given by the dashed lines. The grade specific means under the Normal (top row) assumption are well estimated by all three methods, our model providing standard errors only slightly larger than the classical ones while those from the LRB model are about the double. The means under the skew-Normal (bottom row) are not as accurately estimated but still remain, for the majority, within the 95% confidence band. This might be explained by the difference in complexity of both models, an optimisation over 76 parameters is performed when a Normal distribution is assumed whereas there are 114 parameters when assuming a skew-Normal distribution. The right panels highlight the inability of the LRB method

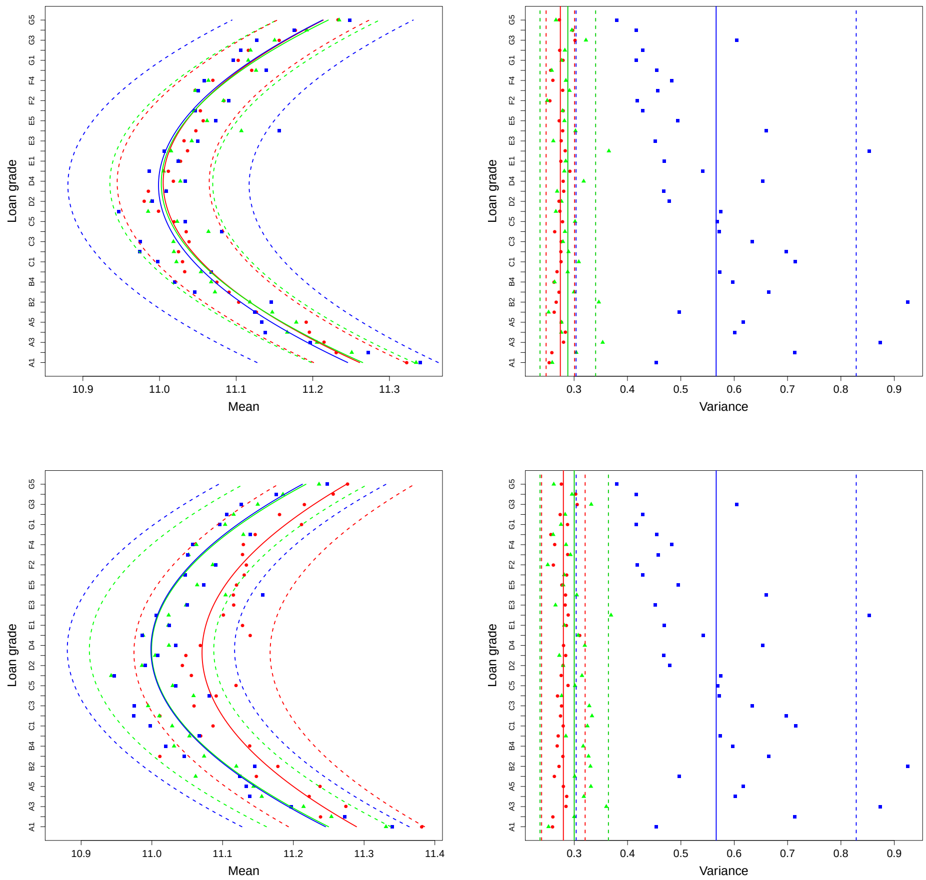

Figure 4.4 – Fitted group means and variances when the underlying distribution is Normal (top) and skew-Normal (bottom) using the classical (red) and symbolic (green) likelihoods and LRB approach (blue).

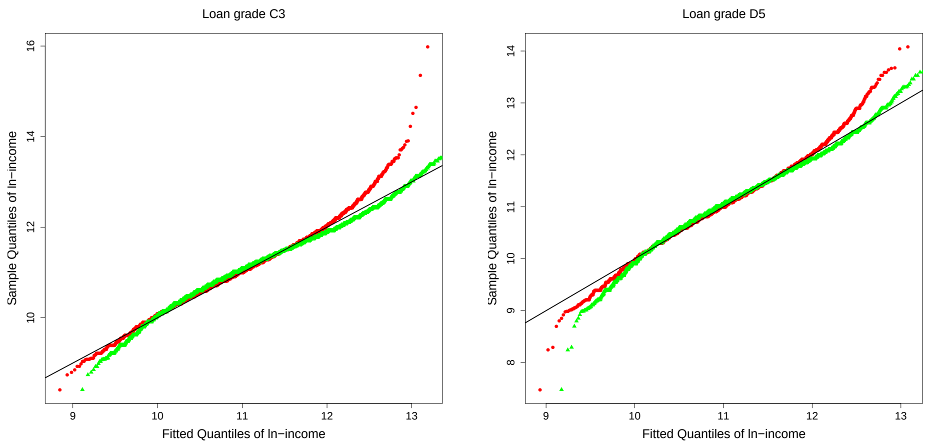

Figure 4.5 – Estimated log-income quantiles using histogram-valued symbols assuming Normal (red) and skew-Normal (green) distribution for loan grade C3 and D5.

to correctly estimate the variances whereas out method performs well.

One of the main advantages of our proposed methodology over the one of Le-Rademacher and Billard (2011) is the ability to make predictions at the data level which allows in the current example to gain insights about the income of borrowers depending on their loan

	Classical	Our method	LRB
N	0.9537	1.7914	0.4379
	(0.0707)	(0.0851)	(0.0014)
SN	144.2014	11.4321	0.4379
	(0.5578)	(0.1534)	(0.0014)

Table 4.3 – Mean (s.e.) evaluation time (ms) of the hierarchical model, based on 1,000 replicates.

grade. Figure 4.5 examines the performance of the estimates of the proposed distributions by grades, through qq-plots of the sample quantiles versus Normal and skew-Normal quantiles (resp. red and green) estimated from histogram-valued symbols. The grades C3 and D5 correspond to large and medium sizes, respectively $n_{\text{C3}} = 50,161$ and $n_{\text{D5}} = 21,389$. Both distributions appear to provide identical results for central values whereas Skew-normal quantiles are closest to the sample quantiles for larger values of the log-income where the Normal quantiles over estimate the sample quantiles. This shows the presences of skewness to the right and reinforces the need for a model that includes asymmetry.

Table 4.3 provides the average time, and standard error, in millisecond of a single evaluation of the hierarchical model obtained from 1,000 randomly generated sets of parameters. It highlights the computational gain of using symbolic based methods when the distribution is skew-Normal. In this scenario, for any symbol type, even though our method is about 20 times slower than the LRB method, it provides an improvement of the computation time by a factor 14 compared to using the full dataset for a comparable quality of fit. Note that for distributions such as the Normal, the likelihood can be written as a function of the sample mean and variance which reduces the computational load but this doesn't apply to the skew-Normal.

In conclusion, in this real data example a hierarchical model was fitted to the log-income of loan borrowers, taking into consideration their loan grade. It was shown that our methodology produces comparable results with those obtained when considering the full dataset. Moreover, in the context of large data, a substantial reduction of the computational times was established for the preferred model. The advantages of the proposed methodology over the commonly used LRB method were clearly demonstrated, particularly for the estimation of grade-level variances.

4.4 Discussion

In this article we have introduced a new framework for the analysis of data that have been summarised into distributional forms. For the general statistical analyst, this method opens up the use of SDA as a broadly applicable statistical technique for analysing large and complex datasets with the potential for large data-storage and computational savings. Within the SDA setting, the fundamentally different approach taken – that of specifying probability models for the data underlying a symbol and deriving the resulting model at the symbolic level, rather than direct model specification at the symbolic level – provides one way to resolve many long-standing methodological weaknesses regarding statistical inference within the field. The resolved problems include the difficulty of specifying meaningful models at the symbolic level, avoidance of the routinely violated uniformity-within-symbols assumption, the ability to perform accurate inference at the level of the underlying data, including model choice, and providing a means to construct and analyse multivariate symbols. As a result, we have been able to expose many weaknesses of current symbol design, and have introduced several new more efficient symbol constructions.

While providing a step forwards, our approach is not without some caveats. Most obviously, the symbolic likelihood function (4.1) requires enumeration of the integral over the underlying data space, which may be problematic in high dimensions. For many standard classes of models, including those considered here, distribution functions $G_{\boldsymbol{X}}(\boldsymbol{x};\theta)$ will be available in closed form. In other cases, numerical or approximation methods may be required, such as quadrature, Monte Carlo techniques (Andrieu and Roberts, 2009), or factorisation of $g_{\boldsymbol{X}}(\boldsymbol{x};\theta)$ to reduce the dimension of the integral.

The symbolic likelihood is clearly an approximation of the classical likelihood as it is based on summary data, and so there will likely be some information loss. While it is possible to approach the accuracy of the classical data model by letting the symbols approach the classical data (e.g. by letting the number of random histograms bins $B \to \infty$), this may not be viable in practice, and in the extreme (e.g. with very large numbers of bins) the computational overheads could exceed that required for the classical data analysis. It is therefore of interest, and the subject of future research, to understand the quality of the approximation. It is possible that some of the theory supporting approximate Bayesian computation (e.g. Sisson et al., 2018), which is also based on computation via summary statistics, could be useful here.

Within this context there is immense scope for optimum symbol design, whereby the symbols are constructed to provide maximal information for a specific analysis or for a family of analyses that may be performed in the future. New symbolic types could also be developed such as Gaussian- or other continuous distribution-based symbols, which may additionally enable direct integration of the integral in (4.1) through conjugacy.

The explosive emergence of the data-rich biome – the *infome* – in which we now reside, since Schweizer (1984)'s 35-year old prediction that "distributions are the numbers of the future", clearly substantiates the potential for symbolic data analysis to become a powerful everyday tool for the statistical analyst. Schweizer (1984)'s future is very much here.

Chapter 5

Bayesian semi-parametric modelling of ultrafine particle number concentration using symbolic data analysis

5.1 Introduction

Ultrafine particles (UFPs), whose diameters are less than 100 nm are ubiquitous in urban air and are acknowledged to have adverse risk to climate, visibility and human health (?). Due to their negligible mass compared with larger-sized particles (such as PM_{10} and $PM_{2.5}$), UFPs are commonly evaluated through measurements of particle number concentration (PNC) (Harrison et al., 2000; Kumar et al., 2010). UFPs are known to undergo physical and chemical transformation which affect their number and size distributions, which contribute significantly to temporal (Shah et al., 2008) and spatial variability (Heal et al., 2012). Understanding this variability is key to quantifying human exposure and designing effective motoring strategies.

In this chapter, we concentrate our attention on developing a flexible statistical model that is able to uncover the dynamic temporal evolution of PNC at a given location. To be more specific, we focus exclusively on modelling and forecasting temporal variability of PNC collected as part of a measurement campaign by the International Laboratory for Air Quality and Health (ILAQH) entitled "Ultrafine Particle Emissions from Traffic and Child Health" (UPTECH). The UPTECH study followed a spilt panel design which involved the short-term measurement of PNC at each of 25 government primary schools in the Brisbane Metropolitan Area along with 3 long-term monitoring sites. To assist the better understanding of aerosol dynamic processes, a PNC measurement was taken every five minutes (resulting in 12 observations per hour) continuously over a 2-week period at each primary school.

It is well-known that continuously measured time series may possess characteristics such as regular temporal trends and non-linear dependence on covariates (and interactions thereof) (Clifford et al., 2012b). The data complexity motivates the desire to search for a flexible regression model that is capable of capturing these features without specifying the functional form of the relationship *a priori*. Splines are commonly used for non- and semi-parametric modelling smooth curves and surfaces (Silverman, 1985), time series (Wahba et al., 1990) and non-linear covariate effects (Lin and Zhang, 1999) due to their simple to construct bases (De Boor et al., 1978). Therefore, we adopt B-splines to approximate the underlying temporal effect which is assumed to be a smooth function over time and cannot be directly observed from the data. The flexibility of B-splines is further enhanced by incorporating a set of prior beliefs to express one's uncertainty about how smooth the fitted functions should be (Lang and Brezger, 2004).

PNC in Brisbane have been shown to exhibit daily and weekly trends, which may be additive functions of the hour of the day and day of the week or some joint, non-separable functions (Morawska et al., 2002; Mejia et al., 2007). As a result, we derive a covariate to represent the interaction between daily and weekly temporal effects. It is assumed that this joint daily-weekly temporal effect varies hour-by-hour but has a periodic pattern which repeats weekly. To ensure the smoothness, a cyclic random walk prior on the precision matrix of the covariates is adopted, leading to similarity in successive covariates in the random walk model (Lang and Brezger, 2004; Rue and Held, 2005; Rue et al., 2009). In addition to a periodic joint daily-weekly temporal trend, there might be a slowly-changing annual trend (day of the year effect) which is modelled by a cyclic B-spline with a smoothing penalty on the prior distribution of the B-spline coefficients.

The above approaches are extracted from a previous analysis performed by Clifford et al. (2012a) where the author developed a Bayesian semi-parametric additive model for the log of particle number concentration (log(PNC), henceforth) in Brisbane. However, measurements were aggregated to every 5 minutes for previous analyses and then to hourly for spatio-temporal modelling in Clifford et al. (2012a). The analysis of Clifford et al. (2012a) is fine for looking at average relationships but does not well characterise the uncertainty in the hourly levels of PNC. Recent studies by Kumar et al. (2011) established an association between excess mortality and human exposure to traffic-derived UFPs in urban areas. Given their adverse impacts on human health, it would be desirable to obtain a richer understanding by modelling and predicting from the full distribution of PNC. This requires a model that takes into account not only the mean but also the variance of the response. However, it is recognised that analysing the full data is challenging. An appealing alternative is to consider a sufficient representation of the data. This motivates representing log(PNC) observations as histogram-valued symbolic data as defined in the symbolic data analysis literature (see e.g., (Billard and Diday, 2006; Billard, 2011; Noirhomme-Fraiture and Brito, 2011) for a comprehensive introduction to symbolic data). The idea here is to aggregate the underlying data consists of individual 5-minute log(PNC) observations into hourly histograms according the hour when it was measured. Then we use the symbolic likelihood function for histogram-valued data with random bins to estimate the parameters associated with the underlying data distribution as defined by Beranger et al. (2018).

The construction of the symbolic likelihood function requires a specification of the distribution underlying the observations within symbols. In this situation, it would be a distribution that can adequately describe the temporal features of all measured log(PNC) observations. Previous analysis by Clifford et al. (2012a) proposed a model with a single Gaussian likelihood. In practice, however, log(PNC) observations are likely to come from heterogeneous sources and thus the resulting distribution will often have multiple modes. Whitby and McMurry (1997); Hussein et al. (2005) and Wraith et al. (2011) proposed to represent particle size distribution at any time point as a set of individual typically normal distributions or modes. Inspired by their approaches, the underlying log(PNC) observations are modelled by a finite Gaussian mixture model. Traditionally, mixture models have been applied in the standard setting where random samples are independent (Marin et al., 2005). Equivalent mixture models have also been developed for data that are spatially and/or temporally correlated (Dunson, 2006; Alston et al., 2007; Caron et al., 2012; Ji, 2009; Fernández and Green, 2002; Green and Richardson, 2002). For the PNC example considered in this chapter, given they were collected regularly and frequently, it is likely that parameters of the mixture model at each time point are correlated with neighbouring time points. In addition, it is also of interest to study how the underlying distribution evolves over time. As a result, the mixture model incorporates time-varying mixture locations and mixing weights with time-invariant mixture scales to model the dynamic processes of log(PNC) over time.

The time-varying mixture locations capturing the temporal effects are modelled as described above by two temporal components: a joint daily-weekly temporal effect and a slowly-changing day of the year effect. To allow for temporal correlation in the for mixture weights, periodic B-splines are used with a first-order random walk prior imposed on the coefficients. This prior penalises large changes in subsequent mixture weights, ensuring a smooth transition in mixture weights. Although the underlying data are dependent over time, in this case, we adopt a simplifying assumption that given the time-correlated parameters of the assumed Gaussian mixture model for the underlying data, the hourly histogram-valued observations are conditionally independent from each other. As a result, we use the symbolic likelihood for independent histogram-valued data.

The rest of the chapter is organised as follows. In Section 5.2, we review the existing methods of estimating log(PNC) using a Bayesian spline-based semi-parametric regression model proposed by Clifford et al. (2012a). We then discuss the potential drawback of the existing method which motivates the proposal of the current model construction. Next, in Section 5.3 we propose the equivalent symbolic version of the proposed models. The method is applied to a number of simulations in order to show its efficacy in Section 5.4 and to the PNC data in Brisbane, Australia, to demonstrate its use in tackling real world problems in Section 5.5. Finally, we discuss some interesting aspects of the model in light of the simulations and real data analysis and conclude the chapter in Section 5.6.

We use the following notation throughout the paper.

- $t = 1:336$ is the time index, denoting the hour in a 2-week period (i.e.$t = 1:24 \equiv$ day 1; $t = 25:48 \equiv$ day 2..., $t = 145:168 \equiv$ day 7..., $t = 313:336 \equiv$ day 14). T=336 is the end of the time index.

- y_{it} is the observed log(PNC) at the i^{th} 5-minute interval within an hour t.

- α is the overall temporal mean in the 1-component Gaussian mixture model.
  α_1 and α_2 are the corresponding mixture locations in the 2-component Gaussian mixture model.

- β_t represents the marginal daily-weekly temporal effects. This temporal covariate is a vector consisting of 168 terms representing hour of the week. For example, $\beta_1 : \beta_{24}$ are 24 hours on Monday, $\beta_{25} : \beta_{48}$ are 24 hours on Tuesday...$\beta_{145} : \beta_{168}$ are 24 hours on Sunday.
  $\beta_t = (\beta_{t1}, \beta_{t2})$ is the corresponding mixture marginal daily-weekly effects in the 2-component Gaussian mixture model.
  The model assumes that the temporal effect β_t in the 1-component model or (β_{t1}, β_{t2}) in the 2-component model has a periodic pattern that repeats weekly. In addition, given the model explicitly models α, the mean temporal level of PNC, we then constrain $\sum_{t=1}^{T} \beta_t = 0$.

- B is a B-spline basis matrix.

- $\boldsymbol{\theta}$ is a vector containing the coefficients of a B-spline basis functions

- Q is the chosen number of ordered quantiles for each histogram bin interval.

- s_t is a vector containing Q order quantiles selected from log(PNC) observed in the t^{th} hour.

- λ_t is the mixing weights of the 2-component mixture model.

- ζ_t is the logarithm of the odds $\frac{\lambda_t}{1-\lambda_t}$.

- σ is the standard deviation associated with a Gaussian model.

- μ_1 and μ_2 are mixture locations in a general 2-component mixture model where $\mu_2 > \mu_1$ to address identifiability issue in finite mixture model.

- D represents a data matrix.

5.2 Construction of the classical data model

5.2.1 The existing Bayesian semi-parametric additive model with a Gaussian likelihood fitted with hourly averaged log(PNC)

As discussed briefly in Section 5.1, the existing method proposed by Clifford et al. (2012a) only models on the hourly averaged log(PNC), though there are 12 observations per hour.

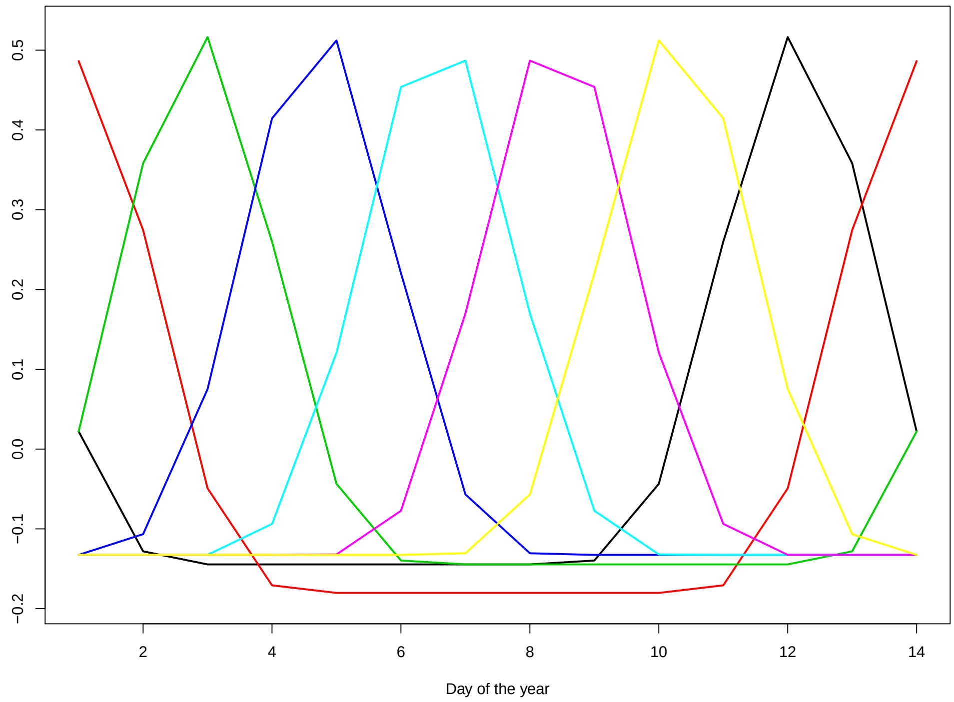

Figure 5.1 – Seven zero mean periodic cubic B-spline basis functions for estimating a smooth function of day of the year, currently only explicitly modelling 14 days.

The model is specified as follows:

$$\begin{aligned} \bar{y}_t &= \alpha + \beta_t + (B\boldsymbol{\theta})_t + \epsilon_t \\ E(\bar{y}_t) &= \alpha + \beta_t + (B\boldsymbol{\theta})_t \\ \epsilon_t &\sim \mathcal{N}(0, \sigma^2) \end{aligned} \tag{5.1}$$

with the hourly averaged data $\bar{y}_t = \frac{\sum_{i=1}^{12} y_{it}}{12}$.

To ensure identifiability, $\sum_{t=1}^{T} \beta_{t,j} = 0$, and $\sum_{t=1}^{T} (B\theta)_t = 0$ are constrained to sum to zero. For the latter term, the B-spline basis matrix B for modelling the slowly-changing marginal annual effects (day of the year) is chosen to be the same as Clifford et al. (2012a). It is constructed with a cubic B-spline from a recursive algorithm (Eilers and Marx, 1996) defined over a grid of ten knots, yielding seven cubic B-spline basis vectors. Given the current data set, we only explicitly model 14 days in a year; the resultant basis function is visualised in Figure 5.1. Each column in this basis matrix is constrained to sum to zero by subtracting each column element from its corresponding column mean. In addition, the B-spline coefficients $\boldsymbol{\theta}$ are also constrained to sum to zero.

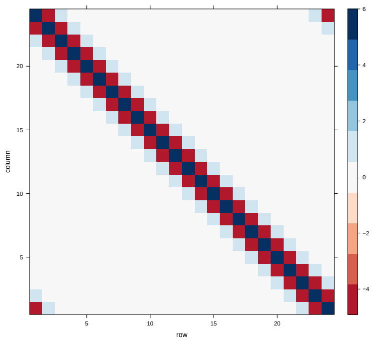

Figure 5.2 – A second-order cyclic random walk penalty matrix of dimensions 24×24 for hour of the day effect. Each square represents a value in this penalty matrix with colours representing different values.

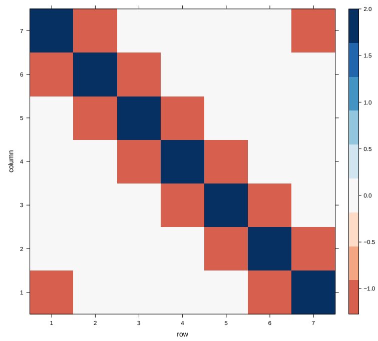

Figure 5.3 – A first-order cyclic random walk penalty matrix of dimensions 7×7 for day of the week effect. Each square represents a value in this penalty matrix with colours representing different values.

The prior distributions for the model parameters are specified as:

$$\begin{aligned}
\alpha \quad & \sim Cauchy(0, 10) \\
\beta|\sigma \quad & \sim Multivariate\ Normal(0, \sigma^2 \times K^{-1}) \\
\sigma \quad & \sim Cauchy(0, 2.5) \\
\boldsymbol{\theta}|\tau_{\boldsymbol{\theta}} \quad & \sim Multivariate\ Normal(0, (\tau_{\boldsymbol{\theta}})^{-1} I) \\
\tau_{\boldsymbol{\theta}} \quad & \sim Gamma(1, 0.05)
\end{aligned}$$

The prior for the overall temporal effect α is modelled by a weakly informative Cauchy distribution as suggested by Gelman et al. (2008). The prior for the derived marginal joint daily-weekly β is a multivariate Gaussian with a customised penalty precision matrix K, which can be visualised in Figure 5.4. It is obtained as a Kronecker product of a second order penalty matrix for hour of the day (Figure 5.2) and a first order penalty matrix for day of the week (Figure 5.3, Marx and Eilers (2005)). In this case, the matrix in Figure 5.2 is equivalent to a periodic version of a second order random walk model and it is chosen to yield smooth estimates of daily trend. On the other hand, the matrix in Figure 5.3 is equivalent to a first order cyclic random walk assuming that while there is day-to-day variation, the mean level on Wednesday for example, is only related to Monday through the mean on Tuesday. The model scale parameter σ serves as a penalty parameter which controls the degree of smoothness. For $\sigma \to \infty$, there is no smoothing which is equivalent to assuming independence among successive hour of the week covariates. The degree of smoothness increases with decreasing σ and in the extreme case where $\sigma = 0$, $\beta_t = \beta'$, for $t = 1...T$ and a constant β'. To ensure a proper prior, a small value (e.g. 0.00001) is added to the diagonal elements of the matrix K. The coefficients of the cyclic B-spline basis matrix $\boldsymbol{\theta}$ are assigned a weakly informative Gaussian prior with the precision parameter $\tau_{\boldsymbol{\theta}}$ given a weakly informative Gamma prior.

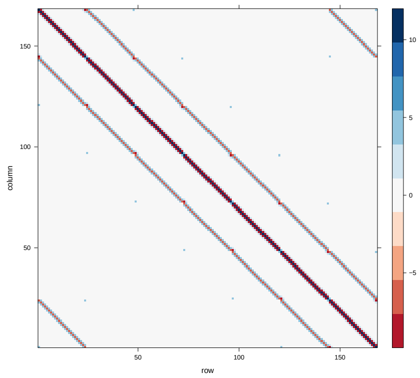

Figure 5.4 – A168 × 168 dimensional Kronecker product of the hour of the day penalty matrix (Figure 5.2) and the day of the week penalty matrix (Figure 5.3). Each square represents a value in this penalty matrix with colours representing different values.

5.2.2 The existing Bayesian semi-parametric additive model with a Gaussian likelihood fitted with individual log(PNC)

Given there are only 12 observations per hour, the same model structure defined in Section 5.2.1 in Equation (5.1) could have been fitted on y_{it} instead of $\bar{y}$.

A similar model that takes into account all of the observations within an hour can be specified as:

$$\begin{aligned} y_{it} &= \alpha + \beta_t + (B \times \theta)_t + \epsilon_{it} \\ E(y_{it}) &= \alpha + \beta_t + (B \times \theta)_t \\ \epsilon_{it} &\sim \mathcal{N}(0, \sigma^{2*}) \end{aligned} \tag{5.2}$$

Where $\sigma^{2*} = \frac{\sigma^2}{12}$. With the same prior distributions as in Equation (5.1). This is the parsimonious 1-component Gaussian model fitted with underlying classical data, that is a data matrix of size 12×336, containing 12 log(PNC) observations per hour over 336 hours (a 2-week period). The assumption here is that a Gaussian describes the distribution of the underlying data, which may not be true in practice.

5.2.3 The proposed Bayesian semi-parametric additive model with a finite mixture Gaussian likelihood

One of the advantages of treating log(PNC) observations as coming from a finite mixture model is that it accounts for heterogeneity and multimodality in the underlying distribution. In a standard setting in which random samples are independent, a 2-component Gaussian mixture model can be specified as follows:

$$\begin{aligned} p(y_{it}|\mu_{t1}, \mu_{t2}, \lambda_t, \sigma) &\sim \lambda_t \mathcal{N}(\mu_{t1}, \sigma) + (1 - \lambda_t)\mathcal{N}(\mu_{t2}, \sigma) \\ \mu_{t1} &= \alpha_1 + \beta_{t1} + (B\boldsymbol{\theta})_t \\ \mu_{t2} &= \alpha_2 + \beta_{t2} + (B\boldsymbol{\theta})_t \end{aligned} \tag{5.3}$$

Mixture models are known to suffer from inherent combinatorial non-identifiability when data generating processes are degenerate (Betancourt, 2017). To ensure identifiability of the mixture model, an ordering constraint on $\mu_{t2} > \mu_{t1}$ for $t = 1...T$ is adopted (Wraith et al., 2011). Time-varying mixture locations μ_{t1} and μ_{t2} are decomposed to component-wise time invariant intercept terms α_1 and α_2 to account for the overall mean per mixture component along with component-wise time-varying marginal daily-weekly trends β_{t1} and β_{t2} respectively. In addition, the marginal annual trends $B\boldsymbol{\theta}$ are shared by both mixture components with a time-invariant scale parameter σ shared between mixture components.

Given log(PNC) observations are temporally correlated, parameters of the mixture model at each time point are also likely to be correlated with the neighbouring time points. Therefore, it is important to consider the correlated nature of the parameters in a mixture model setting. Popular approaches that account for the dependence association of the mixture parameters, both within and across periods, include Dependent Dirichlet Process mixture models (DDPM) and (spatial) dynamic factor models (SDFM) (MacEachern, 2001; Dunson, 2006; Caron et al., 2012; Ji, 2009; Strickland et al., 2008). However, as discussed by Wraith et al. (2011) the DDPM assumes a non-parametric process and thus provides a less intuitive interpretation of the mixture parameters, while a successful implementation of SDFM generally requires a relatively long time series. As a result, Wraith et al. (2011) proposed four types of temporal prior to link mixture parameters $(\mu_{t1}, \mu_{t2}, \lambda_t)$ over time. The first type is the independent prior where the correlated nature of the data is ignored completely. This prior is commonly used in the conventional mixture model setting where observations are independent random samples. The second, third and fourth are termed the "informed prior", "penalised prior" and "hierarchical informed prior". The "penalised prior" on λ at time t incorporates information over all the past time periods. The idea is based on Gustafson and Walker (2003)'s proposal to use an independent prior in conjunction with a penalty term, penalising large changes in probabilities in neighbouring time periods. In the case of a 2-component mixture model, the mixing weights are $(\lambda, 1-\lambda)$ and a standard independent prior for this 2-dimensional simplex is a beta distribution. Its "penalised" reparametrisation version can be specified:

$$p(\lambda|\alpha_\lambda, \beta_\lambda) \quad \propto Beta(\alpha_\lambda, \beta_\lambda)\exp(-\frac{1}{\sigma_\lambda}\sum_{t=1}^{T}||\lambda_t - \lambda_{t-1}||^2)$$

where σ_λ serves as a penalty term with smaller values indicating greater smoothing. $\alpha_\lambda, \beta_\lambda$ and σ_λ are assigned weakly informative hyperpriors. Prior distributions assigned directly on α_λ and β_λ are less intuitive and thus they are reparameterised into ϕ and η representing a prior mean and prior counts on λ. The prior distributions for these hyperparameters are:

$$\begin{aligned}
\phi &\sim \textit{Unif}(0,1) \\
\eta &\sim \textit{Gamma}(1,0.05) \\
\sigma_\lambda &\sim \textit{Cauchy}(0,2.5)
\end{aligned}$$

where $\phi = \alpha_\lambda + \beta_\lambda$ and $\eta = \frac{\alpha_\lambda}{\alpha_\lambda+\beta_\lambda}$.

It is assumed that the mixing weights have the same temporal pattern as β_t, that is the mixing weights vary hour-by-hour within a week and repeat for all weeks in a year. Using the above construction for λ_t, that there are 168 parameters to be explicitly modelled. Given a small sample size within an hour (only 12 observations), so λ_t might not be adequately modelled. It is not unreasonable to assume that the mixing weights change smoothly and slowly over time. As a result, an alternative way of modelling time-varying mixing weights is by a univariate spline with a basis of six second order cyclic B-splines denoted as B_λ. To alleviate the restricted range of $\lambda_t \in (0,1)$ for all t, we decide to model on $\zeta_t = log(\frac{\lambda_t}{1-\lambda_t}) \in \mathbb{R}$. Given B-splines are local bases that form the splines for the *log* odds of the mixing weights λ, if the coefficients of nearby B-splines are close to each other then there will be less local variability in the resulting ζ_t. This motivates the use of priors to enforce smoothness across the coefficients, β_ζ. As a result, a first order random walk prior is adopted for the B-spline coefficients β_ζ,

$$\begin{aligned}
\beta_{\zeta,i} &= \beta_{\zeta,i-1} + \tau_{\zeta,i} \\
\beta_{\zeta,1} &\sim \mathcal{N}(0,1) \\
\beta_{\zeta,i} &\sim \mathcal{N}(\beta_{\zeta,i-1}, \tau_\zeta) \\
\tau_{\zeta,i} &\sim \mathcal{N}(0,1)
\end{aligned}$$

where $i = 2:6$. The initial value $\beta_{\zeta,1}$ is given a weakly informative prior centred at 0, implying that the mixing weight in the first period is centred at 0.5. The Gaussian error $\tau_{\zeta,i}$ is assigned a weakly informative prior to obtain a proper posterior for β_ζ. In this way, only 6 B-spline coefficients for ζ need to be explicitly modelled and overcome the problem of data insufficiency.

The time-dependence structure for the mixture locations $\mu_t = (\mu_{t1}, \mu_{t2})$ is explicitly modelled by the overall mean trend $\alpha = (\alpha_1, \alpha_2)$, the marginal daily-weekly $\beta = (\beta_{t1}, \beta_{t2})$ and a common marginal annual trend $B\theta$ with specific prior distributions assigned below to ensure smoothness over time. In addition, the scale parameter σ is assumed to be time-invariant and is assigned a weakly informative Cauchy prior.

$$
\begin{aligned}
\alpha_1, \alpha_2 \quad & \sim Cauchy(0, 10) \\
\beta_{t1}, \beta_{t2} | \sigma \quad & \sim \textit{Multivariate Normal}(0, \sigma^2 \times K^{-1}) \\
\sigma \quad & \sim Cauchy(0, 2.5) \\
\boldsymbol{\theta} | \tau_{\boldsymbol{\theta}} \quad & \sim \textit{Multivariate Normal}(0, (\tau_{\boldsymbol{\theta}})^{-1} I) \\
\tau_{\boldsymbol{\theta}} \quad & \sim Gamma(1, 0.05)
\end{aligned}
$$

5.3 The motivation of Symbolic Data Analysis

One of the main disadvantages of Equation (5.1) is that it only uses part of the whole data information – the hourly averaged log(PNC) observations $\bar{y}_t$. As a result, it will also be unable to construct predictions on the level of the individual 5-minute measurements (only their mean). Based on Equation (5.2), the whole data matrix containing individual data point y_{it} has to be fitted. Computation of intensity increases with increasing matrix size and the complexity of the model. As both the data matrix and the number of mixture components representing the underlying structure increase, fitting the model using Stan Team (2016) means that it would have to evaluate likelihood function at every single data point as there are no sufficient statistics. Consequently, it would lead to excessive long computation time and costs. As a result, we are in search of a model construction that is able to handle data of potentially large size and fit a practical model. SDA offers a solution to big and complex data challenges as big data can be reduced and summarised by "classes". In this case, we aggregate individual log(PNC) according to the hour (the "class") it was measured to become hourly histogram-valued symbolic data. These symbolic data are constructed from 5-quantile (order statistics) of the individual log(PNC) in that hour. To complete the construction of symbolic likelihood functions for histogram-valued data with random bins , we assume the underlying data are likely to come from 2 different Gaussian distributions as described in Equation (5.3). In the following Section 5.3.1, we outline the estimation of the parameters associated with the underlying data distribution using a symbolic likelihood function for univariate histogram-valued data.

5.3.1 Estimating the parameters from the proposed model using symbolic likelihood function

As discussed in Section 5.1 and also in Section 5.3, it may be preferable to utilise the information contained in the full observed data matrix rather than just using the hourly mean values. It is reasonable to assume that successive log(PNC) observations within a given hour that are only 5-minutes apart possess similar physical and chemical properties. In other words, there is not much change in their dynamics between successive observations in any given hour. As we are interested in modelling the dynamic systems of PNC over time while incorporating the entire observed data matrix ($D = 12 \times 336$, as there are 12 observations per hour over a 2-week of 336 hours) then SDA offers a way to aggregate the underlying hourly log(PNC) data to hourly histograms constructed from quantiles of the underlying data.

The idea is to first order 12 5-minute log(PNC) observations and then choose $Q \in \{1, ..., 12\}$ order-statistics based quantiles. For illustrative purposes, the following example chooses 5 quantiles based on order statistics ($Q = 5$). In this case, the $(1^{st}, 4^{th}, 7^{th}, 10^{th}, 12^{th})$ ordered observations will be chosen from the t^{th} column in log(PNC) data matrix and labelled as $s_t = (s_{1t}, s_{2t}, s_{3t}, s_{4t}, s_{5t})$. Further, assuming the underlying log(PNC) data come from the model described in Equation (5.3), the task then changes to estimate the model parameters using the hourly histogram-valued log(PNC) represented by the order statistic based quantiles $s_t = (s_{1t}, s_{2t}, s_{3t}, s_{4t}, s_{5t})$ rather than using the individual log(PNC) observations y_{it}.

According to the symbolic likelihood function for histograms by Beranger et al. (2018) where bins are random and constructed from 5 quantiles at each time point is:

$$\begin{aligned}\mathcal{L}(s_t;\kappa) \quad \propto \quad & \frac{12!}{3 \times 2! \times 1!} g_y(s_{1t};\kappa) g_y(s_{2t};\kappa) g_y(s_{3t};\kappa) g_y(s_{4t};\kappa) g_y(s_{5t};\kappa) \\ & (G_Y(s_{5t};\kappa) - G_Y(s_{4t};\kappa))(G_Y(s_{4t};\kappa) - G_Y(s_{3t};\kappa))^2 (G_Y(s_{3t};\kappa) - G_Y(s_{2t};\kappa))^2 \\ & (G_Y(s_{2t};\kappa) - G_Y(s_{1t};\kappa))^2 \end{aligned} \tag{5.4}$$

where G_y and g_y are the c.d.f. and p.d.f. of the modelled distribution for the underlying classical data, in this example, it is a 2-component Gaussian mixture model defined in Equation (5.3). $\kappa = (\alpha_1, \alpha_2, \beta, \lambda, \theta, \sigma)$ is a vector containing all parameters in the mixture model.

Assuming independence between successive hourly histograms, conditioning on the mixture parameters of the underlying model, the overall likelihood function for a time series of histograms becomes:

$$\begin{aligned}\mathcal{L}(s_1, ..., s_T;\kappa) = \Pi_{t=1}^{T} \mathcal{L}(s_t;\kappa) \quad \propto \quad & \Pi_{t=1}^{T} [\frac{12!}{3 \times 2! \times 1!} g_y(s_{1t};\kappa) g_y(s_{2t};\kappa) g_y(s_{3t};\kappa) g_y(s_{4t};\kappa) g_y(s_{5t};\kappa) \\ & (G_Y(s_{5t};\kappa) - G_Y(s_{4t};\kappa))(G_Y(s_{4t};\kappa) - G_Y(s_{3t};\kappa))^2 \\ & (G_Y(s_{3t};\kappa) - G_Y(s_{2t};\kappa))^2 (G_Y(s_{2t};\kappa) - G_Y(s_{1t};\kappa))^2]. \end{aligned}$$

Since the underlying data are assumed to come from the mixture model defined in Equation (5.3), the interval probability for example $(G_Y(s_{4t};\kappa) - G_Y(s_{3t};\kappa))$ in the above likelihood function can be explicitly written as:

$$\begin{aligned}(G_Y(s_{4t};\kappa) - G_Y(s_{3t};\kappa)) \quad = \quad & (\lambda_t \times \Phi(s_{4t};\alpha_1 + \beta_{t1} + (B\boldsymbol{\theta})_t, \sigma) + (1 - \lambda_t) \times \Phi(s_{4t};\alpha_2 + \beta_{t2} + (B\boldsymbol{\theta})_t, \sigma)) \\ - \quad & (\lambda_t \times \Phi(s_{3t};\alpha_1 + \beta_{t1} + (B\boldsymbol{\theta})_t, \sigma) + (1 - \lambda_t) \times \Phi(s_{3t};\alpha_2 + \beta_{t2} + (B\boldsymbol{\theta})_t, \sigma)). \end{aligned}$$

Similarly, the density function $g_y(s_{4t};\kappa)$ can be written as:

$$g_y(s_{4t};\kappa)) \quad = \quad \lambda_t \times \phi(s_{4t};\alpha_1 + \beta_{t1} + (B\boldsymbol{\theta})_t, \sigma) + (1 - \lambda_t) \times \phi(s_{4t};\alpha_2 + \beta_{t2} + (B\boldsymbol{\theta})_t, \sigma).$$

The prior distributions assigned to this vector are those described in Section 5.2.3.

5.4 Simulation

To check that the proposed model can capture the dynamic evolution of PNC, some data are simulated from the model. We then check to see if we can recover the (known) model parameters.

To mimic the real data set, all of the following simulation studies are set up to have the same data structure as the real data. To be more specific, we simulate 12 observations per hour over a 2-week period. In other words, the classical data matrix D for y_{it} is of dimension 12×336, totalling 4032 numbers of classical observations.

All the models are fitted using Stan (Team, 2016) in R, running 4 chains with 2000 iterations each. The first 500 are considered as burn-in and therefore discarded. The Rhat values, the effective sample size, and the traceplots of the model parameters are checked to ensure the model has converged and is reliable.

5.4.1 Handling missing data

In the real data, there are 103 missing observations and thus the equivalent data are removed from the simulated data matrix accordingly. In summary, there are 3929 observations recorded over a 2-week period.

We adopt the same assumption as in the previous model of Clifford et al. (2012a) that the observations are missing completely at random. For illustrative purposes, assume that a part of the data matrix looks like this:

$$D = \begin{bmatrix} 1 & 2.3 & NA & 4.1 \\ 3.1 & NA & NA & 1.15 \\ NA & 2.2 & NA & NA \end{bmatrix}$$

Where each cell in the above matrix represents a log(PNC) value- y_{it}. The coding of the matrix D will no longer work as there is no support within Stan Team (2016) for R's NA values, so this data structure cannot be used directly. Instead, the matrix is converted to a "long form" as described in Wickham et al. (2014), with columns indicating the j and k indexes along with the value, shown in Table 5.1. This says that $y_{1,1} = 1, y_{1,2} = 2.3$, and so on, up to $y_{3,2} = 2.2$, with all other entries undefined and not modelled. Therefore, Table 5.1 containing individual observation y_{it} is used when fitting 1 or 2-component Gaussian mixture model to the classical data.

A slight modification is made when fitting 1 or 2-component models with 5 quantiles based on order statistics based ($Q = 5$) symbolic data. As described in Section 5.3.1, we choose the $(1^{st}, 4^{th}, 7^{th}, 10^{th}, 12^{th})$ ordered observations from the t^{th} column in the log(PNC) data matrix. If in the t^{th} column there are between 2 and 11 (inclusive) missing observations, then from the remaining non-missing observations, we select minimum, first quantile, median, third quantile and maximum observations. This procedure is carried for all missing columns where there are at least 5 observations and within each select 5 ordered observations roughly corresponding to (minimum, first quantile, median, third quantile, maximum). For columns with no observations, we treat these 5 ordered observations from

j	k	y_{it}
1	1	1
1	2	2.3
1	4	4.1
2	1	3.1
2	4	1.15
3	2	2.2

Table 5.1 – Example of coding data matrix D with "NA" values in Stan Team (2016). The first two columns, j and k, denote the indexes and the final column, y_{it}, the value. For example, the fifth row of the database-like data structure on the right indicates that $y_{2,4} = 1.15$.

each column as 5 ordered parameters that are estimated from the posterior predictive distribution.

5.4.2 Fitting 1-component Gaussian model

5.4.2.1 Data setup

Firstly, we simulate data from a single Gaussian model, acknowledging that this model may under-represent the type of behaviour of aerosol particle number concentration observed in the UPTECH project.

$$\begin{aligned}
y_{it} &= \alpha + \beta_t + (B\boldsymbol{\theta})_t + \epsilon_t \\
\alpha &= 9 \\
\beta_{1:168} &= \frac{\sin(\frac{2\pi h}{168})}{2}, h = 1..., 168 \\
\boldsymbol{\theta} &= (\theta_1, \theta_2, ...\theta_7) \sim N(0,1) \\
\epsilon_t &\sim N(0, 0.1)
\end{aligned}$$

for $i = 1..., 12, t = 1..., 336$. Notice that only 168 terms of marginal daily-weekly effects β_t are simulated, because we assume that these effects repeat for all weeks in a year and thus they are the same for the second week. 103 observations are removed from the simulated dataset.

5.4.2.2 Results

Given the simulated data, we fit a 1-component Gaussian model to estimate time-correlated model parameters (α, β, θ) and time-independent scale parameter σ using the classical data likelihood function in Equation (5.2) and symbolic likelihood function in Equation (5.3.1) respectively.

Figure 5.5 shows 8 randomly chosen (consecutive in time) hourly histogram densities for the simulated log(PNC) observations with the "true" density superimposed (in red). The green line represents the posterior preditive density fitted using individual observations within an hour, while the blue line represents the model fitted with hourly-histograms. It is apparent that the symbolic model follows the classical data model fairly well and so we are satisfied that 5 quantiles are sufficient to fit this model well. This is also true for all the other parameters of the underlying data distribution which can be seen in Figure C.1, Figure C.2 and Figure C.3 in Appendix C.1. Different numbers of quantiles Q were explored with $Q = 5$ representing the lowest value such that the quality of the model fit was compatible to the classical data fit.

5.4.3 2-component Gaussian model

5.4.3.1 Data setup

In this section, with a 2-component Gaussian mixture model, data are simulated from the equation below

$$y_{it} = \begin{cases} \alpha_1 + \beta_{t1} + (B\boldsymbol{\theta})_t + \epsilon_t, & \text{with probability } \lambda_t \\ \alpha_2 + \beta_{t2} + (B\boldsymbol{\theta})_t + \epsilon_t, & \text{with probability } 1 - \lambda_t. \end{cases}$$

$$\begin{aligned}
\alpha_1 &= \quad 9 \\
\alpha_2 &= \quad 9.5 \\
\beta_{1:168,1} &= \frac{\sin(\frac{2\pi h}{168})}{24}, h = 1..., 168 \\
\beta_{1:168,2} &= \frac{\cos(\frac{2\pi h}{168})}{24}, h = 1..., 168 \\
\boldsymbol{\theta} &= (\theta_1, \theta_2, ...\theta_7) \sim N(0,1) \\
\lambda_t &= log(\frac{\zeta_t}{1-\zeta_t}) = log(\frac{(B_\zeta \zeta)_t}{1-(B_\zeta\zeta)_t}) \\
\zeta_j &\sim (0,1) \\
\epsilon_t &\sim N(0, 0.1)
\end{aligned}$$

For $\zeta_j, j = 1...6$, are coefficients for B-spline basis function (B_ζ) for logarithm of mixing weights λ.
It is highly likely that the real data are more likely to be represented by a 2-component Gaussian mixture model than 1-component.

5.4.3.2 Results

Figure 5.6 shows 8 randomly chosen (consecutive in time) hourly histogram density for the simulated log(PNC) observations with the "true" density superimposed (in red). The green and the blue lines are the posterior predictive density fitted using all 12 observations within an hour estimated using the regular classical data likelihood function and the corresponding hourly histogram posterior predictive densities are estimated using the symbolic likelihood function. As before, it is apparent that the symbolic model follows

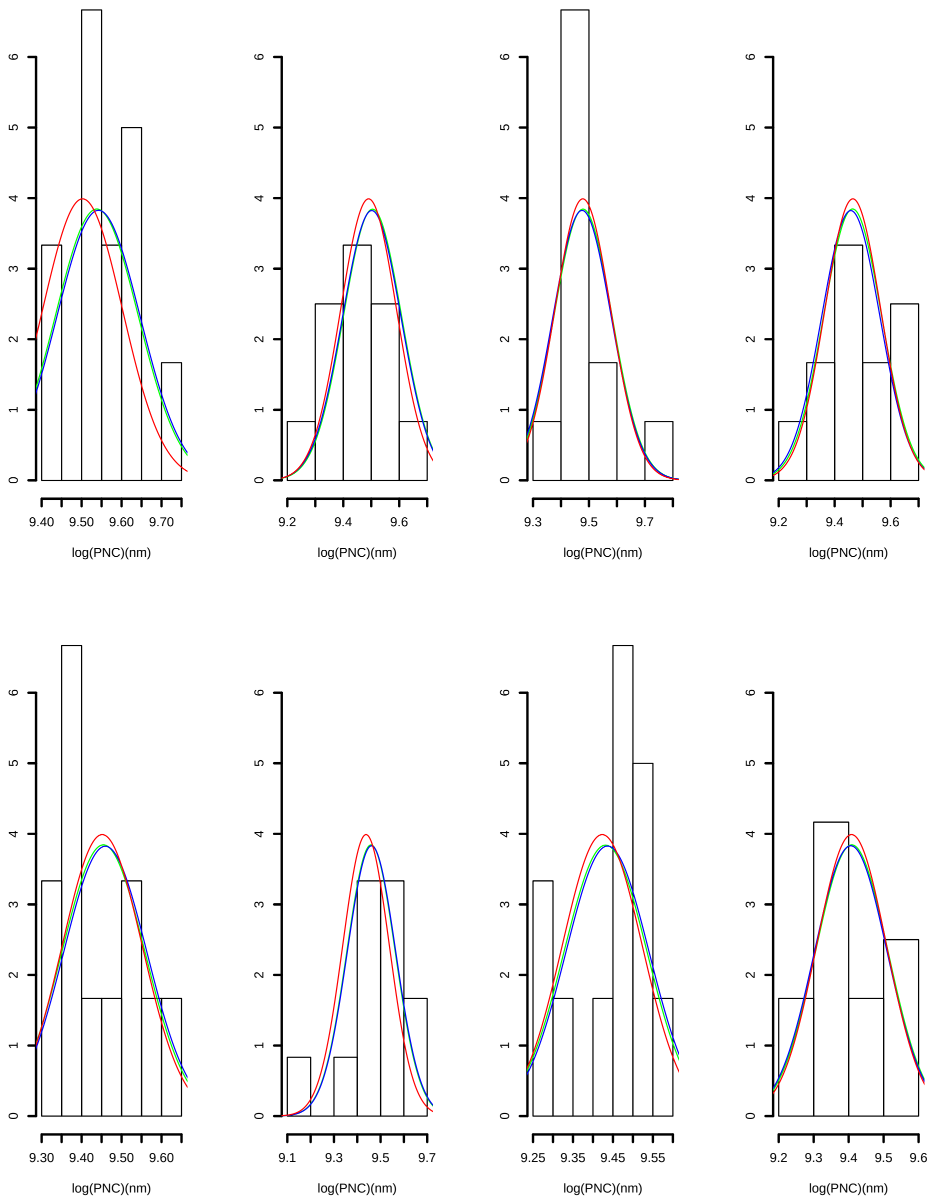

Figure 5.5 – 8 randomly chosen (consecutive in time) hourly density shown in histograms with simulated true density shown in red. The green line is the posterior density from the Equation (5.2). The blue line is the posterior density from the Equation (5.3.1).

the classical data model fairly well. This is also true for all temporal parameters of the underlying data distribution which can be seen in Figure C.4 and Figure C.5 for the 2-component mixture overall means (μ_{t1}, μ_{t2}), in Figure C.6 and Figure C.8 for the marginal daily-weekly effects (β_{t1}, β_{t2}), and a marginal annual effect ($B\boldsymbol{\theta}$) shared by both mixture components in Appendix C.2.

5.4.4 2-component Gaussian model

5.4.4.1 Data setup

In this section, for a 2-component Gaussian mixture model, data are simulated from the equation below

$$y_{it} = \begin{cases} \alpha_1 + \beta_{t1} + (B\boldsymbol{\theta})_t + \epsilon_t, & \text{with probability } \lambda_t \\ \alpha_2 + \beta_{t2} + (B\boldsymbol{\theta})_t + \epsilon_t, & \text{with probability } 1 - \lambda_t. \end{cases}$$

$$\begin{aligned}
\alpha_1 = &\quad 9 \\
\alpha_2 = &\quad 9.5 \\
\beta_{1:168,1} &= \frac{\sin(\frac{2\pi h}{168})}{24}, h = 1..., 168 \\
\beta_{1:168,2} &= \frac{\cos(\frac{2\pi h}{168})}{24}, h = 1..., 168 \\
\boldsymbol{\theta} &= (\theta_1, \theta_2, ...\theta_7) \sim N(0,1) \\
\lambda_t &= log(\frac{\zeta_t}{1-\zeta_t}) = log(\frac{(B_\zeta \zeta)_t}{1-(B_\zeta \zeta)_t}) \\
\zeta_j &\sim (0,1) \\
\epsilon_t &\sim N(0, 0.1)
\end{aligned}$$

For $\zeta_j, j = 1...6$, are coefficients for B-spline basis function (B_ζ) for logarithm of mixing weights λ.
It is highly likely that the real data are more likely to be represented by a 2-component Gaussian mixture model than 1-component.

5.4.4.2 Results

Figure 5.6 shows 8 randomly chosen (consecutive in time) hourly histogram density for the simulated log(PNC) observations with the “true” density superimposed (in red). The green and the blue lines are the posterior predictive density fitted using all 12 observations within an hour estimated using the regular classical data likelihood function and the corresponding hourly histogram posterior predictive densities are estimated using the symbolic likelihood function. As before, it is apparent that the symbolic model follows the classical data model fairly well. This is also true for all temporal parameters of the underlying data distribution which can be seen in Figure C.4 and Figure C.5 for the 2-component mixture overall means (μ_{t1}, μ_{t2}), in Figure C.6 and Figure C.8 for the marginal

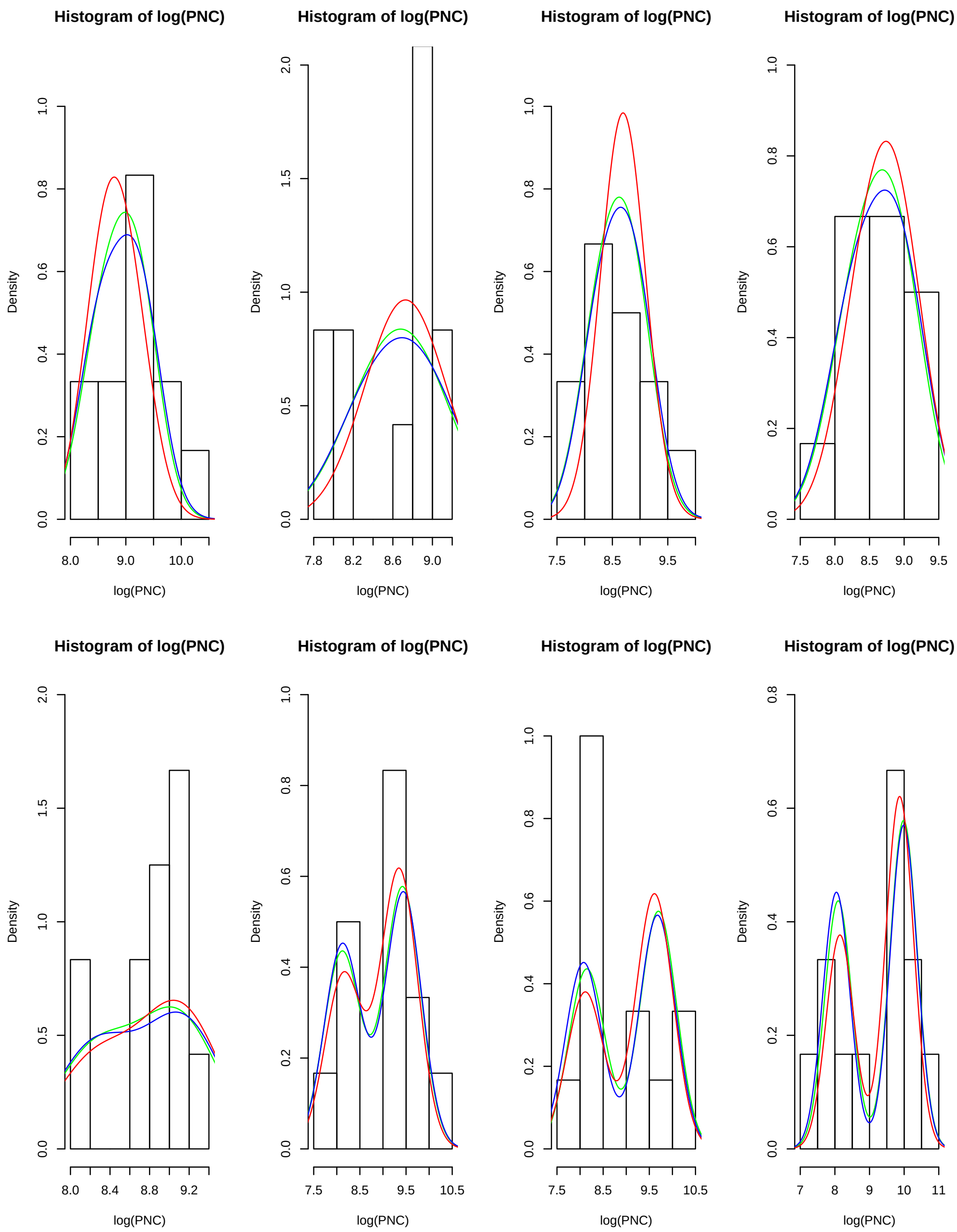

Figure 5.6 – 8 randomly chosen (consecutive in time) hourly density shown in histograms with simulated true density shown in red. The green line is the posterior density from the Equation (5.3). The blue line is the posterior density from the Equation (5.3.1).

daily-weekly effects (β_{t1}, β_{t2}), and a marginal annual effect ($B\boldsymbol{\theta}$) shared by both mixture components in Appendix C.2.

5.4.5 Model Selection based on WAIC information criterion

When analysing real data, we do not know in advance how many components are there in the underlying classical data. Based on a preliminary exploratory analysis of the real data we are going to examine in Section 5.5, it is a reasonable assumption that the underlying log(PNC), y_{it} comes from one of 2 data generating processes, either a 1-component or a 2-component Gaussian mixture model.

As a result, when modelling the real data, it is reasonable to fit both 1-component Gaussian model specified in Equation 5.2 and its 2-component counterpart described in Equation 5.3. We quantitatively assess the model fit between 1-component and 2-component models by using an information criterion measurement. Vehtari and Gelman (2014) proposed the Watanabe-Akaike information criterion (WAIC), which measures posterior predictive accuracy and can be viewed as a tool for model comparison, selection or averaging. It is viewed as an improvement on the traditionally popular deviance information criterion (DIC) for Bayesian models as it is being fully Bayesian rather than based on a point estimate. In addition, it is invariant to parametrisation and also works for singular models. As a result, WAIC is chosen to assist the model selection decision.

The following simulations are set up to test the reliability of WAIC as a model selection criterion. Two simulations are conducted, in the first simulation, the underlying data are simulated from a single Gaussian distribution and we fit both the correct 1-component Equation 5.2 and the incorrect 2-component Equation 5.3. Both models are fitted with classical data and with histogram-valued symbolic data constructed from the 5 observed data quantiles. In the second simulation, the underlying data are simulated from a 2-component Gaussian distribution with overlapping mixture components. In total, 4 models (1-component classical, 2-component classical, 1-component 5-quantile and 2-component 5-quantile) are fitted.

Fitted \ Truth	1-component	2-component
1-component	-12136.01	8412.029
2-component	-12122.71	7089.083

Table 5.2 – WAIC for one and two-component Gaussian mixture models when fitted data generated from each model. All models are fitted using classical data

Table 5.2 shows WAIC values for each various simulation scenarios when the model is fitted with classical data y_{it}. In both cases, the WAIC of the correct model (the diagonal values in the table) is smaller than that of the incorrect models (the off-diagonal values in the table).

In terms of the model fitted with 5-quantile histogram-valued symbolic data, Table 5.3 likewise reveals that the correct models (the diagonal values in the table) have lower WAIC values than the incorrect models (the off-diagonal values in the table). In conclusion, when modelling the real data in Section 5.5, the number of components (either 1 or 2) can be selected based on WAIC.

It is worth mentioning that the values in Table 5.2 and in Table 5.3 are in different scales and they are not comparable. This is expected as they are computed using different datasets.

In terms of posterior fit of the simulated data, in the situations where the "truth" is a 1-component model and both classical and symbolic models fit the correct model shown in Figure C.9 in Appendix C.3, the symbolic model follows the classical data model well. When the "truth" is a 1-component model while both classical and symbolic models over-fit with 2-component are shown in Figure C.10 in Appendix C.3, the symbolic model again follows the classical data model well. The corresponding two cases for the "truth" as a 2-component can be seen in Figure C.11 and Figure C.12 in Appendix C.3. In summary, regardless of whether the correct number of components are fitted, model parameters estimated by the symbolic model follows the classical data model fairly closely.

Truth / Fitted	1-component	2-component
1-component	7936.503	8548.240
2-component	8103.836	7736.981

Table 5.3 – WAIC for one and two-component Gaussian mixture models when fitted data generated from each model. All models are fitted using 5-quantile histogram-valued symbolic data

5.4.6 Time comparisons

In this section, we wish to demonstrate the computational advantange of the SDA method and so we record average time to simulate 1000 posterior samples for 1- and 2-component models fitted with classical and symbolic data. We consider fitting 4 models (1-component classical, 2-component classical, 1-component 5-quantile and 2-component 5-quantile) with 3 different sizes of data matrix. The first one is a replicate of the real data matrix $D \in \mathbb{R}^{12\times336}$ (for 12 observations per hour over a two-week period). The second matrix $D \in \mathbb{R}^{101\times336}$ and the last matrix $D \in \mathbb{R}^{201\times336}$. There are altogether 12 cases and we repeat each case 50 times. The "true" parameters for the 1-component Gaussian model are the same as described in Section 5.4.2.1 and for the 2-component Gaussian mixture model are the same as described in Section 5.4.4.1

It is expected that in the case of fitting a 1-component Gaussian model, the classical model would take less time to sample than the model fitted with symbolic likelihood function. This is because to estimate the parameters in a 1-component Gaussian model, the modelling software just needs to know the sufficient statistics–the sample mean for each column–and evaluates the Gaussian probability density function at each column mean. Therefore, the increment in time is expected to be relatively small, and stable with increasing numbers of observations per column. On the other hand, regardless of how many observations per hour, the symbolic likelihood has to evaluate the differences between two Gaussian cumulative density functions 5 times, which is more complex than that of the classical counterpart. Although the increases in time with increasing number of observations per symbol will also be stable. Therefore, in the case of fitting a 1-component Gaussian model, the classical model should outperform the symbolic model in terms of run time. This expectation can be verified in Table 5.4 where the mean time for these 50 runs is shown for each model with different sample sizes. In addition, the corresponding standard error of each mean time is shown in brackets. For a given sample size, it can be seen that the model fitted with classical Gaussian likelihood function outperforms the model fitted with symbolic likelihood function for histogram.

In the case of fitting a 2-component Gaussian mixture model, the model fitted with the symbolic likelihood outperforms the model fitted with the individual observations with increasing number of observations. This is because, there are no low dimensional sufficient statistics for a finite Gaussian mixture model beyond the full dataset and as a result, fitting the classical model means that the program has to evaluate the mixture Gaussian density at every single observation and the time increases with increasing number of observation. On the contrary, the symbolic model only needs to evaluate the differences between 2 cumulative density functions of a mixture models 5 times, regardless of how many observations are there. Therefore, we expect to see an advantage in run time using the symbolic model when fitting complex models such as the mixture models in comparison to its classical counterparts. Table 5.5 again verifies our expectation where for a given sample size, the classical data model requires more time than the symbolic model. In addition, the run time increases with sample size while the run time remains roughly stable for the symbolic model.

	Classical	5-quantile
N=12	261.1882 (3.8304)	301.7358 (3.5793)
N=101	774.0944 (33.4542)	894.9764 (37.2043)
N=201	778.2316 (29.5631)	832.4824 (33.5813)

Table 5.4 – Mean Run time (in seconds) with standard error over 50 runs shown in bracket for one-component Gaussian model when fitted with classical and symbolic likelihood with $N = 12, 101, 201$ in data matrix D.

5.5 A Bayesian semi-parametric additive model with a finite Gaussian likelihood using SDA and its application

The simulations in Section 5.4 demonstrated the capacity of symbolic likelihood functions for histograms to recover the posterior distribution associated with the underlying data distribution in a way which may be less computationally intensive. Therefore, in this

	Classical	5-quantile
N=12	8,292.353 (388.2184)	7,827.154 (391.5197)
N=101	27,325.2 (1265.4610)	11,688.77 (374.4998)
N=201	59,822.24 (975.9697)	11,562.3 (232.7065)

Table 5.5 – Mean Run time (in seconds) with standard error over 50 runs shown in bracket for two-component Gaussian model when fitted with classical and symbolic likelihood with $N = 12, 101, 201$ in data matrix D.

section, the same sets of models are fitted to log(PNC) collected from a split panel design in Brisbane, Australia over a 2-week period. The data are collected in 25 primary schools and augmented by 3 long-term monitoring sites. We fit the model to a single school site 7 located in an inner suburban area near elevated freeways with prevailing south to west winds. The two-week PNC measurement campaign was conducted from 30/05/2011 to 12/06/2011. It is noted that there are some hours where one or several 5-minute interval measurements were missing. This is not an issue for the previous analysis performed by Clifford et al. (2012a) as only the hourly averaged size fractionated log(PNC) is fitted with an implicit assumption of observations missing completely at random (MCAR). Similarly, MCAR assumption is adopted here. The approach to deal with missing observations when fitting classical model and 5-quantile histogram-valued models are described respectively in Section 5.4.1.

5.5.1 1-component Gaussian model

It can be shown in Figure 5.7 that the 1-component classical data Equation (5.2) can broadly capture the overall temporal patterns of log(PNC) in school 7 over the 2-week period. The same model structure but estimated using the symbolic likelihood function produces very similar results. Figure 5.8 similarly shows the marginal daily-weekly effect, assumed to be the same for both weeks during the measurement campaign. Figure 5.9 shows the marginal day of the year effect, which starts off low and peaks on Thursday in both weeks. Figure 5.10 shows 8 randomly chosen (consecutive in time) observed hourly densities, overlaid with the posterior density fitted with the classical data likelihood (in green) and with symbolic likelihood (in blue). In all cases the symbolic likelihood based analysis produces results similar to the classical data analysis.

5.5.2 2-component Gaussian model

It can be seen in Figure 5.11 that the 2-component classical data Equation (5.3) can broadly capture the temporal fluctuations of log(PNC) in school 7 over the two-week period. The same model structure but estimated using the symbolic likelihood function model produces very similar results. Figure 5.12 shows that the time-varying mixing

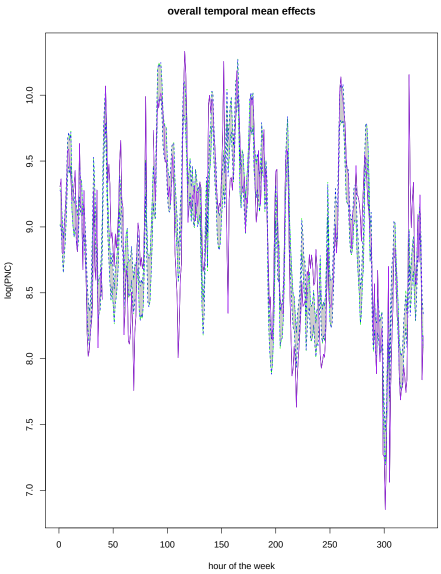

Figure 5.7 – Estimated posterior distribution for overall temporal mean $\alpha + \beta_t + (B\boldsymbol{\theta})_t$ in Equation 5.2. The green dashed lines are bounds of 95% credible interval from Equation 5.2 with credible interval region shaded in grey. The blue dashed lines are bounds of 95% credible interval from Equation 5.3.1. The purple solid line is the observed hourly averaged values of log(PNC) with "NA" values removed.

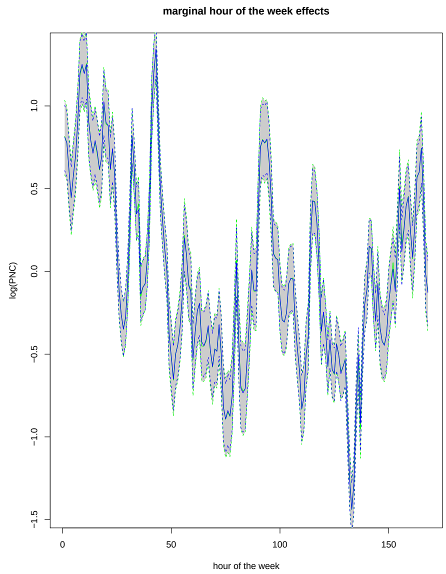

Figure 5.8 – Estimated posterior distribution for marginal daily-weekly effects β_t. The green dashed lines are bounds of 95% credible interval from Equation (5.2) with 95% credible interval region shaded in grey. The blue dashed lines are bounds of 95% credible interval from Equation (5.3.1). The green solid line is the posterior mean of Equation (5.2). The blue solid line is the posterior mean of Equation (5.3.1).

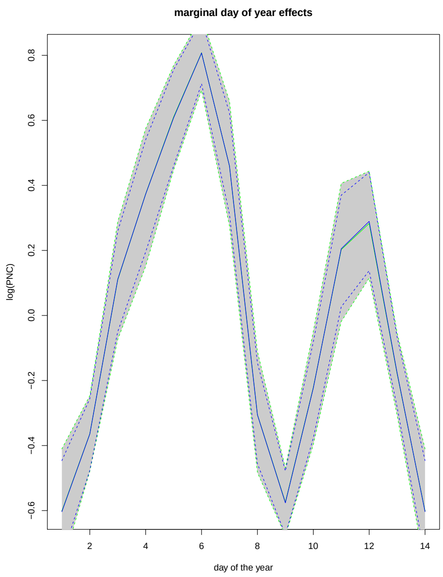

Figure 5.9 – Estimated posterior distribution for marginal day of year effects $(B\boldsymbol{\theta})_t$. The green dashed lines are bounds of 95% credible interval from Equation (5.2) with 95% credible interval region shaded in grey. The blue dashed lines are posterior bounds of 95% credible interval from Equation (5.3.1).The green solid line is the posterior mean of Equation (5.2). The blue solid line is the posterior mean of Equation (5.3.1).

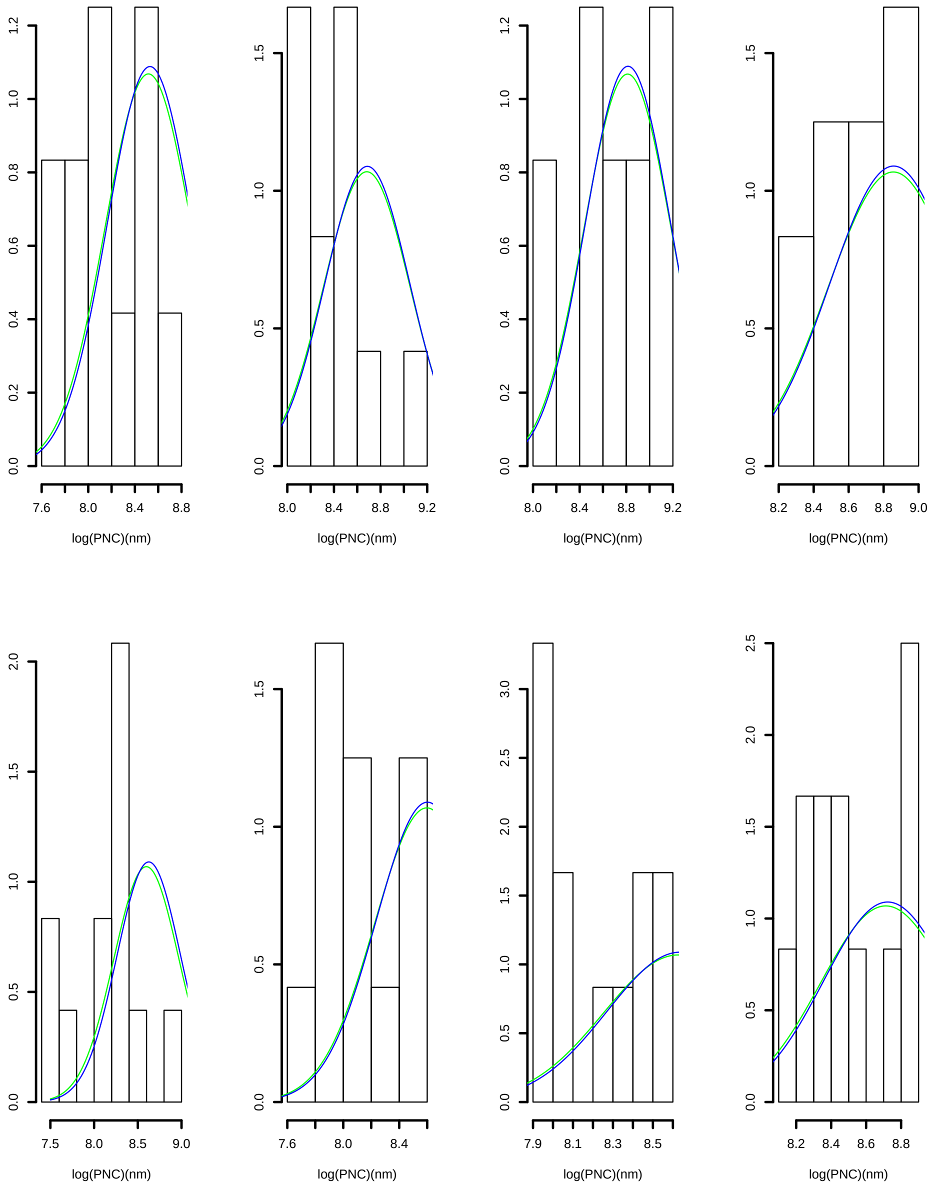

Figure 5.10 – 8 observed hourly histograms with posterior predictive density from Equation 5.2 shown in green. The blue line is the posterior predictive density from 2-component mixture model in Equation 5.2 estimated using the symbolic likelihood function in Equation 5.3.1.

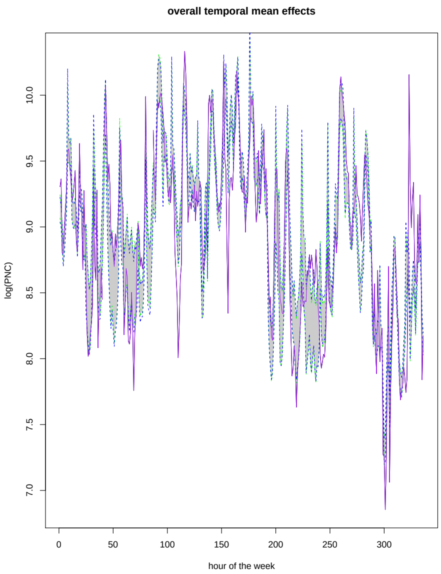

Figure 5.11 – Estimated posterior distribution for overall mixture mean temporal trend $\lambda_t(\alpha_1+\beta_{t1}+(B\boldsymbol{\theta})_t)+(1-\lambda_t)(\alpha_2+\beta_{t2}+(B\boldsymbol{\theta})_t)$ for school 7 over a two-week period. The green dashed lines are bounds of 95% credible interval from Equation (5.3) with 95% credible interval region shaded in grey. The blue dashed lines are bounds of 95% credible interval from Equation (5.3.1). The purple solid line is the observed hourly averaged values of log(PNC) with “NA” values removed.

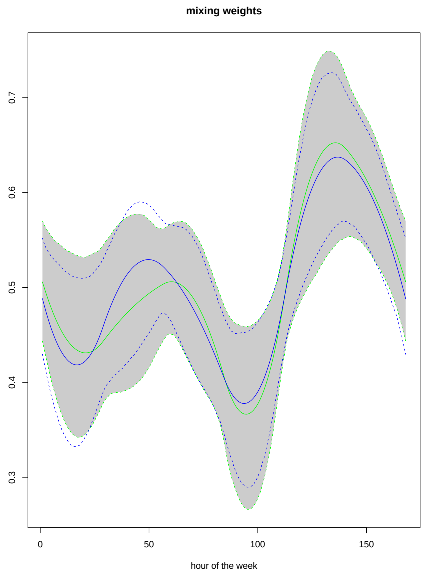

Figure 5.12 – Estimated posterior distribution for mixing weights λ_t. The green dashed lines are bounds of 95% credible interval from Equation (5.3) with 95% credible interval region shaded in grey. The blue dashed lines are bounds of 95% credible interval from Equation (5.3.1). The green solid line is the posterior mean from Equation (5.3). The blue solid line is the posterior mean of the mixture model with symbolic likelihood in Equation (5.3.1).

weights appear to evolve slowly and smoothly hour by hour over 1-week period. These weights are assumed to repeat for all weeks in a year. Figure 5.13 and Figure 5.14 illustrate the marginal daily-weekly effects in mixture components 1 and 2. As with the the 1-component Gaussian model, these hour of the week effects are assumed to be the same for both weeks. The shape of the marginal annual effects are similar to Figure 5.9. The level of log(PNC) was low at the beginning of the week with a peak occurring on Thursday and gradually decreased over the weekend and repeated the pattern for the second week. On average, log(PNC) is higher during the first week of the measurement campaign. In Figure 5.16 shows that the symbolic mixture hourly density follows the classical mixture model hourly density fairly well.

Model	1-component	2-component
Classical Data	3405.738	1601.711
Symbolic Histogram Data	13722.73	12561.28

Table 5.6 – WAIC for real data model comparison

Table 5.6 shows the WAIC values for the 2 types of models (1-component versus 2-component) fitted with classical data and histogram-valued symbolic data respectively. In

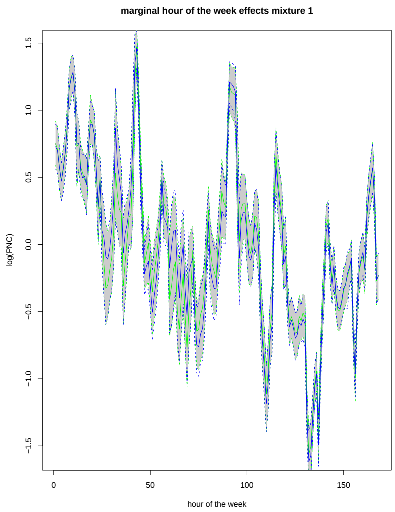

Figure 5.13 – Estimated posterior distribution for marginal daily-weekly effects β_{t1}. The green dashed lines are bounds of 95% credible interval from Equation (5.3) with 95% credible interval region shaded in grey. The blue dashed lines are bounds of 95% credible interval from Equation (5.3.1). The green solid line is the posterior mean from Equation (5.3). The blue solid line is the posterior mean of the mixture model with symbolic likelihood in Equation (5.3.1).

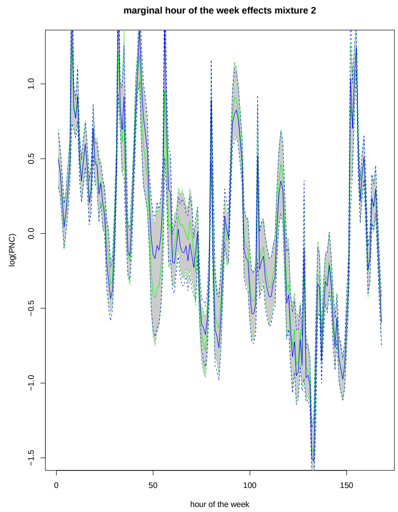

Figure 5.14 – Estimated posterior distribution for marginal daily-weekly effects β_{t2}. The green dashed lines are bounds of 95% credible interval from Equation (5.3) with 95% credible interval region shaded in grey. The blue dashed lines are bounds of 95% credible interval from Equation (5.3.1). The green solid line is the posterior mean from Equation (5.3). The blue solid line is the posterior mean of the mixture model with symbolic likelihood in Equation (5.3.1).

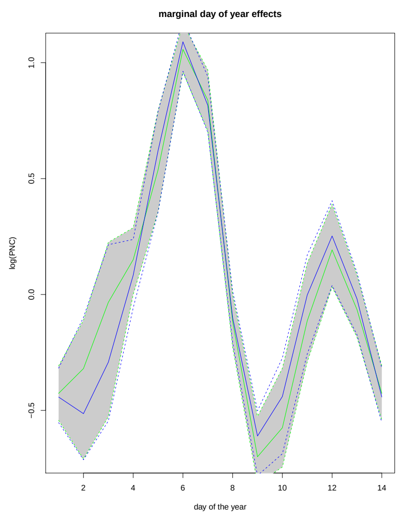

Figure 5.15 – Estimated posterior distribution for marginal day of the year effects $B\boldsymbol{\theta}$. The green dashed lines are bounds of 95% credible interval from Equation (5.3) with 95% credible interval region shaded in grey. The blue dashed lines are bounds of 95% credible interval from Equation (5.3.1). The green solid line is the posterior mean from Equation (5.3). The blue solid line is the posterior mean of the mixture model with symbolic likelihood in Equation (5.3.1).

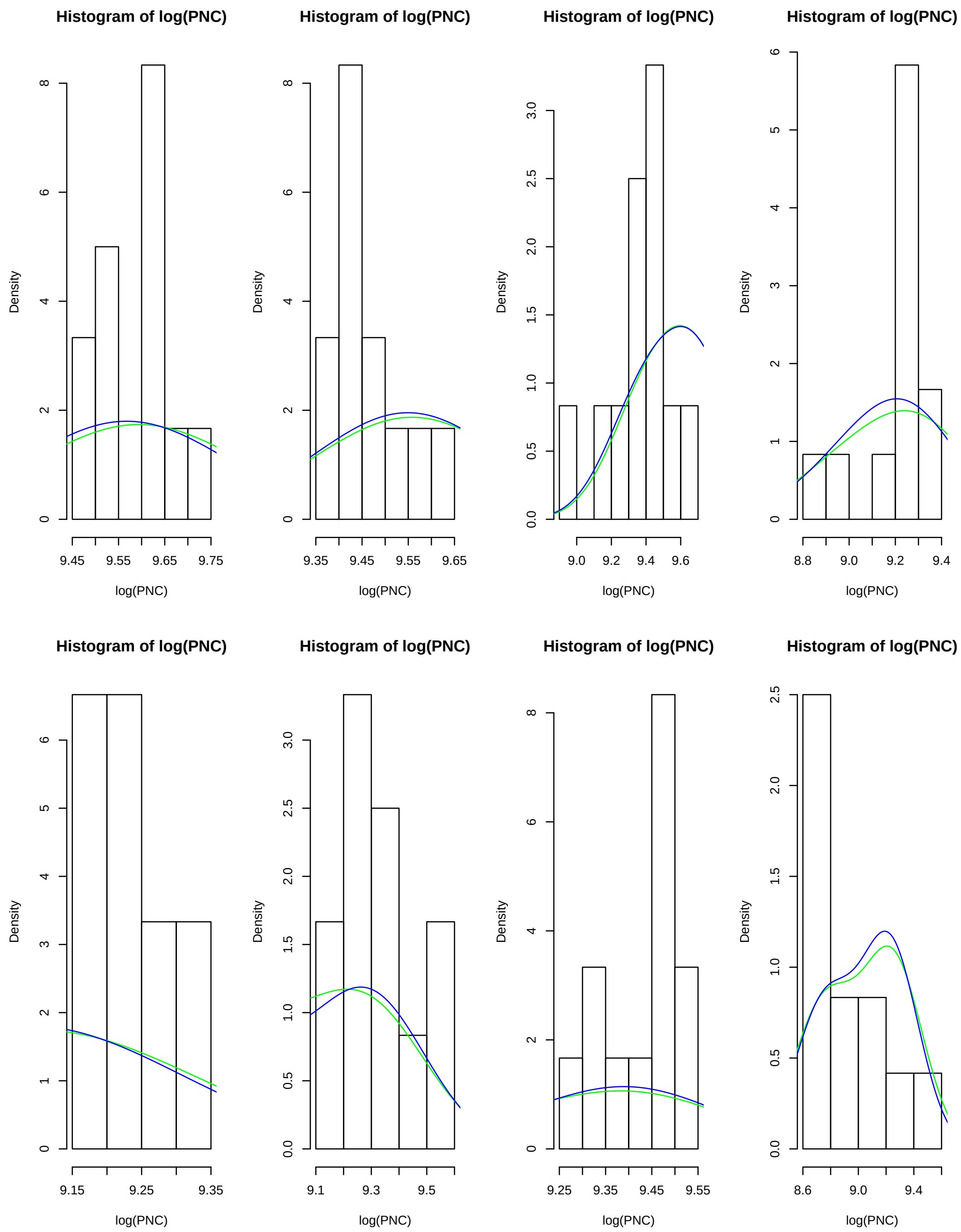

Figure 5.16 – 8 observed hourly histograms with posterior predictive density from Equation 5.3 shown in green. The blue line is the posterior predictive density from 2-component mixture model in Equation 5.3 estimated using the symbolic likelihood function in Equation 5.3.1.

can be seen that, regardless of the type of data fitted, the WAIC favours a 2-component Gaussian model, although few differences are shown in the posterior distributions of the fitted overall mean in Figure 5.7 and in Figure 5.11.

According to Vehtari and Gelman (2014), the WAIC estimates out-of-sample pointwise predictive accuracy using posterior simulations and thus it is more reasonable to visually compare both 1-component and 2-component models using posterior predictive distributions. In addition, it might be of interest for future analysis to set up a protocol for the maximum permitted level of PNC around primary schools based on individual 5-minute observations. It is not possible to investigate the level of prediction using the mean-level model of Clifford et al. (2012a). For demonstrative purposes, we naively set the maximum permitted level to be the 95^{th} percentile of the observed log(PNC) = $9.98cm^{-3}$. Figure 5.17 shows the 95% posterior predictive distribution from a 1-component Gaussian model fitted with classical data shown in green and the symbolic data model shown in blue and the observed data are plotted as black dots. The purple line represents the protocol level. It is evident that the 95% posterior predictive distribution from the symbolic data model follows the one fitted with classical data model closely. However, the 95% posterior predictive distributions from both models fail to capture some observations with extremely high log(PNC).

Figure 5.18 shows the hourly posterior predictive probability for exceeding the level defined by log(PNC) = $9.98cm^{-3}$ from 1-component fitted with classical data in green with the symbolic data model superimposed in blue. The x-axis labelling 1:336 represents the hour during this 2-week period, with 1 refers to 1am on Monday in the first week and 336 refers to 12am on Sunday in the second week. Based on this figure, it can be seen that the probability of exceeding the set level varies at different hours of the day during this 2-week period. However, it can be seen that the peak (the highest probability of exceeding the maximum permitted level) at roughly the same location on Monday, Tuesday, Thursday and Friday in both weeks, while it is least likely to exceed on both Wednesdays. It is interesting to note that the weekend pattern of the first week does not repeat on the second week. The probability of exceeding the pre-determined level is fairly unlikely in the second weekend.

On the contrary, when examining the same set of plots from a 2-component Gaussian model, Figure 5.19 shows that the 2-component model is better at capturing the outlying observations with high values. The posterior predictive distribution from both 2-component classical and symbolic data models tend to capture a wider range comparing to their 1-component counterparts. Figure 5.20 illustrates a slightly different story to that in Figure 5.18. The pattern of the exceedance probability of weekdays in week 1 repeats in week 2. In addition, the pattern of the exceedance probability of Sunday in week 1 also repeats in the following week. While the probability of exceeding the level is fairly likely in the first Saturday, it is almost unlikely that it would exceed in the second Saturday.

In addition, it is worth mentioning that the exceedance probability is on average higher than in Figure 5.18 when the underlying data fitted with a single Gaussian distribution. The same conclusion that school 7 exhibits an evening peak hour level is also found in

Clifford et al. (2012a)'s model.

In conclusion, based on WAIC and the graphs of posterior predictive plots, to adequately model this set of data, we need a 2-component Gaussian model. Fitting a naive single Gaussian distribution would potentially lead to misleading results and underestimate the probability of exceeding a pre-set PNC level.

5.6 Discussion

Atmospheric particulate matter (PM) is one of the main pollutants that directly affects air quality and climate. While the air quality standards have been set up for the mass concentration of PM_{10} and $PM_{2.5}$, less attention has been paid to ultrafine particles (UFPs) (Cheung et al., 2013). Therefore, it was a primary goal of this chapter to develop a statistical model that can represent aerosol dynamic process of UFPs through the evaluation of their particle number concentration (PNC). The developed model was then applied to measurements of particle number concentration in Brisbane, Australia, collected as part of the UPTECH project.

An existing Bayesian semi-parametric additive model with a Gaussian likelihood for modelling hourly averaged log(PNC) was proposed by Clifford et al. (2012a). In this chapter, we restrict our attention to modelling temporal aspect of log(PNC). Due to the fact that aerosol particles are governed by formation and transformation processes, they are likely to form modal features. Therefore, we represented this distinct feature by adopting a finite Gaussian mixture model for the underlying data. Given the temporal dependence exhibited in the underlying data, we allowed parameters in the mixture model to be correlated and to smoothly vary over time. Lastly, rather than just use the summary statistics of underlying data-hourly averaged measurement, we used concepts from SDA to represent and aggregate the whole data matrix into histogram-valued symbolic data from which subsequent analysis were then performed. This allowed us to construct efficient models for the full distributions of the observed data, rather than just the mean level, as with Clifford et al. (2012a).

Based on a series of simulation studies and a real data analysis, it has demonstrated advantage of the proposed 2-component Gaussian mixture model as opposed to fitting a single Gaussian likelihood model for the underlying data. In addition, new method of model fitting for histogram-valued symbolic data based on fitting to the underlying data was proven to be as good as the model fitted with classical data in terms of parameter estimation and model fit. In addition, we illustrated the use of WAIC in assisting us choosing the number of mixture components. While in this case, with only 12 observations per histogram, there was no major advantage in implementing the model with symbolic likelihood function over the classical data model, for other datasets with larger numbers of data points per histogram, the symbolic data approach will be much more efficient as discussed in Section 5.4.6.

However, there are some caveats of the current model fit for the real data. Firstly, the proposed model focussed only on temporal variability. A more comprehensive model could have been built to incorporate spatial variability which would offer more insights into how

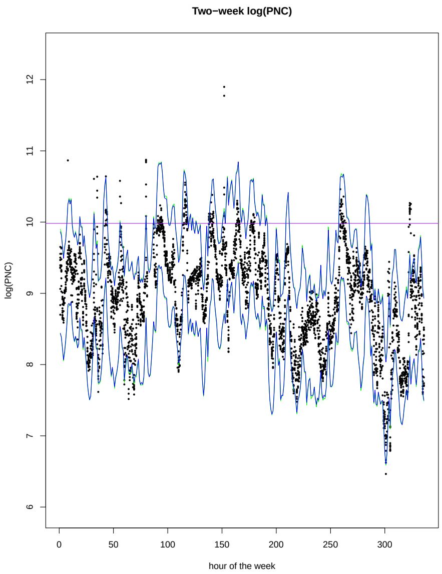

Figure 5.17 – The hourly posterior predictive probability for exceeding the level defined by log(PNC) = $9.98 cm^{-3}$. Individual observations are shown in black dots with a predetermined threshold level drawn in purple.
95% posterior predictive intervals of 1-component classical data model shown in blue and 1-component symbolic data model shown in green.

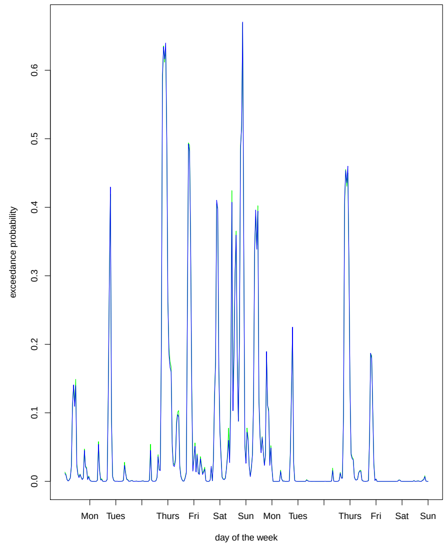

Figure 5.18 – The hourly posterior predictive probability for exceedance of high level from 1-component classical data model shown in blue and the corresponding probability from 1-component symbolic data model.

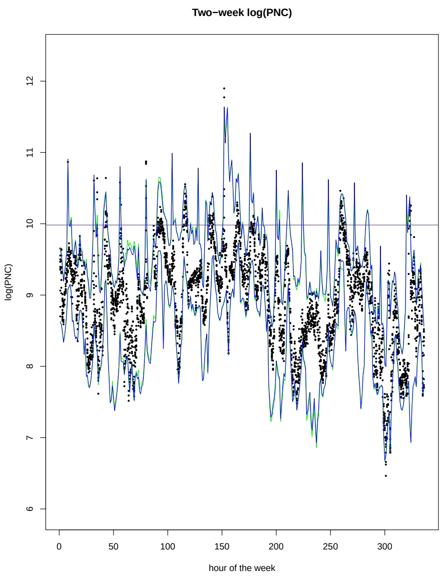

Figure 5.19 – The hourly posterior predictive probability for exceeding the level defined by log(PNC = $9.98 cm^{-3}$. Individual observations are shown in black dots with a predetermined threshold level drawn in purple.
95% posterior predictive intervals of 2-component classical data model shown in blue and 2-component symbolic data model shown in green.

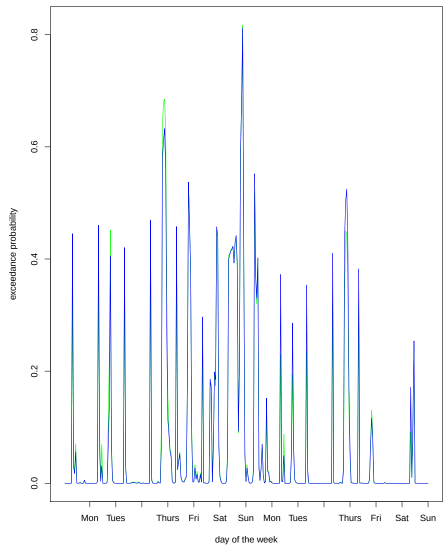

Figure 5.20 – The hourly posterior predictive probability for exceedance of high level from 2-component classical data model shown in blue and the corresponding probability from 2-component symbolic data model.

log(PNC) differs with respect to different the school locations. In addition, by considering multiple locations, it would help improve the estimation of temporal variability as one can break down and specify an overall temporal effect common to all sites and site-specific marginal temporal effects. Secondly, meteorological covariates such as wind speed, humidity, temperature should be considered and included in the model construction, although we would then need joint histogram of PNC and meteorological explanatory variables. The elaborated model consisting of temporal, spatial and meteorological variability would provide a fuller picture. In addition, when this model is integrated into the UPTECH project, it would be better at assessing the association between exposure to UPFs and respiratory health in primary school children in Brisbane, Australia, which is the ultimate goal of the project (Ezz et al., 2015; Clifford et al., 2018). Lastly, we acknowledge the fact that there are only 12 observations per hour and one could have fitted the same model 5.3 with all data without resorting to SDA techniques. However, this data was used merely as an illustration of the efficacy of the new method of model fitting for symbolic data to underlying data. Assuming PNC is measured at even smaller time intervals, the computational time would increase exponentially for the model fitted with all the classical data. On the other hand, the model fitted with histogram-valued symbolic data would have constant computational overheads as the number of classical data points increased. It is highly likely that much of the computational time would be attributed to aggregating underlying data into symbols rather than on model fitting using the histogram-valued symbols.

In spite of the above deficiencies, the proposed 2-component finite mixture with time-dependent parameters estimated using the new method introduced in SDA literature by Beranger et al. (2018) provides a flexible way of modelling, and has been shown to have the same estimation results as the classical data model, while using a small fraction of the observed data in distributional form.

Chapter 6

Discussion and Future Work

Although SDA is a relatively new field in statistics, there has been considerable development in this area, evident by a large number of publications, reports and developments of software tools. SDA extends statistics and multivariate data analysis to deal with data structured in distributional form with complex internal variations. Under the umbrella of SDA, one is required to think and aggregate data points into “classes” of interest. As a result, it reduces the volume and complexity of the data and it is particularly useful to glean information in a world full of “big” data and uncertainty.

Among the different types of symbolic variables described in Chapter 2, interval-valued symbolic variables have drawn the most attention from researchers who work in this area while some have focussed on methods for analysing histogram-valued variables. Nevertheless, existing SDA methods dealing with both, e.g., univariate or multivariate descriptive statistics, similarity and dissimilarity measures, clustering, discrimination methods and linear regression models have proceeded largely based on the assumption of uniformity within each symbol (Bock, 2008). The uniform distribution assumption is overly simplified and thus inappropriate, as acknowledged by Kosmelj et al. (2014) in their analysis of meteorological data in Slovenia.

In spite of the fact that non-parametric descriptive approaches are prevalent in the SDA literature, there have been some advances in proposing parametric models for symbolic data. Bock (2008) first proposed a probabilistic modelling for symbolic data with a focus on probabilistic clustering of interval-valued data. Le-Rademacher and Billard (2011) were among the first ones to propose likelihood functions for interval and histogram-valued symbolic variables. Brito and Duarte Silva (2012) presented a probabilistic model for interval variables which involves a reparametrisation of interval variables into bivariate vectors. The authors assumed either a normal or a skew-normal distribution for these bivariate random variables on which standard maximum likelihood estimation were then applied. Furthermore, the authors proposed different configurations of the global covariance matrix, which offers a flexible way of modelling the relationship that may exist between the bivariate vector of the same or different interval variables. However, the above approaches are still based on a uniformity-within-symbols assumption and are built directly at the symbol level and thus have potentially limited usage. Zhang and Sisson (2016) have made promising steps towards building a probabilistic framework that considers both intraand inter- symbol relationships for interval-valued variables. Likelihood-based statistical analysis remains an important challenge in SDA and much is yet to be explored.

Within ecology there has been a need for better estimates of global and individual taxa species richness, that is the number of species found in a community or ecosystem. This problem becomes even more challenging when ecologists come up with estimates recorded in different forms. These different forms of data have greatly limited the analysts' ability to combine them to reduce parameter estimation uncertainty. In Chapter 3, we tackle this challenge through combining three statistical approaches. Firstly, the uncertainty in estimates is reduced through a meta-analysis approach such that it enables building on knowledge gained through previous attempts to species estimation. Secondly, previous species estimates data recorded in three different forms are reconciled and combined by representing them as interval-valued symbolic variables where some of them are partially observed. Lastly, using the likelihood-based approach of Brito and Duarte Silva (2012), we are able to construct a Bayesian hierarchical model that provides logically consistent estimates. Furthermore, this model permits us to estimate species that are not directly observed through pooling knowledge from other species categories.

As discussed above and in Chapter 1 and Chapter 2 that there are some serious drawbacks associated with the existing likelihood-based SDA approaches. In Chapter 4, we introduced a new method of model fitting for symbolic interval-valued and histogram-valued data based on fitting to the underlying micro data rather than fitting to summary statistics of symbols. The new likelihood-based method is constructed based on an underlying distribution from which classical data are thought to come from, together with an aggregation process that determines the 'class' membership of each classical data point. In this manner, this method permits likelihood-based statistical inferences to be made at the classical level while retaining the capability to obtain symbol level inference. Besides, this construction loses the assumption of uniformity-within-symbols. In addition, the proposed method provides a natural way to specify models for symbolic data, which is not so obvious for the existing methods. As illustrated in Chapter 2 that single-valued quantitative and qualitative classical variables are in fact special cases of corresponding symbolic variables, our method respects this fact and reduces to standard likelihood-based inference in the limit as symbol approaches classical data. Within this unified construction framework, our method is also shown to recover several existing methods (e.g., McLachlan and Jones (1988) for 1-dimensional histogram-valued symbolic variable) while offering models for symbols that have not been previously considered (e.g., multivariate intervals, multivariate histograms). More importantly, in Chapter 4, the new method shed some light on more informative ways of constructing symbols from underlying data. It has shown that the use of minimum and maximum order statistics for intervals is statistically inefficient in terms of capturing the internal distribution and thus leads to inaccurate inference. In addition, current design for multivariate interval symbols leads to a weak to non-identifiable dependence/ correlation structure within symbols.

In Chapter 5, the newly proposed symbolic likelihood function was applied to a Bayesian semi-parametric additive model with a finite mixture Gaussian model for modelling log(PNC). In contrast to a previous analysis (Clifford et al., 2012a), the proposed model was able to address future predictions at the individual 5-minute interval level. Secondly, the representation of log(PNC) as a finite mixture model with time-varying mixture parameters enabled us to describe the evolution of log(PNC) over a 2-week period.

Recognising the need for a unified likelihood-based approach in SDA, this thesis has proposed a new general construction tool for interval-valued and histogram-valued symbolic variables, which has been applied to model complex real-world data. The novelty of this new method lies in, firstly, unlike existing methods where uniformity distribution assumption is strictly required, the choice of distribution of the underlying data is entirely at the analysts' discretion for flexible modelling. Secondly, through fitting to the underlying data, our method allows statistical inference to be made at the classical data level that would otherwise not be possible using the existing likelihood-based approaches. Lastly, in almost all of the current SDA analyses, univariate symbols are adopted while our method opens up a vast number of opportunities to use higher-dimensional symbols.

However, much remains to be done. Firstly, it is found during developing the likelihood-based function for intervals that quantile representations are more informative about the distribution of underlying data than minimum and maximum order statistics conventionally used in SDA. In addition, for histogram-valued symbols, current construction decisions regarding the number of bins or bin locations are naively chosen. As a result, it maybe interesting yet challenging to establish a framework that can systematically design symbols tailored to individual analysis. In Chapter 5, the relationship between histogram-valued observations are assumed to be conditionally independent given the parameters of the underlying data model. However, it would be more advantageous to directly introduce the dependence relationship between successive histogram-valued observations. As early as in 1984, Schweizer (1984) foresaw that 'distributions are the numbers of the future'. To the best of our knowledge, little has been done for more general distributional types of symbols, for example, symbols in the forms of a mean vector and a covariance matrix of a Gaussian distribution and we recommend future research to realise his saying and pursue methods for general distributional types of symbols. We anticipate that solving the above research problems may potentially revolutionise how statistical inference has been implemented in the SDA framework.

Appendix A

Chapter 3: Real and posterior data tables

A.1 Real Data Table

A.2 Posterior Point Estimates for Species Richness

A.3 Prior sensitivity analysis

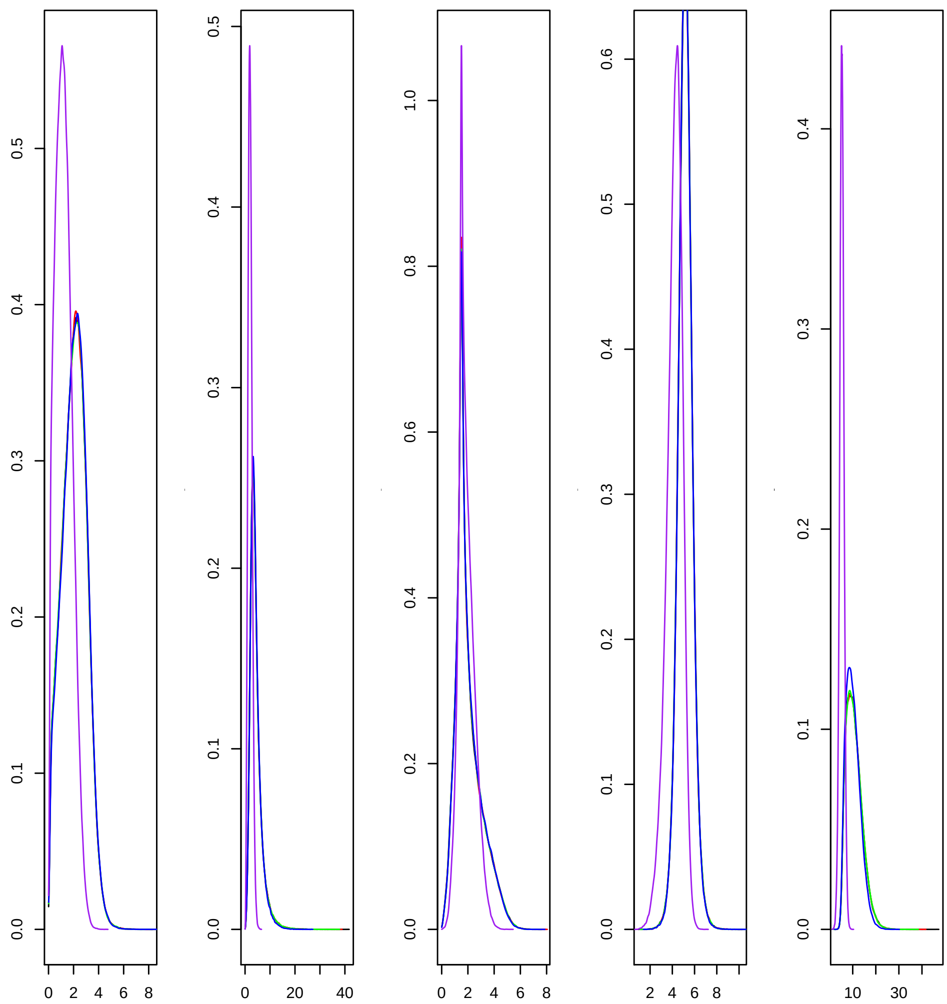

Figure A.1 – Posterior distribution for μ_{mj} when fitted with different scale values in the hyper-prior distribution $\mu_{mj} \sim N(0, \tau)I(\mu_m > 0)$. The black line represents $\tau = 10,000$, the green line represents $\tau = 1000$, the red line represents $\tau = 100$, the blue line represents $\tau = 10$ while the purple line represents $\tau = 1$. From the left to the right, the posterior distribution is for the j^{th} "parent" species beetles, coral reefs, marine, insects and arthropods respectively.

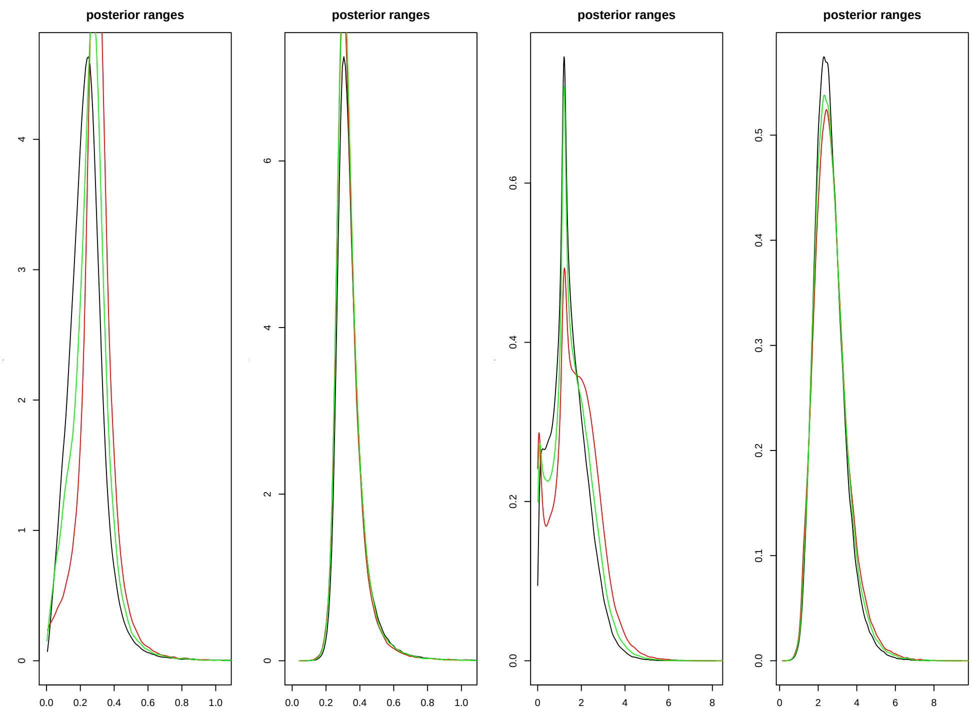

Figure A.2 – Posterior distribution for μ_{rj} when fitted with different scale values in the hyper-prior distribution $\mu_{mj} \sim N(0, \alpha)I(\mu_r << \log(2\mu_m))$. The black line represents $\alpha = 1.5$, the green line represents $\alpha = 2.5$, the red line represents $\alpha = 5$. From the left to the right, the posterior distribution is for the j^{th} "parent" species beetles, coral reefs, marine and insects respectively.

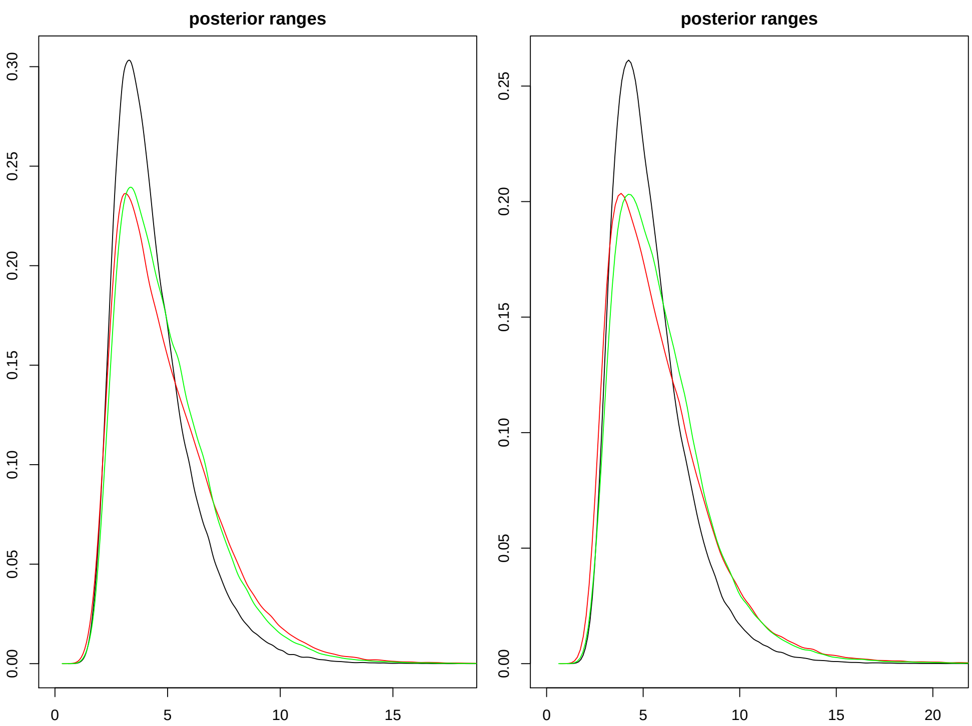

Figure A.3 – Posterior distribution for μ_{rj} when fitted with different scale values in the hyper-prior distribution $\mu_{mj} \sim N(0, \alpha)I(\mu_r < \log(2\mu_m))$. The black line represents $\alpha = 1.5$, the green line represents $\alpha = 2.5$, the red line represents $\alpha = 5$. From the left to the right, the posterior distribution is for the j^{th} "parent" species arthropods and global respectively.

Species Category	Lower (a)	Point estimate (x)	Upper (b)	Source
Beetles	1.5		2.1	Stork et al. (2015)*
Coral Reefs	0.62		9.5	Reaka-Kudla et al. (1996)
Coral Reefs	1.7		3.2	Small et al. (1998)
Coral Reefs	1	2	3	Reaka-Kudla (2005)
Coral Reefs	0.49		10	Knowlton et al. (2010)
Global	3		4	Raven (1983)
Global	10		100	Ehrlich and Wilson (1991)
Global		100		May (1992b)
Global		11.6		Cracraft and Grifo (1999)
Global	5	7^a	15	Raven et al. (2000)
Global		14		Groombridge and Jenkins (2002)
Global	7.4	8.7	10	Mora et al. (2011)
Global	1.8		2	Costello et al. (2011)
Global	2	5	8	Costello et al. (2013)
Marine		5		May and Beverton (1990)
Marine		10		Grassle and Maciolek (1992)
Marine		100		Lambshead (1993)
Marine		0.5		Raven et al. (2000)
Marine	1.4		1.6	Bouchet and Duarte (2006)
Marine	2.02	2.2	2.38	Mora et al. (2011)
Marine		0.3		Costello et al. (2011)
Marine	0.7		1	Appeltans et al. (2012)
Terrestrial		10		May (1992b)†
Terrestrial (Arthropods)		30		Erwin (1982)
Terrestrial (Arthropods)	10		80	Stork (1988)
Terrestrial (Arthropods)		6.6		Basset et al. (1996)
Terrestrial (Arthropods)	5		10	ØDegaard (2000)
Terrestrial (Arthropods)		3.7		Novotny et al. (2002)
Terrestrial (Arthropods)		5.9		Novotny et al. (2002)
Terrestrial (Arthropods)	3.6	6.1^a	11.4	Hamilton et al. (2010)
Terrestrial (Arthropods)	3.7	7.8^a	13.7	Hamilton et al. (2010)
Terrestrial (Arthropods)	2.9		12.7	Hamilton et al. (2013)
Terrestrial (Arthropods)	5.9	6.8^a	7.8	Stork et al. (2015)*
Terrestrial (Insects)	2.5		10	Sabrosky (1953)
Terrestrial (Insects)	3		5	May and Beverton (1990)
Terrestrial (Insects)	4.9		6.6	Stork and Gaston (1990)
Terrestrial (Insects)	1.84		2.57	Hodkinson and Casson (1991)
Terrestrial (Insects)	5		10	Gaston (1991)
Terrestrial (Insects)		8		Hammond (1995)
Terrestrial (Insects)	3		6	Raven et al. (2000)
Terrestrial (Insects)		4		Raven et al. (2000)
Terrestrial (Insects)		2		Nielsen and Mound (2000)
Terrestrial (Insects)		8		Groombridge and Jenkins (2002)
Terrestrial (Insects)	5		6	Raven and Yeates (2007)
Terrestrial (Insects)	2.6	5.5^a	7.8	Stork et al. (2015)*

Table 1 – Point (x) and interval (a, b) estimates of species diversity from 45 previously published studies. Diversity estimates are measured in millions. These data were originally collated by Caley et al. (2014) with the exception of those in Stork et al. (2015), as indicated by asterisks *. † indicates that this datapoint was not used in this analysis as it is strongly inconsistent with all other estimates. ^{a} indicates that the point estimate is asymmetric with respect to the interval, so that $x \neq (a+b)/2$.

Species Category	Intervals		Midpoints	
	Mean lower bound	Mean upper bound	Mean	95% HPD
Arthropods	8.51	12.87	10.69	(5.09, 17.30)
Beetle	1.30	2.74	2.02	(0.40, 4.20)
Coral Reefs	1.95	2.19	2.07	(0.16, 3.68)
Insects	3.87	6.46	5.16	(3.86, 6.46)
Marine	4.12	4.47	4.30	(0.69, 9.04)
Other Arthropods	4.64	6.41	5.53	(0.08, 11.90)
Other Insects	2.57	3.72	3.15	(0.71, 5.07)
Other Global	7.62	8.27	7.94	(0.03, 19.50)
Other Marine	2.17	2.28	2.22	(0.01, 6.71)
Global	20.25	25.61	22.93	(11.24, 36.26)

Table 2 – Posterior point estimate summaries of species numbers in each category for both intervals and midpoints. Interval estimates are the posterior mean lower and upper interval bound. Midpoint estimates are the posterior mean and the 95% highest posterior density (HPD) interval. Point estimates are measured in millions.

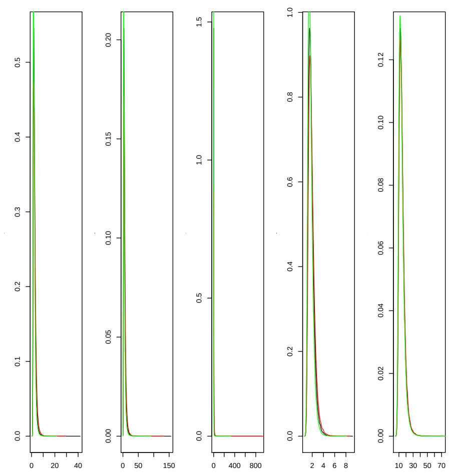

Figure A.4 – Posterior distribution for σ_{mj} fitted with 3 values of scale parameter (A) in the Half-Cauchy distribution. The black line represents $A = 2.5$, the red line represents $A = 5$ and the green line represents $A = 1.25$. From the left to the right, the posterior distribution is for the j^{th} "parent" species beetles, coral reefs, marine, insects and arthropods respectively.

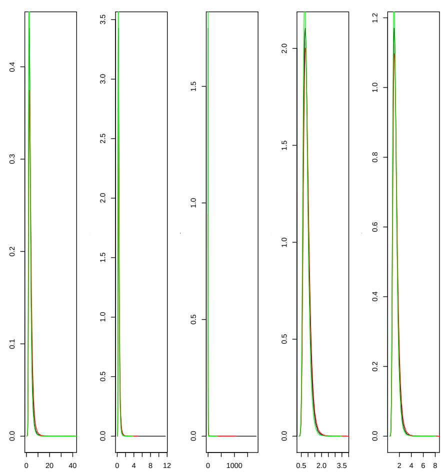

Figure A.5 – Posterior distribution for σ_{rj} fitted with 3 values of scale parameter (A) in the Half-Cauchy distribution. The black line represents $A = 2.5$, the red line represents $A = 5$ and the green line represents $A = 1.25$. From the left to the right, the posterior distribution is for the j^{th} "parent" species beetles, coral reefs, marine, insects and arthropods respectively.

Appendix B

Chapter 4 Supporting information

B.1 Proofs

B.1.1 Univariate intervals - Proof of Lemma 4.2

The main challenge is to derive the conditional density $f_{S|\boldsymbol{X}=\boldsymbol{z}}(s;\vartheta)$ in order to apply the methodology given in Proposition 4.1. Based on the the aggregation function (4.2) we have $S = (S_l, S_u, N) = (X_{(l)}, X_{(u)}, N)$ and thus the role of the conditional distribution of S given $\boldsymbol{X} = \boldsymbol{z}$ is to ensure that $s_l = z_{(l)}$ and $s_u = z_{(u)}$. As a consequence we can write

$$f_{S|\boldsymbol{X}=\boldsymbol{z}}(s;\vartheta) = \delta_{z_{(l)},z_{(u)}}(s_l, s_u) = \delta_{z_{(l)}}(s_l)\,\delta_{z_{(u)}}(s_u),$$

meaning that $l-1$ points of $\boldsymbol{z}$ belong to $(-\infty, s_l)$, one is at s_l, $u-l-1$ belong to (s_l, s_u), one is at s_u and $n-u$ belong to (s_u, ∞). As there is $n!/((l-1)!(u-l-1)!(n-u)!)$ possible combinations to arrange n points in such a way, the likelihood function can then be written as

$$\begin{aligned}
\mathcal{L}(s_l, s_u, n;\theta) =& \frac{n!}{(l-1)!(u-l-1)!(n-u)!}\left(\int_{-\infty}^{s_l} g_X(z;\theta)\mathrm{d}z\right)^{l-1}\int_{-\infty}^{+\infty} g_X(z;\theta)\delta_z(s_l)\mathrm{d}z \\
&\times\left(\int_{s_l}^{s_u} g_X(z;\theta)\mathrm{d}z\right)^{u-l-1}\int_{-\infty}^{+\infty} g_X(z;\theta)\delta_z(s_u)\mathrm{d}z\left(\int_{s_u}^{\infty} g_X(z;\theta)\mathrm{d}z\right)^{n-u} \\
=& \frac{n!}{(l-1)!(u-l-1)!(n-u)!}\left[G_X(s_l;\theta)\right]^{l-1}\left[G_X(s_u;\theta) - G_X(s_l;\theta)\right]^{u-l-1} \\
&\times\left[1 - G_X(s_u;\theta)\right]^{n-u} g_X(s_l;\theta) g_X(s_u;\theta),
\end{aligned}$$

using the independence between the n replicates $X_1, \ldots, X_n$.

B.1.2 Multivariate intervals - Details on Lemma 4.3 and Corollary 4.4

For simplicity, consider bivariate intervals, identical arguments can be applied to the multivariate setting. As X is here a bivariate random vector with p.d.f. $g_X(\cdot;\theta)$, its marginal and conditional p.d.f. are respectively denoted by $g_{X_i}(\cdot;\theta), i = 1,2$ and $g_{X_i|X_j}(\cdot;\theta), i,j = 1,2; i \neq j$. The conditional distribution of S given $\boldsymbol{X} = \boldsymbol{z} \in \mathbb{R}^2$ is obtained from the aggregation function (4.3). When $S_p = 2$, $S_{I_p} = \{\text{bltr}\}$ or $\{\text{tlbr}\}$ depending wether the two pints are in the bottom-left and top-right corners or top-left and bottom-right corners. Define $s_a = (s_{a_1}, s_{a_2})$ and $s_b = (s_{b_1}, s_{b_2})$ which take values $(s_{\min,1}, s_{\min,2})$ and $(s_{\max,1}, s_{\max,2})$ if $S_{I_p} = \{\text{bltr}\}$ and $(s_{\min,1}, s_{\max,2})$ and $(s_{\max,1}, s_{\min,2})$ if $S_{I_p} = \{\text{tlbr}\}$. We can then write

$$f_{S|\boldsymbol{X}=\boldsymbol{z}}(s;\vartheta) = \begin{cases} \delta_{z_{(1),1}, z_{(1),2}, z_{(n),1}, z_{(n),2}}\left(s_{a_1}, s_{a_2}, s_{b_1}, s_{b_2}\right) \\ \text{or} \\ \delta_{z_{(1),1}, z_{(n),2}, z_{(n),1}, z_{(1),2}}\left(s_{a_1}, s_{a_2}, s_{b_1}, s_{b_2}\right) \end{cases}.$$

Straightforwardly this ensures that two points gives the marginal minima and maxima and the remaining points are within the interval. Thus there are $n(n-1)$ possible combinations to arrange n points in such a way and the likelihood function is

$$\begin{aligned} \mathcal{L}(s;\theta) =& n(n-1)\left(\int_{s_{\min}}^{s_{\max}} g_X(z;\theta)\mathrm{d}z\right)^{n-2} \int_{\mathbb{R}^2} g_X(z;\theta)\delta_{s_a}(z)\mathrm{d}z \int_{\mathbb{R}^2} g_X(z;\theta)\delta_{s_b}(z)\mathrm{d}z \\ =& n(n-1)\left(\int_{s_{\min}}^{s_{\max}} g_X(z;\theta)\mathrm{d}z\right)^{n-2} g_X(s_a;\theta)g_X(s_b;\theta). \end{aligned}$$

When $S_p = 3$ and if $S_{I_p} = \{\text{bl}\}$ meaning that a single point is the minimum in both components and assuming its coordinate to be $s_c = s_{\min}$, then

$$f_{S|\boldsymbol{X}=\boldsymbol{z}}(s;\vartheta) = \delta_{z_{(1),1}, z_{(1),2}}(s_c)\delta_{(s_{\min,1}, s_{\max,1}),(s_{\min,2}, s_{\max,2})}\left(z_{j,1}|z_{j,2} = s_{\max,2}, z_{j,2}|z_{j,1} = s_{\max,1}\right),$$

where the second delta function is the product of Dirac measures defined for subset $A \subset \mathbb{R}$ by $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise. There are $n(n-1)(n-2)$ possible combinations to arrange n points such that one is at a corner, two are on two different edges and the rest is inside the interval. The likelihood is then

$$\begin{aligned} \mathcal{L}(s;\theta) =& n(n-1)(n-2)\int_{\mathbb{R}^2} g_X(z;\theta)\delta_{s_{\min}}(z)\mathrm{d}z \left(\int_{s_{\min,1}}^{s_{\max,1}} g_{X_1|X_2=s_{\max,2}}(z_1;\theta)\mathrm{d}z_1\right) g_{X_2}(s_{\max,2};\theta) \\ &\times \left(\int_{s_{\min,2}}^{s_{\max,2}} g_{X_2|X_1=s_{\max,1}}(z_2;\theta)\mathrm{d}z_2\right) g_{X_1}(s_{\max,1};\theta)\left(\int_{s_{\min}}^{s_{\max}} g_X(z;\theta)\mathrm{d}z\right)^{n-3} \\ =& n(n-1)(n-2)g_X(s_{\min};\theta)\left(\int_{s_{\min}}^{s_{\max}} g_X(z;\theta)\mathrm{d}z\right)^{n-3} \\ &\times \left[G_{X_1|X_2=s_{\max,2}}(s_{\max,1};\theta) - G_{X_1|X_2=s_{\max,2}}(s_{\min,1};\theta)\right] g_{X_2}(s_{\max,2};\theta) \\ &\times \left[G_{X_2|X_1=s_{\max,1}}(s_{\max,2};\theta) - G_{X_2|X_1=s_{\max,1}}(s_{\min,2};\theta)\right] g_{X_1}(s_{\max,1};\theta). \end{aligned}$$

Finally when $S_p = 4$ then

$$\begin{aligned} f_{S|\boldsymbol{X}=\boldsymbol{z}}(s;\vartheta) =& \delta_{(s_{\min,1}, s_{\max,1}),(s_{\min,1}, s_{\max,1})}\left(z_{j,1}|z_{j,2} = s_{\min,2}, z_{j,1}|z_{j,2} = z_{\max,2}\right) \\ &\times \delta_{(s_{\min,2}, s_{\max,2}),(s_{\min,2}, s_{\max,2})}\left(z_{j,2}|z_{j,1} = s_{\min,1}, z_{j,2}|z_{j,1} = s_{\max,1}\right), \end{aligned}$$

and there are $n(n-1)(n-2)(n-3)$ possible combinations to arrange four points on different edges and the rest inside the interval. The likelihood is then

$$\begin{aligned}
\mathcal{L}(s;\theta) =& n(n-1)(n-2)(n-3)\left(\int_{s_{\min}}^{s_{\max}} g_X(z;\theta)\mathrm{d}z\right)^{n-4} \\
&\times\left(\int_{s_{\min,1}}^{s_{\max,1}} g_{X_1|X_2=s_{\min,2}}(z_1;\theta)\mathrm{d}z_1\right) g_{X_2}(s_{\min,2};\theta) \\
&\times\left(\int_{s_{\min,1}}^{s_{\max,1}} g_{X_1|X_2=s_{\max,2}}(z_1;\theta)\mathrm{d}z_1\right) g_{X_2}(s_{\max,2};\theta) \\
&\times\left(\int_{s_{\min,2}}^{s_{\max,2}} g_{X_2|X_1=s_{\min,1}}(z_2;\theta)\mathrm{d}z_2\right) g_{X_1}(s_{\min,1};\theta) \\
&\times\left(\int_{s_{\min,2}}^{s_{\max,2}} g_{X_2|X_1=s_{\max,1}}(z_2;\theta)\mathrm{d}z_2\right) g_{X_1}(s_{\max,1};\theta) \\
=& n(n-1)(n-2)(n-3)\left(\int_{s_{\min}}^{s_{\max}} g_X(z;\theta)\mathrm{d}z\right)^{n-4} \\
&\times\left[G_{X_1|X_2=s_{\min,2}}(s_{\max,1};\theta) - G_{X_1|X_2=s_{\min,2}}(s_{\min,1};\theta)\right] g_{X_2}(s_{\min,2};\theta) \\
&\times\left[G_{X_1|X_2=s_{\max,2}}(s_{\max,1};\theta) - G_{X_1|X_2=s_{\max,2}}(s_{\min,1};\theta)\right] g_{X_2}(s_{\max,2};\theta) \\
&\times\left[G_{X_2|X_1=s_{\min,1}}(s_{\max,2};\theta) - G_{X_2|X_1=s_{\min,1}}(s_{\min,2};\theta)\right] g_{X_1}(s_{\min,1};\theta) \\
&\times\left[G_{X_2|X_1=s_{\max,1}}(s_{\max,2};\theta) - G_{X_2|X_1=s_{\max,1}}(s_{\min,2};\theta)\right] g_{X_1}(s_{\max,1};\theta).
\end{aligned}$$

B.1.3 Multivariate histograms with fixed bins

Details about the likelihood function of Lemma 4.9 are given here. Analogously to the case of intervals, the knowledge of the aggregation function π is used to define the conditional density $f_{S|\boldsymbol{X}=\boldsymbol{z}}(s;\vartheta)$ required to apply Proposition 4.1. As $S=\pi(\boldsymbol{X})$ by definition, then when $\boldsymbol{X}$ is known to be taking value $\boldsymbol{z}$, then $f_{S|\boldsymbol{X}=\boldsymbol{z}}$ can only exist if, for all $\boldsymbol{b}=\boldsymbol{1},\dots,\boldsymbol{B}$, $s_{\boldsymbol{b}}=\sum_{i=1}^{n}\mathbb{I}\{z_i\in\mathcal{B}_{\boldsymbol{b}}\}$, which is equivalent to writing

$$f_{S|\boldsymbol{X}=\boldsymbol{z}}(s;\vartheta)=\prod_{\boldsymbol{b}=\boldsymbol{1}}^{\boldsymbol{B}}\delta_{\sum_{i=1}^{n}\mathbb{I}\{z_i\in\mathcal{B}_{\boldsymbol{b}}\}}(s_{\boldsymbol{b}}).$$

The number of combinations to arrange $z_1,\dots,z_n$ into the $B_1\times\cdots\times B_{\boldsymbol{B}}$ bins is the multinomial coefficient equal to $n!/\prod_{\boldsymbol{b}} s_{\boldsymbol{b}}!$ and the likelihood function (4.1) becomes

$$\begin{aligned}
\mathcal{L}(s;\theta) =& \frac{n!}{s_{\boldsymbol{1}}!\cdots s_{\boldsymbol{B}}}\int_{\mathbb{R}^{n\times d}}\delta_{z_1}(\mathcal{B}_{\boldsymbol{1}})\cdots\delta_{z_{s_{\boldsymbol{1}}}}(\mathcal{B}_{\boldsymbol{1}})\cdots\delta_{z_{n-s_{\boldsymbol{B}}+1}}(\mathcal{B}_{\boldsymbol{B}})\cdots\delta_{z_n}(\mathcal{B}_{\boldsymbol{B}})\prod_{i=1}^{n}g_X(z_i;\theta)\mathrm{d}\boldsymbol{z} \\
=& \frac{n!}{s_{\boldsymbol{1}}!\cdots s_{\boldsymbol{B}}}\left(\int_{\mathbb{R}^d} g_X(z;\theta)\delta_z(\mathcal{B}_{\boldsymbol{1}})\mathrm{d}z\right)^{s_{\boldsymbol{1}}}\cdots\left(\int_{\mathbb{R}^d} g_X(z;\theta)\delta_z(\mathcal{B}_{\boldsymbol{B}})\mathrm{d}z\right)^{s_{\boldsymbol{B}}} \\
=& \frac{n!}{s_{\boldsymbol{1}}!\cdots s_{\boldsymbol{B}}}\prod_{\boldsymbol{b}=\boldsymbol{1}}^{\boldsymbol{B}}\left(\int_{\mathcal{B}_{\boldsymbol{b}}} g_X(z;\theta)\mathrm{d}z\right)^{s_{\boldsymbol{b}}}.
\end{aligned}$$

B.1.4 Histograms with fixed counts

Let $k=(k_1,\dots,k_B), 1\le k-1\le\cdot\le k_B\le n$ and consider the aggregation function (4.15). The histogram construction ensures that the B bins are defined by some order statistics. This implies that the symbol S provides the location of B out of n points and the number of points in between those is fixed and can be derived through k. As a consequence the conditional density $f_{S|\boldsymbol{X}=\boldsymbol{z}}(s;\vartheta)$ is

$$f_{S|\boldsymbol{X}=\boldsymbol{z}}(s;\vartheta) = \prod_{b=1}^{B} \delta_{z_{(k_b)}}(s_b) \prod_{b=1}^{B+1} \prod_{j=k_{b-1}}^{k_b-1} \delta_{z_{(j)}}\left((s_{b-1}, s_b)\right),$$

and there are $n!/\prod_{b=1}^{B+1}(k_b - k_{b-1} - 1)!$ possible combinations to arrange n points this way. Thus the likelihood function is then

$$\begin{aligned}
\mathcal{L}(s;\theta) =& \frac{n!}{\prod_{b=1}^{B+1}(k_b - k_{b-1} - 1)!} \int_{\mathbb{R}^n} \left(\prod_{b=1}^{B} \delta_{z_{(k_b)}}(s_b)\right) \prod_{b=1}^{B+1} \left(\prod_{j=k_{b-1}}^{k_b-1} \delta_{z_{(j)}}\left((s_{b-1}, s_b)\right)\right) \prod_{i=1}^{n} g_X(z_i;\theta)\mathrm{d}\boldsymbol{z} \\
=& \frac{n!}{\prod_{b=1}^{B+1}(k_b - k_{b-1} - 1)!} \prod_{b=1}^{B} \left(\int_{\mathbb{R}} \delta_z(s_b)\, g_X(z;\theta)\mathrm{d}z\right) \prod_{b=1}^{B+1} \left(\int_{s_{b-1}}^{s_b} g_X(z;\theta)\mathrm{d}z\right)^{k_b-k_{b-1}-1} \\
=& \frac{n!}{\prod_{b=1}^{B+1}(k_b - k_{b-1} - 1)!} \prod_{b=1}^{B} g_X(s_b;\theta) \prod_{b=1}^{B+1} \left(G_X(s_b;\theta) - G_X(s_{b-1};\theta)\right)^{k_b-k_{b-1}-1},
\end{aligned}$$

which proves Lemma 4.10.

B.2 Supplementary Material

B.2.1 Estimates of the μ_1, μ_2, σ_1 and σ_2, from Section 4.3.2

		$m = 20$				$m = 50$			
n_s		5	10	50	100	5	10	50	100
$\rho = 0.0$	$\mathcal{L}_4$	1.9991	2.0040	2.0002	2.0057	1.9991	2.0000	2.0024	2.0010
		(0.0506)	(0.0451)	(0.0330)	(0.0347)	(0.0310)	(0.0275)	(0.0189)	(0.0207)
	$\mathcal{L}_\varnothing$	1.9992	2.0042	2.0003	2.0057	1.9991	2.0000	2.0024	2.0010
		(0.0506)	(0.0451)	(0.0331)	(0.0346)	(0.0309)	(0.0274)	(0.0188)	(0.0207)
	$\mathcal{L}_{\text{full}}$	1.9991	2.0039	2.0001	2.0058	1.9991	2.0000	2.0024	2.0010
		(0.0506)	(0.0451)	(0.0329)	(0.0347)	(0.0309)	(0.0275)	(0.0189)	(0.0207)
0.3	$\mathcal{L}_4$	1.9955	1.9956	2.0005	1.9981	1.9962	1.9995	2.0014	1.9961
		(0.0524)	(0.0442)	(0.0339)	(0.0336)	(0.0345)	(0.0280)	(0.0208)	(0.0204)
	$\mathcal{L}_\varnothing$	1.9954	1.9957	2.0004	1.9980	1.9961	1.9994	2.0015	1.9960
		(0.0528)	(0.0441)	(0.0339)	(0.0337)	(0.0344)	(0.0277)	(0.0208)	(0.0204)
	$\mathcal{L}_{\text{full}}$	1.9955	1.9957	2.0007	1.9981	1.9962	1.9997	2.0015	1.9961
		(0.0528)	(0.0443)	(0.0342)	(0.0335)	(0.0344)	(0.0279)	(0.0208)	(0.0203)
0.5	$\mathcal{L}_4$	1.9950	1.9953	2.0001	1.9979	1.9964	1.9995	2.0011	1.9956
		(0.0533)	(0.0439)	(0.0355)	(0.0342)	(0.0346)	(0.0280)	(0.0208)	(0.0207)
	$\mathcal{L}_\varnothing$	1.9946	1.9955	2.0001	1.9979	1.9964	1.9995	2.0011	1.9956
		(0.0537)	(0.0437)	(0.0354)	(0.0342)	(0.0344)	(0.0279)	(0.0209)	(0.0207)
	$\mathcal{L}_{\text{full}}$	1.9956	1.9949	2.0002	1.9981	1.9962	2.0002	2.0011	1.9959
		(0.0536)	(0.0441)	(0.0361)	(0.0339)	(0.0341)	(0.0279)	(0.0210)	(0.0206)
0.7	$\mathcal{L}_4$	1.9943	1.9960	2.0001	1.9968	1.9966	1.9990	2.0008	1.9951
		(0.0539)	(0.0435)	(0.0365)	(0.0339)	(0.0349)	(0.0273)	(0.0215)	(0.0203)
	$\mathcal{L}_\varnothing$	1.9943	1.9959	2.0001	1.9967	1.9966	1.9991	2.0008	1.9951
		(0.0547)	(0.0430)	(0.0366)	(0.0339)	(0.0350)	(0.0277)	(0.0214)	(0.0204)
	$\mathcal{L}_{\text{full}}$	1.9951	1.9953	1.9999	1.9980	1.9970	1.9999	2.0007	1.9953
		(0.0540)	(0.0428)	(0.0369)	(0.0325)	(0.0337)	(0.0273)	(0.0215)	(0.0200)
0.9	$\mathcal{L}_4$	1.9932	1.9985	1.9980	1.9960	1.9969	1.9994	1.9994	1.9951
		(0.0539)	(0.0435)	(0.0366)	(0.0327)	(0.0349)	(0.0268)	(0.0227)	(0.0196)
	$\mathcal{L}_\varnothing$	1.9930	1.9971	1.9978	1.9961	1.9968	1.9990	1.9995	1.9955
		(0.0548)	(0.0437)	(0.0370)	(0.0328)	(0.0350)	(0.0270)	(0.0231)	(0.0193)
	$\mathcal{L}_{\text{full}}$	1.9940	1.9983	1.9997	1.9982	1.9968	2.0012	2.0003	1.9959
		(0.0539)	(0.0408)	(0.0364)	(0.0315)	(0.0345)	(0.0258)	(0.0205)	(0.0191)

Table 1 – Mean estimate (and standard deviation) of the mean μ_1 over 100 replicates using the $\mathcal{L}_4$, $\mathcal{L}_\varnothing$ and $\mathcal{L}_{\text{full}}$ likelihood function with $m = 20$ and 50 symbols aggregating $n_s = 5, 10, 50$ and 100 observations.

B.2.2 Visualisation of the constructions of bivariate intervals from order statistics used in Section 4.3.2

B.2.3 Estimates of $(\sigma_1, \rho, \sigma_2)$, from Section 4.3.2

B.2.4 Visualisation of some symbolic datasets used in Section 4.3.2 when $\rho = -0.7$ and 0

		$m = 20$				$m = 50$			
n_s		5	10	50	100	5	10	50	100
$\rho = 0.0$	$\mathcal{L}_4$	4.9934	5.0004	4.9974	4.9966	4.9976	4.9999	4.9984	4.9965
		(0.0533)	(0.0455)	(0.0358)	(0.0316)	(0.0343)	(0.0263)	(0.0228)	(0.0184)
	$\mathcal{L}_\varnothing$	4.9932	5.0005	4.9974	4.9966	4.9975	4.9998	4.9984	4.9964
		(0.0534)	(0.0455)	(0.0358)	(0.0316)	(0.0347)	(0.0262)	(0.0228)	(0.0184)
	$\mathcal{L}_{\text{full}}$	4.9933	5.0005	4.9975	4.9967	4.9976	4.9999	4.9984	4.9965
		(0.0531)	(0.0456)	(0.0357)	(0.0315)	(0.0341)	(0.0263)	(0.0228)	(0.0184)
0.3	$\mathcal{L}_4$	4.9988	5.0007	4.9972	5.0039	4.9999	5.0024	4.9989	5.0015
		(0.0534)	(0.0451)	(0.0320)	(0.0322)	(0.0327)	(0.0276)	(0.0207)	(0.0174)
	$\mathcal{L}_\varnothing$	4.9987	5.0007	4.9972	5.0039	4.9997	5.0025	4.9990	5.0015
		(0.0530)	(0.0452)	(0.0320)	(0.0322)	(0.0329)	(0.0275)	(0.0207)	(0.0174)
	$\mathcal{L}_{\text{full}}$	4.9989	5.0006	4.9976	5.0039	4.9997	5.0028	4.9991	5.0015
		(0.0533)	(0.0448)	(0.0317)	(0.0322)	(0.0325)	(0.0273)	(0.0205)	(0.0174)
0.5	$\mathcal{L}_4$	4.9982	5.0010	4.9973	5.0043	5.0001	5.0017	4.9981	5.0014
		(0.0531)	(0.0466)	(0.0330)	(0.0319)	(0.0326)	(0.0277)	(0.0209)	(0.0176)
	$\mathcal{L}_\varnothing$	4.9982	5.0011	4.9972	5.0044	5.0003	5.0019	4.9981	5.0014
		(0.0531)	(0.0463)	(0.0329)	(0.0319)	(0.0327)	(0.0275)	(0.0209)	(0.0176)
	$\mathcal{L}_{\text{full}}$	4.9987	5.0008	4.9976	5.0045	4.9996	5.0024	4.9983	5.0016
		(0.0524)	(0.0461)	(0.0332)	(0.0320)	(0.0321)	(0.0271)	(0.0207)	(0.0175)
0.7	$\mathcal{L}_4$	4.9970	5.0013	4.9979	5.0029	5.0001	5.0012	4.9977	5.0010
		(0.0532)	(0.0471)	(0.0343)	(0.0316)	(0.0329)	(0.0283)	(0.0218)	(0.0177)
	$\mathcal{L}_\varnothing$	4.9975	5.0012	4.9980	5.0029	4.9999	5.0012	4.9978	5.0010
		(0.0532)	(0.0473)	(0.0344)	(0.0316)	(0.0331)	(0.0283)	(0.0219)	(0.0177)
	$\mathcal{L}_{\text{full}}$	4.9981	5.0013	4.9981	5.0041	5.0001	5.0023	4.9981	5.0009
		(0.0524)	(0.0464)	(0.0341)	(0.0308)	(0.0315)	(0.0275)	(0.0210)	(0.0177)
0.9	$\mathcal{L}_4$	4.9946	5.0009	4.9980	4.9999	4.9990	5.0003	4.9975	4.9996
		(0.0534)	(0.0475)	(0.0348)	(0.0312)	(0.0337)	(0.0279)	(0.0228)	(0.0174)
	$\mathcal{L}_\varnothing$	4.9944	5.0010	4.9978	4.9999	4.9986	4.9998	4.9974	4.9997
		(0.0536)	(0.0476)	(0.0351)	(0.0317)	(0.0335)	(0.0283)	(0.0226)	(0.0171)
	$\mathcal{L}_{\text{full}}$	4.9953	5.0018	4.9999	5.0017	4.9986	5.0020	4.9991	4.9995
		(0.0524)	(0.0450)	(0.0342)	(0.0290)	(0.0329)	(0.0268)	(0.0211)	(0.0179)

Table 2 – Mean estimate (and standard deviation) of the mean μ_2 over 100 replicates using the $\mathcal{L}_4$, $\mathcal{L}_\varnothing$ and $\mathcal{L}_{\text{full}}$ likelihood function with $m = 20$ and 50 symbols aggregating $n_s = 5, 10, 50$ and 100 observations.

		m = 20				m = 50			
n_s		5	10	50	100	5	10	50	100
$\rho = 0.0$	$\mathcal{L}_4$	0.2470	0.2472	0.2469	0.2498	0.2468	0.2468	0.2485	0.2494
		(0.0394)	(0.0287)	(0.0163)	(0.0152)	(0.0242)	(0.0176)	(0.0106)	(0.0086)
	$\mathcal{L}_\varnothing$	0.2492	0.2471	0.2469	0.2498	0.2482	0.2470	0.2485	0.2493
		(0.0393)	(0.0288)	(0.0163)	(0.0152)	(0.0241)	(0.0177)	(0.0106)	(0.0086)
	$\mathcal{L}_{\text{full}}$	0.2467	0.2471	0.2469	0.2498	0.2467	0.2467	0.2484	0.2494
		(0.0394)	(0.0287)	(0.0163)	(0.0152)	(0.0241)	(0.0176)	(0.0106)	(0.0086)
0.3	$\mathcal{L}_4$	0.2502	0.2552	0.2491	0.2482	0.2464	0.2512	0.2489	0.2491
		(0.0436)	(0.0312)	(0.0167)	(0.0136)	(0.0257)	(0.0177)	(0.0098)	(0.0082)
	$\mathcal{L}_\varnothing$	0.2534	0.2553	0.2491	0.2482	0.2492	0.2512	0.2489	0.2491
		(0.0443)	(0.0311)	(0.0166)	(0.0136)	(0.0259)	(0.0177)	(0.0098)	(0.0082)
	$\mathcal{L}_{\text{full}}$	0.2496	0.2549	0.2491	0.2482	0.2460	0.2510	0.2488	0.2492
		(0.0434)	(0.0311)	(0.0167)	(0.0135)	(0.0256)	(0.0176)	(0.0098)	(0.0082)
0.5	$\mathcal{L}_4$	0.2512	0.2555	0.2495	0.2477	0.2478	0.2517	0.2490	0.2488
		(0.0428)	(0.0306)	(0.0169)	(0.0130)	(0.0251)	(0.0182)	(0.0098)	(0.0082)
	$\mathcal{L}_\varnothing$	0.2566	0.2552	0.2495	0.2477	0.2519	0.2517	0.2491	0.2488
		(0.0445)	(0.0304)	(0.0169)	(0.0130)	(0.0257)	(0.0182)	(0.0098)	(0.0082)
	$\mathcal{L}_{\text{full}}$	0.2497	0.2548	0.2497	0.2477	0.2466	0.2513	0.2490	0.2488
		(0.0424)	(0.0305)	(0.0169)	(0.0129)	(0.0249)	(0.0180)	(0.0098)	(0.0082)
0.7	$\mathcal{L}_4$	0.2531	0.2562	0.2501	0.2477	0.2504	0.2528	0.2494	0.2490
		(0.0419)	(0.0294)	(0.0173)	(0.0125)	(0.0246)	(0.0186)	(0.0099)	(0.0081)
	$\mathcal{L}_\varnothing$	0.2592	0.2552	0.2499	0.2477	0.2578	0.2521	0.2493	0.2489
		(0.0429)	(0.0294)	(0.0173)	(0.0126)	(0.0245)	(0.0185)	(0.0099)	(0.0081)
	$\mathcal{L}_{\text{full}}$	0.2492	0.2546	0.2503	0.2475	0.2471	0.2515	0.2494	0.2490
		(0.0411)	(0.0287)	(0.0173)	(0.0122)	(0.0241)	(0.0180)	(0.0098)	(0.0079)
0.9	$\mathcal{L}_4$	0.2595	0.2583	0.2510	0.2482	0.2584	0.2562	0.2501	0.2494
		(0.0411)	(0.0289)	(0.0172)	(0.0127)	(0.0244)	(0.0197)	(0.0103)	(0.0080)
	$\mathcal{L}_\varnothing$	0.2566	0.2534	0.2481	0.2466	0.2535	0.2513	0.2476	0.2486
		(0.0399)	(0.0273)	(0.0158)	(0.0126)	(0.0223)	(0.0178)	(0.0100)	(0.0079)
	$\mathcal{L}_{\text{full}}$	0.2489	0.2532	0.2507	0.2477	0.2481	0.2512	0.2495	0.2491
		(0.0394)	(0.0267)	(0.0172)	(0.0116)	(0.0234)	(0.0182)	(0.0098)	(0.0077)

Table 3 – Mean estimate (and standard deviation) of the standard deviation σ_1 over 100 replicates using the $\mathcal{L}_4$, $\mathcal{L}_\varnothing$ and $\mathcal{L}_{\text{full}}$ likelihood function with m = 20 and 50 symbols aggregating $n_s = 5, 10, 50$ and 100 observations.

		$m = 20$				$m = 50$			
n_s		5	10	50	100	5	10	50	100
$\rho = 0.0$	$\mathcal{L}_4$	0.2508	0.2512	0.2500	0.2496	0.2511	0.2506	0.2494	0.2502
		(0.0377)	(0.0285)	(0.0155)	(0.0125)	(0.0229)	(0.0194)	(0.0098)	(0.0078)
	$\mathcal{L}_\varnothing$	0.2535	0.2512	0.2499	0.2496	0.2528	0.2505	0.2494	0.2502
		(0.0386)	(0.0286)	(0.0154)	(0.0125)	(0.0232)	(0.0193)	(0.0098)	(0.0078)
	$\mathcal{L}_{\text{full}}$	0.2505	0.2510	0.2499	0.2496	0.2510	0.2505	0.2494	0.2502
		(0.0376)	(0.0285)	(0.0154)	(0.0125)	(0.0230)	(0.0193)	(0.0098)	(0.0078)
0.3	$\mathcal{L}_4$	0.2504	0.2466	0.2491	0.2500	0.2536	0.2497	0.2492	0.2506
		(0.0324)	(0.0262)	(0.0156)	(0.0111)	(0.0241)	(0.0195)	(0.0099)	(0.0074)
	$\mathcal{L}_\varnothing$	0.2536	0.2466	0.2491	0.2499	0.2565	0.2498	0.2492	0.2506
		(0.0330)	(0.0263)	(0.0156)	(0.0112)	(0.0246)	(0.0196)	(0.0099)	(0.0074)
	$\mathcal{L}_{\text{full}}$	0.2497	0.2463	0.2491	0.2500	0.2531	0.2496	0.2492	0.2506
		(0.0325)	(0.0262)	(0.0156)	(0.0112)	(0.0239)	(0.0193)	(0.0099)	(0.0074)
0.5	$\mathcal{L}_4$	0.2515	0.2473	0.2493	0.2502	0.2546	0.2504	0.2493	0.2508
		(0.0327)	(0.0269)	(0.0155)	(0.0106)	(0.0241)	(0.0199)	(0.0097)	(0.0069)
	$\mathcal{L}_\varnothing$	0.2568	0.2470	0.2493	0.2502	0.2586	0.2503	0.2493	0.2508
		(0.0338)	(0.0268)	(0.0155)	(0.0105)	(0.0245)	(0.0201)	(0.0097)	(0.0069)
	$\mathcal{L}_{\text{full}}$	0.2500	0.2466	0.2494	0.2502	0.2532	0.2500	0.2493	0.2508
		(0.0327)	(0.0267)	(0.0156)	(0.0106)	(0.0238)	(0.0196)	(0.0097)	(0.0070)
0.7	$\mathcal{L}_4$	0.2538	0.2486	0.2494	0.2503	0.2565	0.2515	0.2495	0.2509
		(0.0334)	(0.0282)	(0.0153)	(0.0106)	(0.0242)	(0.0203)	(0.0095)	(0.0070)
	$\mathcal{L}_\varnothing$	0.2601	0.2478	0.2493	0.2502	0.2636	0.2506	0.2495	0.2509
		(0.0351)	(0.0285)	(0.0152)	(0.0106)	(0.0241)	(0.0204)	(0.0095)	(0.0070)
	$\mathcal{L}_{\text{full}}$	0.2500	0.2470	0.2496	0.2501	0.2530	0.2502	0.2495	0.2509
		(0.0327)	(0.0274)	(0.0155)	(0.0106)	(0.0236)	(0.0196)	(0.0094)	(0.0071)
0.9	$\mathcal{L}_4$	0.2600	0.2531	0.2499	0.2505	0.2623	0.2555	0.2501	0.2511
		(0.0363)	(0.0300)	(0.0150)	(0.0111)	(0.0244)	(0.0210)	(0.0092)	(0.0073)
	$\mathcal{L}_\varnothing$	0.2571	0.2479	0.2472	0.2488	0.2571	0.2504	0.2476	0.2502
		(0.0351)	(0.0278)	(0.0142)	(0.0109)	(0.0220)	(0.0191)	(0.0092)	(0.0072)
	$\mathcal{L}_{\text{full}}$	0.2492	0.2480	0.2495	0.2501	0.2515	0.2502	0.2493	0.2506
		(0.0338)	(0.0277)	(0.0152)	(0.0105)	(0.0230)	(0.0192)	(0.0091)	(0.0071)

Table 4 – Mean estimate (and standard deviation) of the standard deviation σ_2 over 100 replicates using the $\mathcal{L}_4$, $\mathcal{L}_\varnothing$ and $\mathcal{L}_{\text{full}}$ likelihood function with $m = 20$ and 50 symbols aggregating $n_s = 5, 10, 50$ and 100 observations.

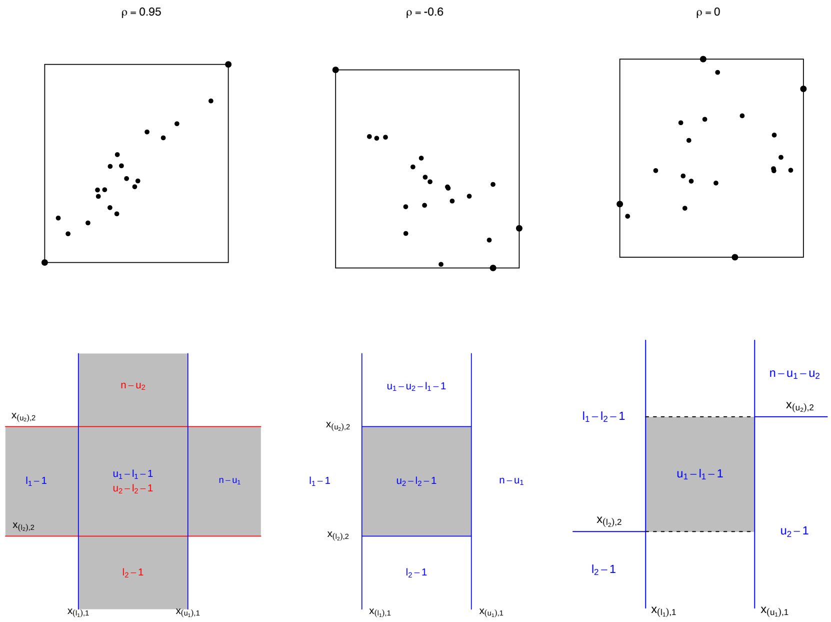

Figure B.1 – Construction methods for bivariate intervals using marginal minima/maxima (top panels) or marginal order statistics (bottom). Top panels: Illustrative random rectangles constructed from 2 points (high correlation), 3 points (moderate correlation) and 4 points (low/no correlation). Bottom panels: Three alternative construction methods: marginal only (left panel), sequential nesting (centre; equation (4.9)) and iterative segmentation (right; equation (4.11)). Values in blue (red) denote the number of observations in the area bounded by blue (red) lines.

		$n_s = 60$			$n_s = 300$		
	Orders	σ_1	ρ	σ_2	σ_1	ρ	σ_2
$\mathcal{L}_{1x}$	$(6, 55, 3, 3)$	0.4975	−0.7133	0.4929	0.4999	−0.7133	0.4896
		(0.0121)	(0.0497)	(0.0393)	(0.0060)	(0.0472)	(0.0320)
	$(16, 45, 10, 2)$	0.4987	−0.7325	0.4966	0.4994	−0.7215	0.4983
		(0.0162)	(0.0932)	(0.0277)	(0.0075)	(0.1051)	(0.0248)
	$(20, 41, 7, 14)$	0.5004	−0.7108	0.4869	0.4993	−0.7128	0.4771
		(0.0180)	(0.0363)	(0.0444)	(0.0080)	(0.0275)	(0.0453)
$\mathcal{L}_{1y}$	$(3, 3, 6, 55)$	0.4900	−0.7130	0.4993	0.4915	−0.7127	0.4984
		(0.0288)	(0.0517)	(0.0147)	(0.0326)	(0.0447)	(0.0061)
	$(10, 2, 16, 45)$	0.4915	−0.7327	0.4982	0.4955	−0.7284	0.4985
		(0.0228)	(0.1020)	(0.0163)	(0.0238)	(0.0999)	(0.0077)
	$(7, 14, 20, 41)$	0.4802	−0.7155	0.4990	0.4850	−0.7096	0.4981
		(0.0424)	(0.0335)	(0.0205)	(0.0401)	(0.0253)	(0.0101)
$\mathcal{L}_{2x}$	$(6, 55, 5, 35)$	0.4974	−0.6912	0.5106	0.4998	−0.6596	0.5040
		(0.0124)	(0.2625)	(0.0472)	(0.0060)	(0.2790)	(0.0410)
	$(16, 45, 6, 24)$	0.4986	−0.6933	0.5289	0.4994	−0.6606	0.5144
		(0.0164)	(0.1854)	(0.0949)	(0.0075)	(0.2146)	(0.0856)
	$(20, 41, 5, 16)$	0.5004	−0.6699	0.5231	0.4993	−0.6790	0.5201
		(0.0184)	(0.1963)	(0.0987)	(0.0080)	(0.1753)	(0.0919)
$\mathcal{L}_{2y}$	$(5, 35, 6, 55)$	0.4979	−0.6423	0.4993	0.5006	−0.6405	0.4984
		(0.0394)	(0.2486)	(0.0148)	(0.0364)	(0.2746)	(0.0061)
	$(6, 24, 16, 45)$	0.5060	−0.6447	0.4981	0.5223	−0.6726	0.4985
		(0.0859)	(0.2168)	(0.0162)	(0.0910)	(0.2231)	(0.0078)
	$(5, 16, 20, 41)$	0.5054	−0.6396	0.4991	0.5141	−0.6451	0.4981
		(0.0999)	(0.1981)	(0.0206)	(0.1018)	(0.2205)	(0.0101)

Table 5 – Mean estimate (and standard deviation) of ($\sigma_1 = 0.5, \rho = -0.7, \sigma_2 = 0.5$) over 100 replicates using the $\mathcal{L}_{1x}, \mathcal{L}_{1y}, \mathcal{L}_{2x}$ and $\mathcal{L}_{2y}$ likelihood functions with $m = 20$ symbols aggregating $n_s = 60$ and 300 observations. The orders are multiplied by 5 for $n_s = 300$.

		$n_s = 60$			$n_s = 300$		
	Orders	σ_1	ρ	σ_2	σ_1	ρ	σ_2
$\mathcal{L}_{1x}$	$(6, 55, 3, 3)$	0.4980	−0.0048	0.4856	0.4998	0.0008	0.4838
		(0.0126)	(0.0519)	(0.0546)	(0.0059)	(0.0205)	(0.0584)
	$(16, 45, 10, 2)$	0.4968	−0.0353	0.4777	0.5001	−0.0260	0.4881
		(0.0157)	(0.0828)	(0.0520)	(0.0076)	(0.0653)	(0.0514)
	$(20, 41, 7, 14)$	0.4957	0.0214	0.4846	0.4990	0.0184	0.4829
		(0.0186)	(0.0618)	(0.0547)	(0.0099)	(0.0566)	(0.0527)
$\mathcal{L}_{1y}$	$(3, 3, 6, 55)$	0.4830	0.0074	0.4984	0.4775	0.0004	0.4995
		(0.0538)	(0.0524)	(0.0141)	(0.0516)	(0.0272)	(0.0058)
	$(10, 2, 16, 45)$	0.5006	−0.0055	0.4984	0.4762	−0.0391	0.4986
		(0.0491)	(0.0752)	(0.0151)	(0.0563)	(0.0743)	(0.0063)
	$(7, 14, 20, 41)$	0.4804	0.0270	0.5005	0.4852	0.0163	0.4984
		(0.0577)	(0.0697)	(0.0174)	(0.0494)	(0.0525)	(0.0089)
$\mathcal{L}_{2x}$	$(6, 55, 5, 35)$	0.4980	0.0183	0.5235	0.4998	−0.0191	0.5216
		(0.0126)	(0.4156)	(0.0322)	(0.0059)	(0.3888)	(0.0301)
	$(16, 45, 6, 24)$	0.4968	0.0670	0.5329	0.5001	−0.0172	0.5307
		(0.0157)	(0.3490)	(0.0612)	(0.0076)	(0.3375)	(0.0572)
	$(20, 41, 5, 16)$	0.4957	0.0847	0.5394	0.4990	−0.0018	0.5355
		(0.0186)	(0.3747)	(0.0551)	(0.0099)	(0.3671)	(0.0508)
$\mathcal{L}_{2y}$	$(5, 35, 6, 55)$	0.5252	0.0024	0.4983	0.5235	0.0261	0.4995
		(0.0412)	(0.4303)	(0.0142)	(0.0306)	(0.4018)	(0.0058)
	$(6, 24, 16, 45)$	0.5382	−0.0048	0.4983	0.5359	0.0240	0.4986
		(0.0532)	(0.3863)	(0.0151)	(0.0558)	(0.3647)	(0.0063)
	$(5, 16, 20, 41)$	0.5343	−0.0024	0.5005	0.5434	0.0080	0.4984
		(0.0586)	(0.3645)	(0.0174)	(0.0569)	(0.3855)	(0.0089)

Table 6 – Mean estimate (and standard deviation) of the $(\sigma_1 = 0.5, \rho = 0, \sigma_2 = 0.5)$ over 100 replicates using the $\mathcal{L}_{1x}, \mathcal{L}_{1y}, \mathcal{L}_{2x}$ and $\mathcal{L}_{2y}$ likelihood functions with $m = 20$ symbols aggregating $n_s = 60$ and 300 observations.

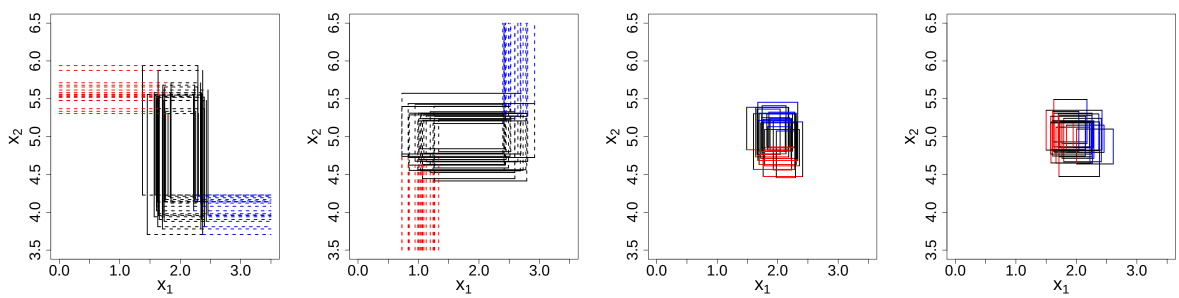

Figure B.2 – Symbolic datasets with $m = 20$ resulting from the aggregation of bivariate normal data with correlation $\rho = -0.7$, using (4.11) (left) and (4.9) (right). The red and blue colours represent $x_{(l_2),2}$ and $x_{(u_2),2}$ the first and third panels and $x_{(l_1),1}$ and $x_{(u_1),1}$ for the second and fourth panels. From left to right, the orders are $(16, 45, 10, 2), (10, 2, 16, 45), (20, 41, 5, 16)$ and $5, 16, 20, 41$.

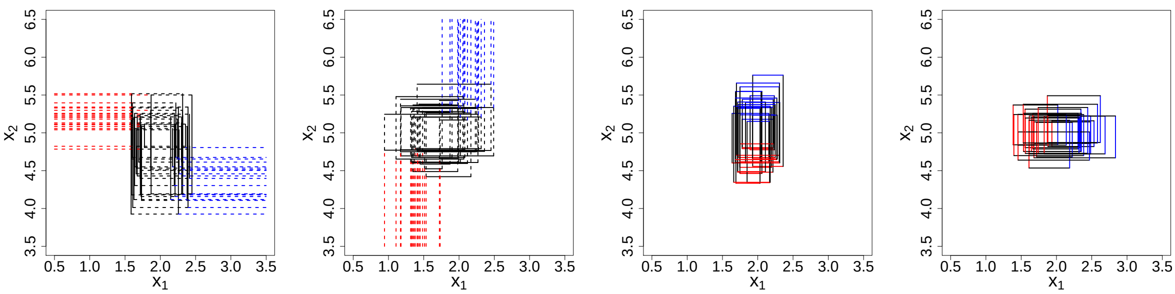

Figure B.3 – Symbolic datasets with $m = 20$ resulting from the aggregation of bivariate normal data with correlation $\rho = 0$, using (4.11) (left) and (4.9) (right). The red and blue colours represent $x_{(l_2),2}$ and $x_{(u_2),2}$ the first and third panels and $x_{(l_1),1}$ and $x_{(u_1),1}$ for the second and fourth panels. From left to right, the orders are $(16, 45, 10, 2), (10, 2, 16, 45), (20, 41, 5, 16)$ and $5, 16, 20, 41$.

Appendix C

Chapter 5: Simulation and Real data analysis graphs

C.1 Simulation outputs for 1-component Gaussian model

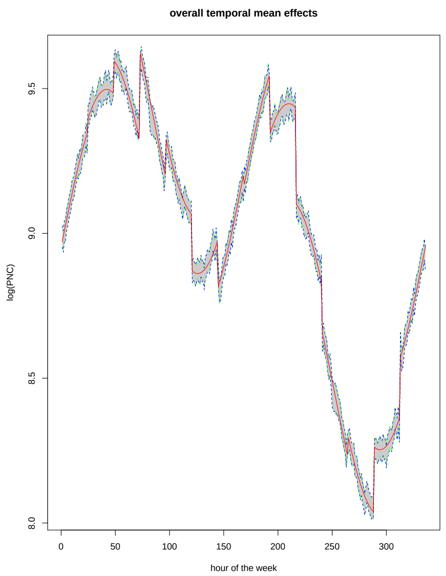

Figure C.1 – The simulated overall mean trend $(\alpha + \beta_t + (B\theta)_t)$, shown in red dashed line
The green dashed lines are posterior bounds of 95% credible interval fitted using Equation (5.2) with 95% credible interval region shaded in grey. The blue dashed lines are posterior bounds of 95% credible interval fitted using symbolic likelihood function in Equation (5.3.1).

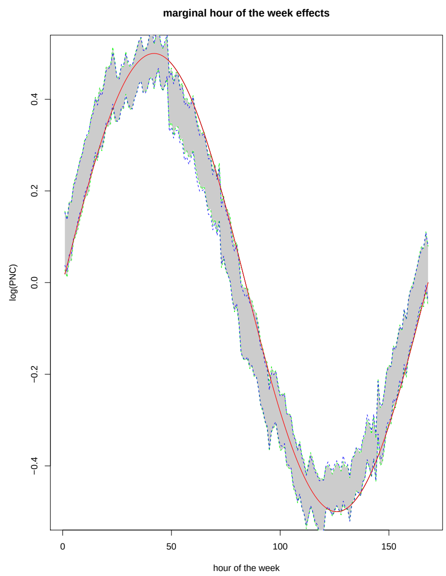

Figure C.2 – The simulated marginal joint daily-weekly effects β_t which repeats weekly, shown in red line
The green dashed lines are posterior bounds of 95% credible interval fitted using Equation (5.2) with 95% credible interval region shaded in grey. The blue dashed lines are posterior bounds of 95% credible interval fitted using symbolic likelihood function in Equation (5.3.1).

C.2 Simulation outputs for 2-component Gaussian model

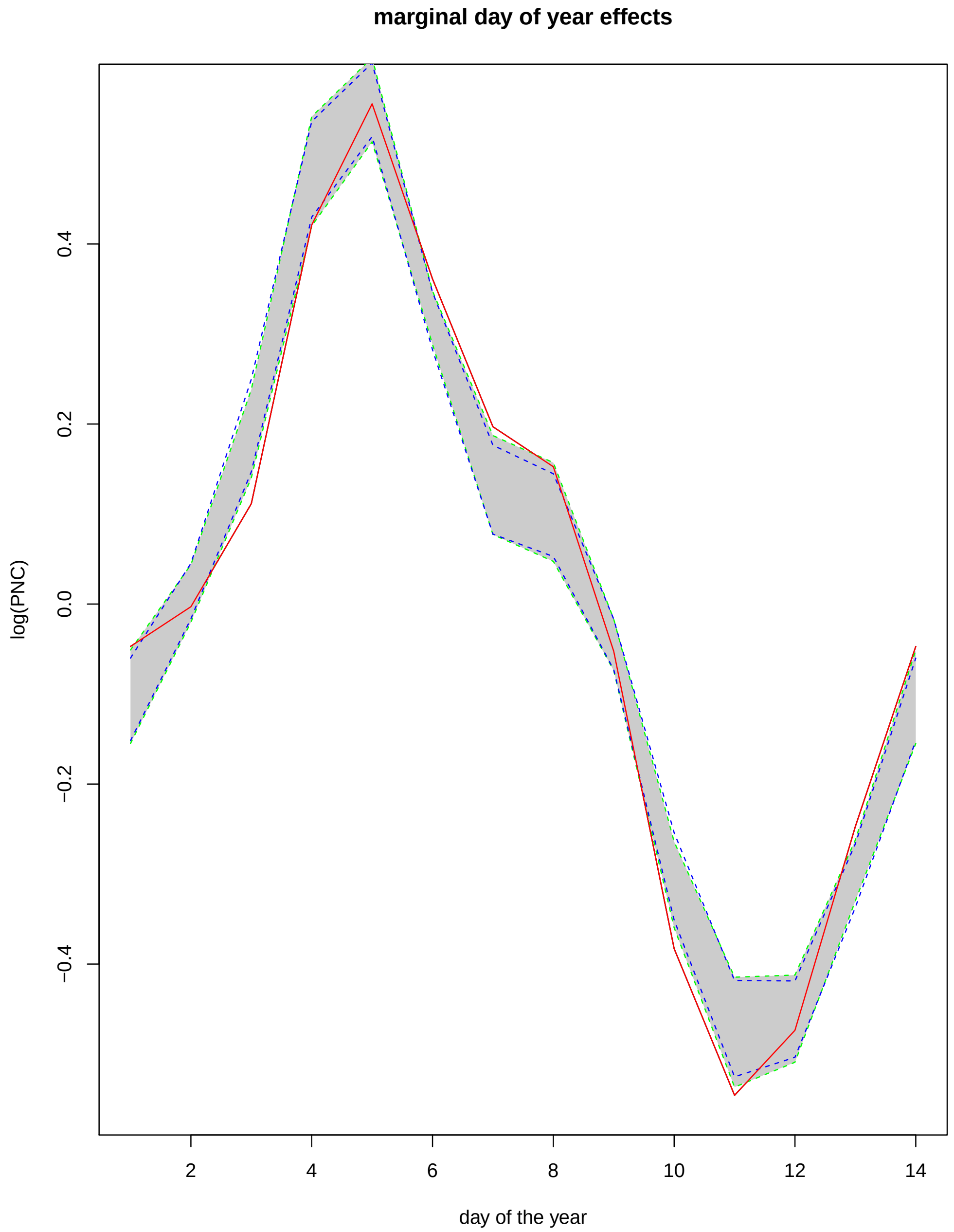

Figure C.3 – The simulated marginal annual effects $(B\theta)_t$ over a 2-week period, shown in red line.
The green dashed lines are posterior bounds of 95% credible interval fitted using Equation (5.2) with 95% credible interval region shaded in grey. The blue dashed lines are posterior bounds of 95% credible interval fitted using symbolic likelihood function in Equation (5.3.1).

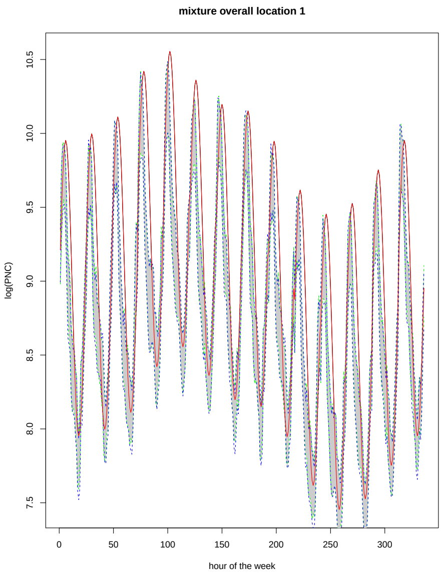

Figure C.4 – The simulated mixture locations 1 $(\alpha_1 + \beta_{t1} + (B\theta)_t)$, shown in red line The green dashed lines are posterior bounds of 95% credible interval from Equation (5.3) with 95% credible interval region shaded in grey. The blue dashed lines are posterior bounds of 95% credible interval fitted using symbolic likelihood function in Equation (5.3.1).

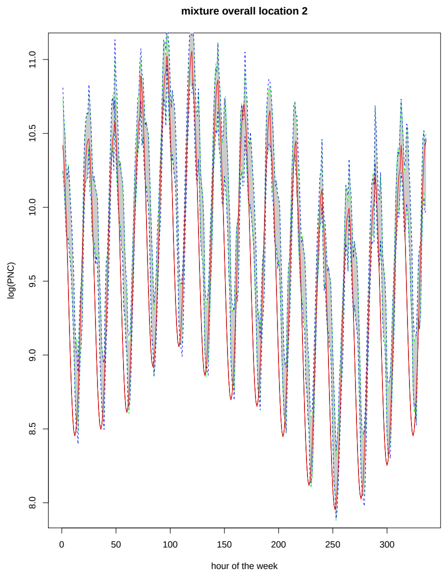

Figure C.5 – The simulated mixture locations 2 $(\alpha_2 + \beta_t + (B\theta)_t)$, shown in red line The green dashed lines are posterior bounds of 95% credible interval from Equation (5.3) with 95% credible interval region shaded in grey. The blue dashed lines are posterior bounds of 95% credible interval fitted using symbolic likelihood function in Equation (5.3.1).

C.3 Simulation outputs for Model Selection

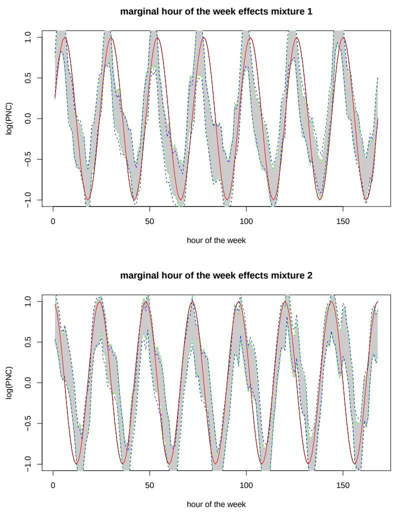

Figure C.6 – The simulated marginal joint daily-weekly $\beta = (\beta_{t1}, \beta_{t2})$, shown in red lines.
The green dashed lines are posterior bounds of 95% credible interval from Equation (5.3) with 95% credible interval region shaded in grey. The blue dashed lines are posterior bounds of 95% credible interval fitted using symbolic likelihood function in Equation (5.3.1).

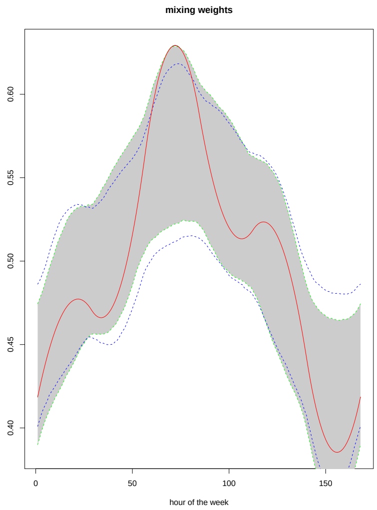

Figure C.7 – The simulated time-varying mixing weights in red line.
The green dashed lines are posterior bounds of 95% credible interval from Equation (5.3) with 95% credible interval region shaded in grey. The blue dashed lines are posterior bounds of 95% credible interval fitted using symbolic likelihood function in Equation (5.3.1).

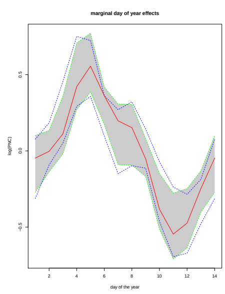

Figure C.8 – The simulated marginal annual effects shown in red. The green dashed lines are posterior bounds of 95% credible interval from Equation (5.3) with 95% credible interval region shaded in grey. The blue dashed lines are posterior bounds of 95% credible interval fitted using symbolic likelihood function in Equation (5.3.1).

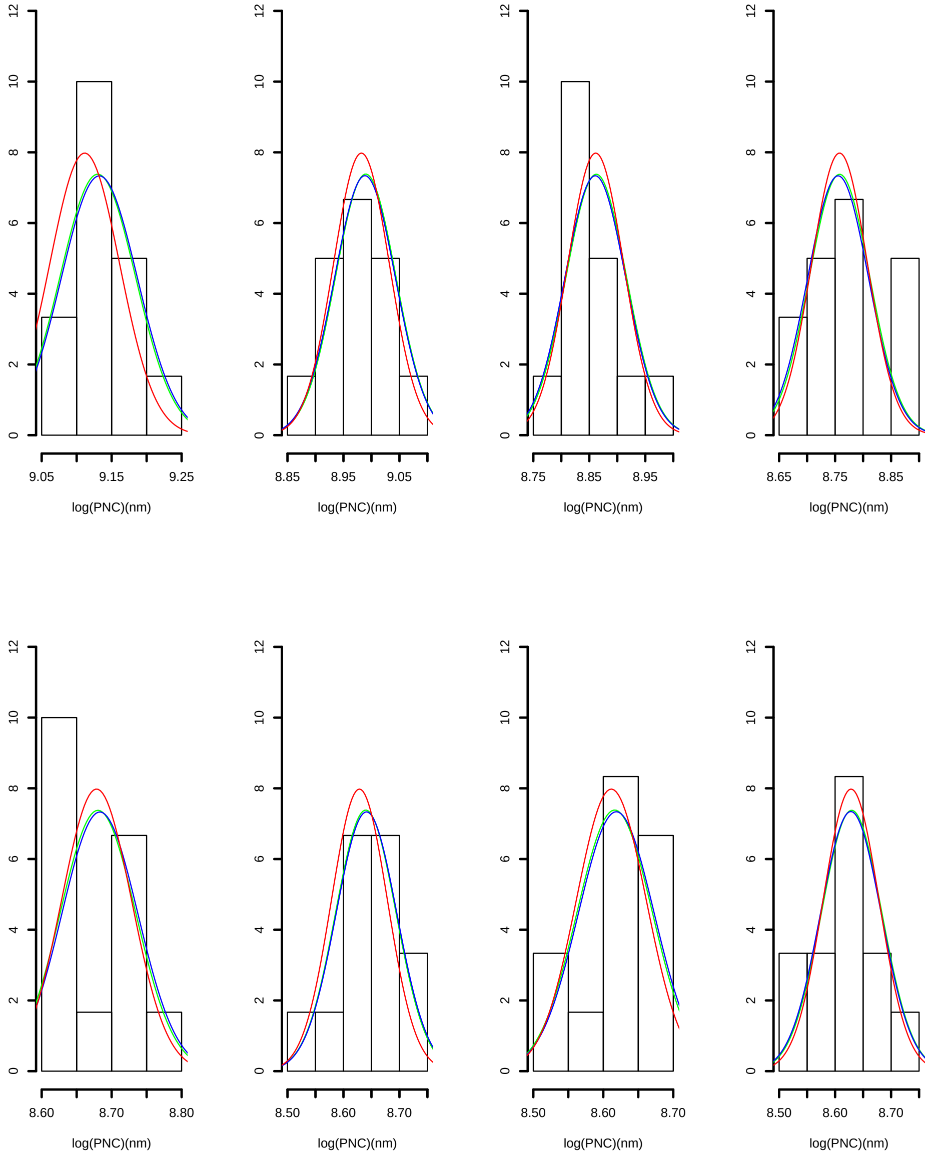

Figure C.9 – The fitted hourly density for simulations described in Section 5.4.5. The red line represents the true density (**1-component Gaussian distribution**) and the green line represents the posterior density from **correct 1-component model fitted with classical data** while the blue line represents the posterior density from **correct 1-component model fitted with 5-quantile histogram**.

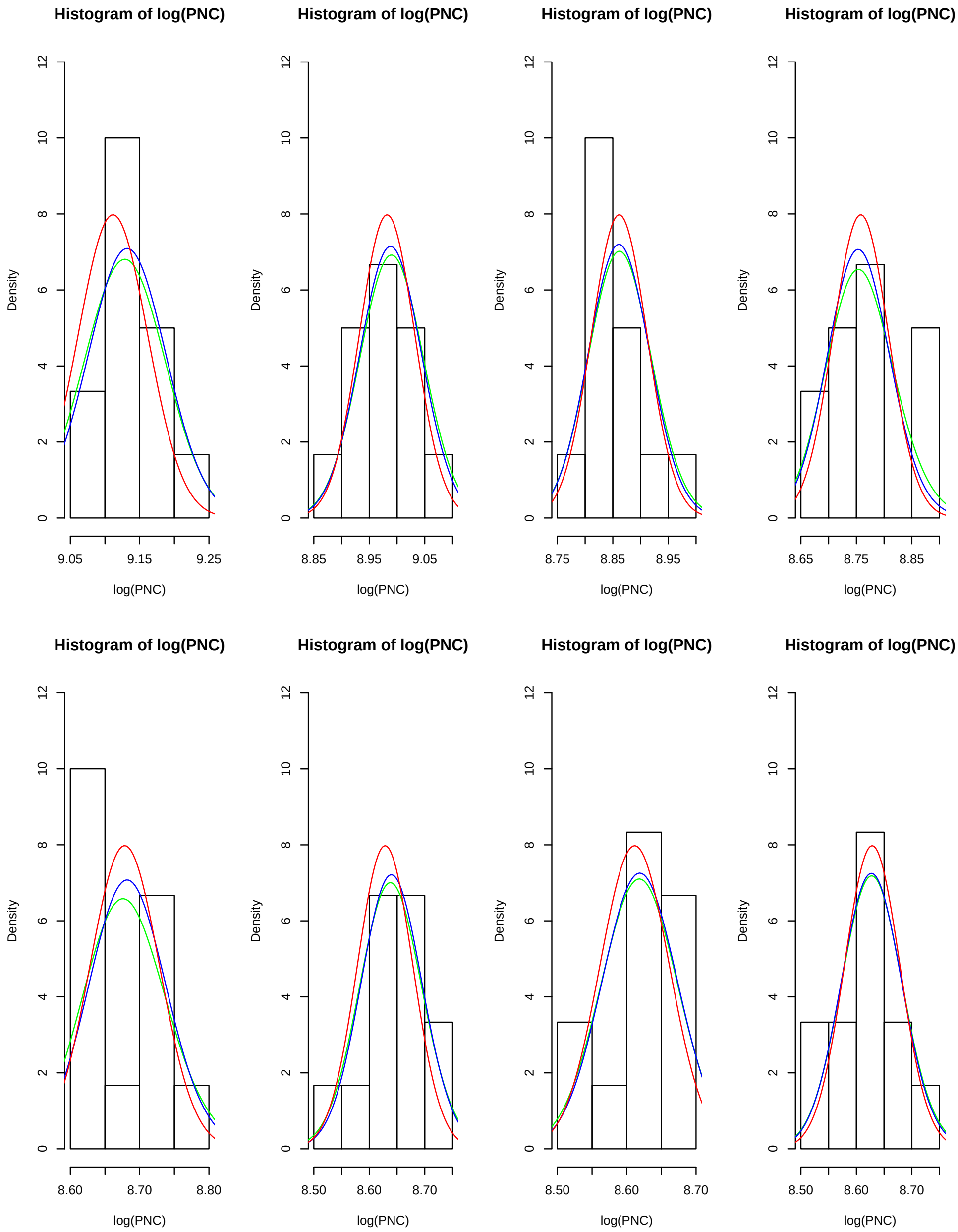

Figure C.10 – 8 randomly chosen hourly densities for simulations described in Section 5.4.5. The red line represents the true density (**1-component Gaussian distribution**) and the green line represents the posterior density from **incorrect 2-component model fitted with classical data** while the blue line represents the posterior density from **incorrect 2-component model fitted with 5-quantile histogram**.

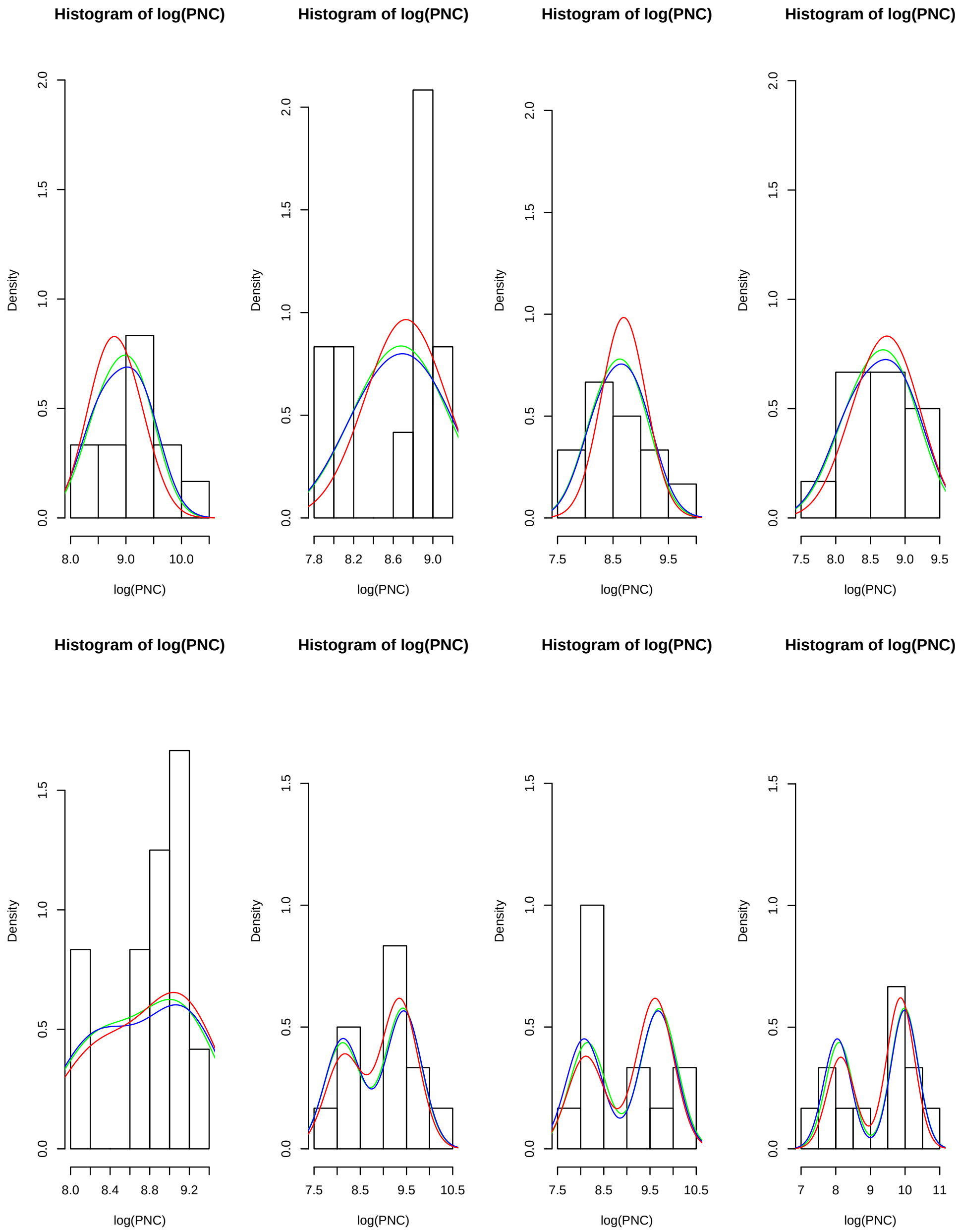

Figure C.11 – 8 randomly chosen hourly densities for simulations described in Section 5.4.5. The red line represents the true density (**2-component Gaussian distribution**) and the green line represents the posterior density from **correct 2-component model fitted with classical data** while the blue line represents the posterior density from **correct 2-component model fitted with 5-quantile histogram**.

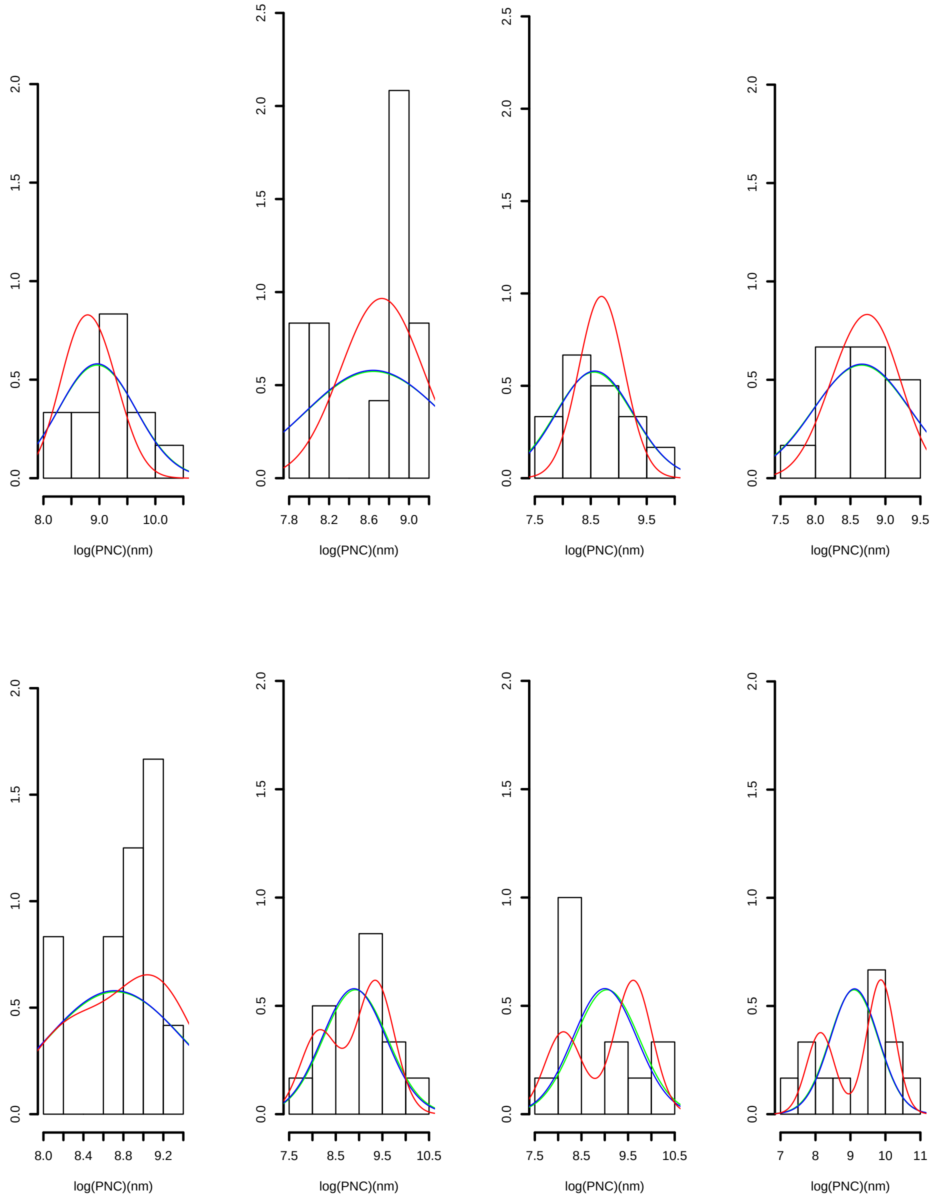

Figure C.12 – 8 randomly chosen hourly densities for simulations described in Section 5.4.5. The red line represents the true density (**2-component Gaussian distribution**) and the green line represents the posterior density from **incorrect 1-component model fitted with classical data** while the blue line represents the posterior density from **incorrect 1-component model fitted with 5-quantile histogram**.

List of Figures

List of Tables

Bibliography

Ahn, J., Peng, M., Park, C., and Jeon, Y. (2012), "A resampling approach for interval-valued data regression," *Statistical Analysis and Data Mining*, 5, 336–348.

Ai, H., Yongmiao, H., Lai, K. K., and Shouyang (2008), "Interval time series analysis with an application to the sterling-dollar exchange rate," *Journal of Systems Science and Complexity*, 21, 558–573.

Alston, C. L., Mengersen, K. L., Robert, C. P., Thompson, J., Littlefield, P., Perry, D., and Ball, A. (2007), "Bayesian mixture models in a longitudinal setting for analysing sheep CAT scan images," *Computational statistics & data analysis*, 51, 4282–4296.

Andrieu, C. and Roberts, G. O. (2009), "The pseudo-marginal approach for efficient Monte Carlo computations," *Annals of Statistics*, 37, 697–725.

Appeltans, W., Ahyong, S. T., Anderson, G., Angel, M. V., Artois, T., Bailly, N., Bamber, R., Barber, A., Bartsch, I., Berta, A., et al. (2012), "The magnitude of global marine species diversity," *Current Biology*, 22, 2189–2202.

Arroyo, J., González-Rivera, G., Maté, C., and Roque, A. M. S. (2011), "Smoothing methods for histogram-valued time series: an application to value-at-risk," *Statistical Analysis and Data Mining*, 4, 216–228.

Arroyo, J. and Maté, C. (2009), "Forecasting histogram time series with k-nearest neighbours methods," *International Journal of Forecasting*, 25, 192–207.

Azzalini, A. (2014), *The skew-normal and related families*, vol. 3 of *Institute of Mathematical Statistics (IMS) Monographs*, Cambridge University Press, Cambridge, with the collaboration of Antonella Capitanio.

Bardenet, R., Doucet, A., and Holmes, C. (2014), "Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach," *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 405–413.

Basset, Y., Samuelson, G., Allison, A., and Miller, S. (1996), "How many species of host-specific insects feed on a species of tropical tree?" *Biological Journal of the Linnean Society*, 59, 201–216.

Benayas, J. M. R., Newton, A. C., Diaz, A., and Bullock, J. M. (2009), "Enhancement of biodiversity and ecosystem services by ecological restoration: a meta-analysis," *Science*, 325, 1121–1124.

Beranger, B., Lin, H., and Sisson, S. A. (2018), "New models for symbolic data analysis," *arXiv preprint arXiv:1809.03659*.

Bertrand, P. and Goupil, F. (2000), "Descriptive statistics for symbolic data," in *Analysis of symbolic data*, Springer, pp. 106–124.

Betancourt, M. (2017), "Identifying Bayesian Mixture Models," [Online; posted February-2017].

Billard, L. (2007), *Dependencies and Variation Components of Symbolic Interval-Valued Data*, Springer Berlin Heidelberg, chap. Selected Contributions in Data Analysis and Classification Part of the series Studies in Classification, Data Analysis, and Knowledge Organization, pp. 3–12.

— (2008), "Sample covariance functions for complex quantitative data," in *Proceedings of World IASC Conference, Yokohama, Japan*, pp. 157–163.

— (2011), "Brief overview of symbolic data and analytic issues," *Statistical Analysis and Data Mining*, 4, 149–156.

Billard, L. and Diday, E. (2000), "Regression analysis for interval-valued data," in *Data Analysis, Classification, and Related Methods*, Springer, pp. 369–374.

— (2003a), "From the statistics of data to the statistics of knowledge: symbolic data analysis," *Journal of the American Statistical Association*, 98, 470–487.

— (2003b), "Symbolic data analysis: Definitions and examples," *Technical Report 62 pages*.

— (2006), *Symbolic data analysis*, Wiley Series in Computational Statistics, John Wiley & Sons, Ltd., Chichester.

Bland, M. (2015), "Estimating mean and standard deviation from the sample size, three quartiles, minimum and maximum," *International Journal of Statistics in Medical Research*, 4, 57–64.

Bock, H.-H. (2008), "Probabilistic modeling for symbolic data," in *COMPSTAT 2008*, Springer, pp. 55–65.

Bock, H.-H. and Diday, E. (eds.) (2000), *Analysis of symbolic data*, Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, Berlin, exploratory methods for extracting statistical information from complex data.

Bouchet, P. and Duarte, C. M. (2006), "The exploration of marine biodiversity: scientific and technological challenges," *Fundación BBVA*, 33.

Brito, P. (1994), "Use of pyramids in symbolic data analysis," in *New Approaches in Classification and Data Analysis*, Springer, pp. 378–386.

— (1995), “Symbolic objects: order structure and pyramidal clustering,” *Annals of Operations Research*, 55, 277–297.

— (2002), “Hierarchical and pyramidal clustering for symbolic data,” *Journal of the Japanese Society of Computational Statistics*, 15, 231–244.

Brito, P. and Chavent, M. (2012), “Divisive monothetic clustering for interval and histogram-valued data,” in *ICPRAM 2012-1st International Conference on Pattern Recognition Applications and Methods*, pp. 229–234.

Brito, P., De Carvalho, F. d. A., Diday, E., and Noirhomme-Fraiture, M. (2008), “Hierarchical and pyramidal clustering,” *Symbolic Data Analysis and the Sodas Software*, 181–203.

Brito, P. and Duarte Silva, A. P. (2012), “Modelling interval data with normal and skew-normal distributions,” *Journal of Applied Statistics*, 39, 3–20.

Brito, P. and Polaillon, G. (2012), “Classification Conceptuelle avec Généralisation par Intervalles.” in *EGC*, pp. 113–118.

Brito, P., Silva, A. P. D., and Dias, J. G. (2015), “Probabilistic clustering of interval data.” *Intell. Data Anal.*, 19, 293–313.

Caley, M. J., Fisher, R., and Mengersen, K. (2014), “Global species richness estimates have not converged,” *Trends in Ecology & Evolution*, 29, 187–188.

Cariou, V. and Billard, L. (2015), “Generalization method when manipulating relational databases,” in *Symbolic Data Analysis and Visualisation: Special Issue of Revue des Nouvelles Technologies de l’Information in honour of Monique Noirhomme-Fraiture*, eds. Brito, P. and Venturini, G., vol. RNTI-E-29, pp. 59–88.

Caron, F., Davy, M., and Doucet, A. (2012), “Generalized Polya urn for time-varying Dirichlet process mixtures,” *arXiv preprint arXiv:1206.5254*.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2015), “Stan: A probabilistic programming language,” *Journal of Statistical Software*, in press.

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., and Palmer, T. M. (2015), “Accelerated modern human–induced species losses: Entering the sixth mass extinction,” *Science Advances*, 1, e1400253.

Chavent, M. and Lechevallier, Y. (2002), “Dynamical clustering of interval data: optimization of an adequacy criterion based on Hausdorff distance,” in *Classification, clustering, and data analysis*, Springer, pp. 53–60.

Cheung, H., Chou, C.-K., Huang, W.-R., and Tsai, C.-Y. (2013), “Characterization of ultrafine particle number concentration and new particle formation in an urban environment of Taipei, Taiwan,” *Atmospheric Chemistry and Physics*, 13, 8935–8946.

Clifford, S., Choy, S. L., Mazaheri, M., Salimi, F., Morawska, L., and Mengsersen, K. (2012a), "A Bayesian spatio-temporal model of panel design data: airborne particle number concentration in Brisbane, Australia," *arXiv preprint arXiv:1206.3833*.

Clifford, S., Mazaheri, M., Salimi, F., Ezz, W. N., Yeganeh, B., Low-Choy, S., Walker, K., Mengersen, K., Marks, G. B., and Morawska, L. (2018), "Effects of exposure to ambient ultrafine particles on respiratory health and systemic inflammation in children," *Environment international*, 114, 167–180.

Clifford, S., Mølgaard, B., Choy, S. L., Corander, J., Hämeri, K., Mengersen, K., and Hussein, T. (2012b), "Bayesian semi-parametric forecasting of ultrafine particle number concentration with penalised splines and autoregressive errors," *arXiv preprint arXiv:1207.0558*.

Costello, M. J., May, R. M., and Stork, N. E. (2013), "Can we name Earth's species before they go extinct?" *Science*, 339, 413–416.

Costello, M. J., Wilson, S., and Houlding, B. (2011), "Predicting total global species richness using rates of species description and estimates of taxonomic effort," *Systematic Biology*, syr080.

Cracraft, J. and Grifo, F. T. (1999), *The living planet in crisis: biodiversity science and policy*, Columbia University Press.

Cuvelier, E. et al. (2009), "QAMML: Probability Distributions for Functional Data," Ph.D. thesis, Ph. D. Thesis, University of Namur, Belgium.

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C., and De Boor, C. (1978), *A practical guide to splines*, vol. 27, Springer-Verlag New York.

de Carvalho, F. d. A., Brito, P., and Bock, H.-H. (2006), "Dynamic clustering for interval data based on L 2 distance," *Computational Statistics*, 21, 231–250.

De Carvalho, F. d. A. and Tenório, C. P. (2010), "Fuzzy K-means clustering algorithms for interval-valued data based on adaptive quadratic distances," *Fuzzy Sets and Systems*, 161, 2978–2999.

de Souza, R. M. and De Carvalho, F. d. A. (2004), "Clustering of interval data based on city–block distances," *Pattern Recognition Letters*, 25, 353–365.

Dias, S. and Brito, P. (2015), "Linear regression model with histogram-valued variables," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 8, 75–113.

— (2016), "Off the beaten track: A new linear model for interval data," *European Journal of Operational Research*.

Diday, E. (1987), "Introduction à l'approche symbolique en analyse des données," *Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle*, 23, 193–236.

— (2016), "Thinking by classes in data science: the symbolic data analysis paradigm," *Wiley Interdisciplinary Reviews: Computational Statistics*, 8, 172–205.

Diday, E. and Bock, H. H. (2000), "Analysis of symbolic data: Exploratory methods for extracting statistical information from complex data," .

Diday, E. and Noirhomme-Fraiture, M. (2008), *Symbolic data analysis and the SODAS software*, John Wiley & Sons.

Diday, E. and Vrac, M. (2005), "Mixture decomposition of distributions by copulas in the symbolic data analysis framework," *Discrete Applied Mathematics*, 147, 27–41.

Dunson, D. B. (2006), "Bayesian dynamic modeling of latent trait distributions," *Biostatistics*, 7, 551–568.

Ehrlich, P. R. and Wilson, E. O. (1991), "Biodiversity studies: science and policy," *Science*, 253, 758.

Eilers, P. H. and Marx, B. D. (1996), "Flexible smoothing with B-splines and penalties," *Statistical science*, 89–102.

Erwin, T. L. (1982), "Tropical forests: their richness in Coleoptera and other arthropod species." *Coleopterists Bulletin*, 36, 74–75.

Ezz, W. N., Mazaheri, M., Robinson, P., Johnson, G. R., Clifford, S., He, C., Morawska, L., and Marks, G. B. (2015), "Ultrafine particles from traffic emissions and children' health (Uptech) in Brisbane, Queensland (Australia): Study design and implementation," *International journal of environmental research and public health*, 12, 1687–1702.

Fernández, C. and Green, P. J. (2002), "Modelling spatially correlated data via mixtures: a Bayesian approach," *Journal of the royal statistical society: series B (Statistical methodology)*, 64, 805–826.

Fisher, R., O'Leary, R. A., Low-Choy, S., Mengersen, K., Knowlton, N., Brainard, R. E., and Caley, M. J. (2015), "Species richness on coral reefs and the pursuit of convergent global estimates," *Current Biology*, 25, 500–505.

Gallardo, J. A. and Jiménez, C. M. (2008), "Métodos de predicción para series temporales de intervalos e histogramas," *Departamento de Organización Industrial, ICAI, Universidad Pontificia Comillas*, 41.

García-Ascanio, C. and Maté, C. (2010), "Electric power demand forecasting using interval time series: A comparison between VAR and iMLP," *Energy Policy*, 38, 715–725.

Gaston, K. J. (1991), "The magnitude of global insect species richness," *Conservation biology*, 5, 283–296.

Gelman, A., Carlin, J. B., and Stern, H. S. (2003), *Bayesian Data Analysis*, Chapman & Hall/CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis*, Chapman and Hall/CRC Press, Boca Raton, 3rd ed.

Gelman, A., Jakulin, A., Pittau, M. G., Su, Y.-S., et al. (2008), "A weakly informative default prior distribution for logistic and other regression models," *The Annals of Applied Statistics*, 2, 1360–1383.

Gelman, A. et al. (2006), "Prior distributions for variance parameters in hierarchical models (Comment on a paper by Browne and Draper)," *Bayesian Analysis*, 1, 515–534.

Gioia, F. and Lauro, C. N. (2005), "Basic statistical methods for interval data," *Statistica Applicata*, 17, 1–27, 1.

— (2006), "Principal component analysis on interval data," *Computational Statistics*, 21, 343–363.

Giordani, P. (2015), "Lasso-constrained regression analysis for interval-valued data," *Advances in Data Analysis and Classification*, 9, 5–19.

Grassle, J. F. and Maciolek, N. J. (1992), "Deep-sea species richness: Regional and local diversity estimates from quantitative bottom samples," *American Naturalist*, 313–341.

Green, P. J. and Richardson, S. (2002), "Hidden Markov models and disease mapping," *Journal of the American statistical association*, 97, 1055–1070.

Groombridge, B. and Jenkins, M. D. (2002), "World Atlas of Biodiversity." *Prepared by the UNEP World Conservation Monitoring Centre. Unversity of California Press, Berkeley.*

Guha, S., Hafen, R., Rounds, J., Xia, J., Li, J., Xi, B., and Cleveland, W. S. (2012), "Large complex data: Divide and recombine (D&R) with RHIPE," *Stat*, 1, 53–67.

Gustafson, P. and Walker, L. (2003), "An extension of the Dirichlet prior for the analysis of longitudinal multinomial data," *Journal of Applied Statistics*, 30, 293–310.

Hamilton, A. J., Basset, Y., Benke, K. K., Grimbacher, P. S., Miller, S. E., Novotnỳ, V., Samuelson, G. A., Stork, N. E., Weiblen, G. D., and Yen, J. D. (2010), "Quantifying uncertainty in estimation of tropical arthropod species richness," *The American Naturalist*, 176, 90–95.

Hamilton, A. J., Novotnỳ, V., Waters, E. K., Basset, Y., Benke, K. K., Grimbacher, P. S., Miller, S. E., Samuelson, G. A., Weiblen, G. D., Yen, J. D., et al. (2013), "Estimating global arthropod species richness: refining probabilistic models using probability bounds analysis," *Oecologia*, 171, 357–365.

Hammond, P. (1995), "Described and estimated species numbers: an objective assessment of current knowledge," *Microbial diversity and ecosystem function*, 29–71.

Harrison, R. M., Shi, J. P., Xi, S., Khan, A., Mark, D., Kinnersley, R., and Yin, J. (2000), "Measurement of number, mass and size distribution of particles in the atmosphere," *Philosophical transactions of the royal society of London A: mathematical, physical and engineering sciences*, 358, 2567–2580.

Heal, M. R., Kumar, P., and Harrison, R. M. (2012), "Particles, air quality, policy and health," *Chemical Society Reviews*, 41, 6606–6630.

Heitjan, D. F. and Rubin, D. B. (1991), "Ignorability and Coarse Data," *The Annals of Statistics*, 19, 2244–2253.

Hodkinson, I. and Casson, D. (1991), "A lesser predilection for bugs: Hemiptera (Insecta) diversity in tropical rain forests," *Biological Journal of the Linnean Society*, 43, 101–109.

Hozo, S. P., Djulbegovic, B., and Hozo, I. (2005), "Estimating the mean and variance from the median, range and the size of a sample," *BMC Medical Research Methodology*, 5, 13.

Hron, K., Brito, P., and Filzmoser, P. (2017), "Exploratory data analysis for interval compositional data," *Advances in Data Analysis and Classification*, 11, 223–241.

Hunter, P. (2007), "The human impact on biological diversity," *EMBO reports*, 8, 316–318.

Hussein, T., Hämeri, K., Aalto, P. P., Paatero, P., and Kulmala, M. (2005), "Modal structure and spatial–temporal variations of urban and suburban aerosols in Helsinki Finland," *Atmospheric Environment*, 39, 1655–1668.

Ichino, M. (2011), "The quantile method for symbolic principal component analysis," *Statistical Analysis and Data Mining*, 4, 184–198.

Ioannidis, Y. (2003), "- The History of Histograms (abridged)," in *Proceedings 2003 VLDB Conference*, eds. Freytag, J.-C., Lockemann, P., Abiteboul, S., Carey, M., Selinger, P., and Heuer, A., San Francisco: Morgan Kaufmann, pp. 19 – 30.

Irpino, A. (2013), "Basic univariate and bivariate statistics for symbolic data: a critical review," *arXiv preprint arXiv:1312.2248*.

Irpino, A. and Verde, R. (2006), *A New Wasserstein Based Distance for the Hierarchical Clustering of Histogram Symbolic Data*, Springer Berlin Heidelberg, pp. 185–192.

— (2015), "Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein Distance," *Advances in Data Analysis and Classification*, 9, 81–106.

Ji, C. (2009), "Advances in bayesian modelling and computation: spatio-temporal processes, model assessment and adaptive mcmc," .

Jordan, M. I., Lee, J. D., and Yang, Y. (2018), "Communication-efficient distributed statistical inference," *Journal of the American Statistical Association*, in press.

Knowlton, N., Brainard, R. E., Fisher, R., Moews, M., Plaisance, L., and Caley, M. J. (2010), "Coral reef biodiversity," *Life in the Worlds Oceans: Diversity Distribution and Abundance*, 65–74.

Kosmelj, K., Le-Rademacher, J., and Billard, L. (2014), "Symbolic Covariance Matrix for Interval-valued Variables and its Application to Principal Component Analysis: a Case Study," *Metodoloski Zvezki*, 11, 1–20, copyright - Copyright Anuska Ferligoj 2014; Document feature - ; Tables; Graphs; Last updated - 2015-11-24.

Kumar, P., Gurjar, B., Nagpure, A., and Harrison, R. M. (2011), "Preliminary estimates of nanoparticle number emissions from road vehicles in megacity Delhi and associated health impacts," *Environmental Science & Technology*, 45, 5514–5521.

Kumar, P., Robins, A., Vardoulakis, S., and Britter, R. (2010), "A review of the characteristics of nanoparticles in the urban atmosphere and the prospects for developing regulatory controls," *Atmospheric Environment*, 44, 5035–5052.

Lambshead, P. (1993), "Recent developments in marine benthic biodiversity research," *Oceanis*, 19, 5–5.

Lang, S. and Brezger, A. (2004), "Bayesian P-splines," *Journal of computational and graphical statistics*, 13, 183–212.

Lauro, C., Verde, R., and Irpino, A. (2008), "Generalized canonical analysis," *Symbolic Data Analysis and the Sodas Software*, 313–330.

Le-Rademacher, J. and Billard, L. (2011), "Likelihood functions and some maximum likelihood estimators for symbolic data," *Journal of Statistical Planning and Inference*, 141, 1593–1602.

— (2012), "Symbolic covariance principal component analysis and visualization for interval-valued data," *Journal of Computational and Graphical Statistics*, 21, 413–432.

— (2013), "Principal component analysis for histogram-valued data," *Advances in Data Analysis and Classification*, 1–25.

Lima Neto, E. d. A. and dos Anjos, U. U. (2015), "Regression model for interval-valued variables based on copulas," *Journal of Applied Statistics*, 42, 2010–2029.

Lin, H., Caley, M. J., and Sisson, S. A. (2017), "Estimating global species richness using symbolic data meta-analysis," *arXiv:1711.03202*.

Lin, W. and González-Rivera, G. (2016), "Interval-valued time series models: Estimation based on order statistics exploring the Agriculture Marketing Service data," *Computational Statistics & Data Analysis*, 100, 694–711.

Lin, X. and Zhang, D. (1999), "Inference in generalized additive mixed modelsby using smoothing splines," *Journal of the royal statistical society: Series b (statistical methodology)*, 61, 381–400.

Luo, D., Wan, X., Liu, J., and Tong, T. (2018), "Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range," *Statistical Methods in Medical Research*, 27, 1785–1805.

MacEachern, S. N. (2001), "Decision theoretic aspects of dependent nonparametric processes," *Bayesian methods with applications to science, policy and official statistics*, 551–560.

Maia, A. L. S. and de Carvalho, F. d. A. (2008), "Fitting a least absolute deviation regression model on interval-valued data," in *Advances in Artificial Intelligence-SBIA 2008*, Springer, pp. 207–216.

Maia, A. L. S., de Carvalho, F. d. A., and Ludermir, T. B. (2008), "Forecasting models for interval-valued time series," *Neurocomputing*, 71, 3344–3352.

Marin, J.-M., Mengersen, K., and Robert, C. P. (2005), "Bayesian modelling and inference on mixtures of distributions," *Handbook of statistics*, 25, 459–507.

Marx, B. D. and Eilers, P. H. (2005), "Multidimensional penalized signal regression," *Technometrics*, 47, 13–22.

May, R. M. (1992a), "Bottoms up for the oceans," *Nature*, 357, 278–279.

— (1992b), "How many species inhabit the earth," *Scientific American*, 267, 42–48.

May, R. M. and Beverton, R. (1990), "How many species?[and discussion]," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 330, 293–304.

McLachlan, G. and Jones, P. (1988), "Fitting mixture models to grouped and truncated data via the EM algorithm," *Biometrics*, 571–578.

Mejia, J., Wraith, D., Mengersen, K., and Morawska, L. (2007), "Trends in size classified particle number concentration in subtropical Brisbane, Australia, based on a 5 year study," *Atmospheric Environment*, 41, 1064–1079.

Montanari, A., Mignani, S., Monari, P., and Vichi, M. (2005), *New Developments in Classification and Data Analysis: Proceedings*, Springer.

Mooney, H., Larigauderie, A., Cesario, M., Elmquist, T., Hoegh-Guldberg, O., Lavorel, S., Mace, G. M., Palmer, M., Scholes, R., and Yahara, T. (2009), "Biodiversity, climate change, and ecosystem services," *Current Opinion in Environmental Sustainability*, 1, 46–54.

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G., and Worm, B. (2011), "How many species are there on Earth and in the ocean?" *PLoS Biol*, 9, e1001127.

Morawska, L., Jayaratne, E., Mengersen, K., Jamriska, M., and Thomas, S. (2002), "Differences in airborne particle and gaseous concentrations in urban air between weekdays and weekends," *Atmospheric Environment*, 36, 4375–4383.

Mousavi, H. and Zaniolo, C. (2011), “Fast and Accurate Computation of Equi-depth Histograms over Data Streams,” in *Proceedings of the 14th International Conference on Extending Database Technology*, New York, NY, USA: ACM, EDBT/ICDT ’11, pp. 69–80.

Neto, E. A. L., Corderio, G. M., and de Carvalho, F. A. T. (2011), “Bivarite symbolic regression models for interval-valued variables,” *Journal of Statistical Computation and Simulation*, 81, 1727–1744.

Neto, E. d. A. L., Cordeiro, G. M., de Carvalho, F. A., dos Anjos, U. U., and da Costa, A. G. (2009), “Bivariate generalized linear model for interval-valued variables,” in *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, IEEE, pp. 2226–2229.

Neto, E. d. A. L., de Carvalho, F. A., and Tenorio, C. P. (2004), “Univariate and multivariate linear regression methods to predict interval-valued features,” in *Australasian Joint Conference on Artificial Intelligence*, Springer, pp. 526–537.

Neto, E. d. A. L. and de Carvalho, F. d. A. (2010), “Constrained linear regression models for symbolic interval-valued variables,” *Computational Statistics & Data Analysis*, 54, 333–347.

Nielsen, E. S. and Mound, L. A. (2000), “Global diversity of insects: the problems of estimating numbers,” *Nature and human society: The quest for a sustainable world*, 213–222.

Noirhomme-Fraiture, M. and Brito, P. (2011), “Far beyond the classical data models: symbolic data analysis,” *Statistical Analysis and Data Mining*, 4, 157–170.

Novotny, V., Basset, Y., Miller, S. E., Weiblen, G. D., Bremer, B., Cizek, L., and Drozd, P. (2002), “Low host specificity of herbivorous insects in a tropical forest,” *Nature*, 416, 841–844.

ØDegaard, F. (2000), “How many species of arthropods? Erwin’s estimate revised,” *Biological Journal of the Linnean Society*, 71, 583–597.

Pimm, S. L., Jenkins, C. N., Abell, R., Brooks, T. M., Gittleman, J. L., Joppa, L. N., Raven, P. H., Roberts, C. M., and Sexton, J. O. (2014), “The biodiversity of species and their rates of extinction, distribution, and protection,” *Science*, 344, 1246752.

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2018a), “Speeding up MCMC by eficient data subsampling,” *Journal of the American Statistical Association*, in press.

Quiroz, M., Tran, M.-N., Villani, M., and Kohn, R. (2018b), “Speeding up MCMC by delayed acceptance and data subsampling,” *Journal of Computational and Graphical Statistics*, 27, 12–22.

Raven, P. H. (1983), “The challenge of tropical biology,” *Bulletin of the Entomological Society of America*, 29, 4–13.

Raven, P. H. and Yeates, D. K. (2007), "Australian biodiversity: threats for the present, opportunities for the future," *Australian Journal of Entomology*, 46, 177–187.

Raven, P. H. et al. (2000), *Nature and human society: the quest for a sustainable world*, National Academies.

Reaka-Kudla (2005), *Biodiversity of Caribbean coral reefs. In: Caribbean Marine Biodiversity: The Known and the Unknown.*, DEStech Publications.

Reaka-Kudla, M. L., Wilson, D. E., and Wilson, E. O. (1996), *Biodiversity II: understanding and protecting our biological resources*, Joseph Henry Press.

Rendell, L. J., Johansen, A. M., Lee, A., and Whiteley, N. (2018), "Global consensus Monte Carlo," *arXiv preprint arXiv:1807.09288*.

Rodrigues, G. S., Nott, D. J., and Sisson, S. A. (2016), "Functional regression approximate Bayesian computation for Gaussian process density estimation," *Computational Statistics & Data Analysis*, 103, 229–241.

Rodrıguez, O., Diday, E., and Winsberg, S. (2000), "Generalization of the principal components analysis to histogram data," in *Workshop on simbolic data analysis of the 4th European Conference on principles and practice of knowledge discovery in data bases, Setiembre*, pp. 12–16.

Rodrıguez, O. and Pacheco, A. (2004), "Applications of histogram principal components analysis," *Symbolic and Spatial Data Analysis: Mining Complex Data Structures*, 101.

Rubin, D. B. (1981), "Estimation in parallel randomised experiments," *Journal of Educational Statistics*, 6, 377–401.

Rue, H. and Held, L. (2005), *Gaussian Markov random fields: theory and applications*, CRC press.

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the royal statistical society: Series b (statistical methodology)*, 71, 319–392.

Sabrosky, C. W. (1953), "How many insects are there?" *Systematic Zoology*, 2, 31–36.

Schweizer, B. (1984), "Distributions are the numbers of the future," in *Proc. Math. Fuzzy Syst. Meeting*, pp. 137–149.

Shah, A. P., Pietropaoli, A. P., Frasier, L. M., Speers, D. M., Chalupa, D. C., Delehanty, J. M., Huang, L.-S., Utell, M. J., and Frampton, M. W. (2008), "Effect of inhaled carbon ultrafine particles on reactive hyperemia in healthy human subjects," *Environmental health perspectives*, 116, 375.

Shi, J., Luo, D., Weng, H., Zeng, X.-T., Lin, L., and Tong, T. (2018), "How to estimate the sample mean and standard deviation from the five number summary?" *arXiv:1801.01267*.

Silva, A. P. D. and Brito, P. (2015), "Discriminant analysis of interval data: An assessment of parametric and distance-based approaches," *Journal of Classification*, 32, 516–541.

Silva, A. P. D., Filzmoser, P., and Brito, P. (2017), "Outlier detection in interval data," *Advances in Data Analysis and Classification*, 1–38.

Silverman, B. W. (1985), "Some aspects of the spline smoothing approach to non-parametric regression curve fitting," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–52.

Sisson, S. A., Fan, Y., and Beaumont, M. A. (eds.) (2018), *Handbook of Approximate Bayesian Computation*, Chapman & Hall/CRC Press.

Small, A. M., Adey, W. H., and Spoon, D. (1998), "Are current estimates of coral reef biodiversity too low? The view through the window of a microcosm," *Atoll Research Bulletin*, 450, 20.

Stork, N. and Gaston, K. (1990), "Counting species one by one," *New Scientist*, 127, 43–47.

Stork, N. E. (1988), "Insect diversity: facts, fiction and speculation," *Biological Journal of the Linnean Society*, 35, 321–337.

Stork, N. E., McBroom, J., Gely, C., and Hamilton, A. J. (2015), "New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods," *Proceedings of the National Academy of Sciences*, 112, 7519–7523.

Strickland, C. M., Simpson, D. P., Turner, I. W., Denham, R., and Mengersen, K. L. (2008), "Fast Bayesian analysis of spatial dynamic factor models," .

Team, S. (2016), "RStan: the R interface to Stan," *R package version*, 2.

Teles, P. and Brito, P. (2005), "Modelling interval time series data," in *Proceedings of the 3rd IASC World Conference on Computational Statistics and Data Analysis, Limassol, Cyprus.*

Vardeman, S. B. and Lee, C.-S. (2005), "Likelihood-based statistical estimation from quantised data," *IEEE Transactions on Instrumentation and Measurement*, 54, 409–414.

Vehtari, A. and Gelman, A. (2014), "WAIC and cross-validation in Stan," *Submitted. http://www. stat. columbia. edu/˜ gelman/research/unpublished/waic_ stan. pdf Accessed*, 27, 5.

Vono, M., Dobigeon, N., and Chainais, P. (2018), "Split-and-augmented Gibbs sampler – application to large-scale inference problems," *arXiv:1804.05809.*

Wahba, G. et al. (1990), "Spline models for observational data. Society for Industrial and Applied Mathematics," .

Wan, X., Wang, W., Liu, J., and Tong, T. (2014), "Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range," *BMC Medical Research Methodology*, 14, 135.

Wang, X., Zhang, Z., and Li, S. (2016), "Set-valued and interval-valued stationary time series," *Journal of Multivariate Analysis*, 145, 208–223.

Whitby, E. R. and McMurry, P. H. (1997), "Modal aerosol dynamics modeling," *Aerosol Science and Technology*, 27, 673–688.

Wickham, H. et al. (2014), "Tidy data," *Journal of Statistical Software*, 59, 1–23.

Wraith, D., Alston, C., Mengersen, K., and Hussein, T. (2011), "Bayesian mixture model estimation of aerosol particle size distributions," *Environmetrics*, 22, 23–34.

Wraith, D., Mengersen, K., Alston, C., Rousseau, J., and Hussein, T. (2014), "Using informative priors in the estimation of mixtures over time with application to aerosol particle size distributions," *Annals of Applied Statistics*, 8, 232–258.

Xu, W. (2010), "Symbolic Data Analysis: Interval-Valued Data Regression," Ph.D. thesis, University of Georgia.

Zhang, X. and Sisson, S. A. (2016), "Constructing Likelihood Functions for Interval-valued Random Variables," *arXiv:1608.00945*.