Innovative methods for the analysis of complex and non-standard data

Tom Whitaker

Supervised by: Prof. Scott Sisson
Co-supervised by: Dr. Boris Beranger

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

UNSW
AUSTRALIA

School of Mathematics and Statistics
Faculty of Science
November 2019

UNSW SYDNEY | Australia's Global University

Thesis/Dissertation Sheet

Surname/Family Name	:	**Whitaker**
Given Name/s	:	**Thomas Grieve**
Abbreviation for degree as give in the University calendar	:	**PhD**
Faculty	:	**Faculty of Science**
School	:	**School of Mathematics and Statistics**
Thesis Title	:	**Innovative methods for the analysis of complex and non-standard data**

Abstract 350 words maximum:

Symbolic Data Analysis (SDA) is an emerging branch of statistics that addresses some of the issues associated with the analysis of non-standard (symbolic) datasets, such as intervals, histograms and lists. Datasets of this nature are useful in preserving the privacy of individual observations, and also for reducing the size and dimension of big datasets. This leads to significant computational benefits if an appropriate symbolic analysis can be derived. The rapidly increasing computational power that is becoming more and more readily available has also led to increasingly common non-standard datasets. Data arriving in a non-standard form often possesses internal variation not seen in pointwise classical observations. This means that existing classical methods of analysis are unsuitable if results are desired that possess an underlying classical interpretation.

Currently, most developed SDA methods focus on an exploratory analysis of the data, with the subsequent results only useful at the symbolic level, and not directly comparable to the complete analysis of the true latent underlying dataset unless some specific assumptions concerning the uniformity of the data within each symbol are met. A common existing symbolic methodology is to perform a classical analysis of features of the non-standard data, such as interval end-points. In this thesis methods of analysis for non-standard data are developed that are interpretable at the underlying classical level. Further, if enough information is retained during the aggregation process, the methods derived for the analyses of non-standard datasets obtain comparable results to the complete classical analysis of the underlying latent dataset. As a result, big datasets that pose computational problems can be analysed using the proposed symbolic methodologies instead of the classical analyses, at a cheaper computational cost. These methods are highly flexible, meaning they don't rely on a uniformity assumption within each symbol, and can be applied to a range of symbolic data. The utility of each symbolic method is demonstrated via simulation studies illustrating the convergence of the results towards the complete analysis with increasing information retention during the aggregation process. Further, each derived method has then been applied to a real dataset in order to demonstrate their real-life application.

COPYRIGHT STATEMENT

Signed …….....

Date …….....

AUTHENTICITY STATEMENT

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 candidate contributed greater than 50% of the content in the publication and is the "primary author", ie. the candidate was responsible primarily for the planning, execution and preparation of the work for publication
- The candidate 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 UNSW 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.

Candidate's Name	**Signature**	**Date (dd/mm/yy)**

POSTGRADUATE COORDINATOR'S DECLARATION *To only 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 UNSW Thesis Examination Procedure

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

Firstly, I would like to express my gratitude to my primary supervisor Professor Scott Sisson for his continuous support and guidance throughout my PhD journey. I am forever grateful to Scott for taking me on as an honours student 5 years ago, and have benefited immensely from his patience and guidance. I've come to really enjoy our weekly meetings, as they are always upbeat and filled with humour, and sometimes hardly seem like work. Under his guidance, I feel I have aquired a strong foundation in symbolic data analysis, and feel well prepared to communicate these as a statistician.

I am truly grateful to Dr Boris Beranger for acting as my secondary supervisor across the course of my PhD. Countless times Boris has been available for me to come and annoy him with questions, and has never hesitated to provide guidance and support. In particular, I really appreciate the hard work Boris put in towards the end of my PhD, working extremely hard to help me prepare each manuscript and the final submission. 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.

Special thanks to the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) for providing opportunities for collaboration and travel. I am also grateful to Dr Markus Donat (Climate Change Research Centre, UNSW Sydney) for providing the climate dataset used in Chaper 3. I would like to thank Dr Ben Hess and Professor Kerrie Mengersen for providing the Landsat satellite data for Chapter 4, and also for offering advice for the initial analysis. I also owe thanks to Dr Eugenie Hunsicker for providing the Athena SWAN dataset that motivated the methods developed in Chapter 6, and for providing valuable insights into the data in a productive collaboration.

In June 2017, I was fortunate to attend the sixth symbolic data analysis workshop held in Ljubljana, Slovenia, where I was fortunate enough to meet some of the leading senior researchers from other countries in the field of Symbolic Data Analysis. I am so grateful for UNSW for funding this travel, and for the SDA community for making it a highly memorable trip.

Special thanks to my fellow PhD students with whom I share the same office. Thank you for the frequent friendly chats and the overall friendly environment of the office. In particular, thank you to my (former) fellow PhD colleagues who were also under the supervision of Scott; Jaslene, Vincent, Guilherme and Xin, who provided valuable support and were always eager to share their own experiences. To all the academic staff I encountered and worked with during my time at UNSW, thank you for your assistance and for providing me with a wonderful opportunity to

work as a statistics tutor for the school. I feel very strongly that this experience of teaching was highly enjoyable, and provided a nice break from PhD life, while also helping me significantly develop my communication skills.

Finally, I cannot thank my family and friends enough. My mother and father have been more supportive than I thought was possible, providing countless emotional (and financial) support. Words can't explain the significance of your constant availability to always talk whenever I felt overly stressed. To my partner Bri, thank you for putting up with me during this period of my life. I don't know what I would have done without your endless support. Finally, to Matt and Simon, and my other friends, these sorts of endeavours would not be possible without a solid group of mates to fall back on. Although my PhD journey is coming to an end, a brand-new chapter in my life is beginning. Thank you immensely to all for joining me on this journey.

Contents

Chapter 1

Introduction

Classical statistical techniques involve the analysis of realisations $\mathbf{x}$ of random variables $\mathbf{X}$ that take the form of single points (possibly multidimensional) within their domain $D_{\mathbf{X}} \subseteq \mathbb{R}^D$. Classical datasets then comprise of N observations, with D measurements each, where D is the number of variables in $\mathbf{X}$. A consequence of the increasing occurence of huge, complex datasets, often referred to as big data (Billard and Diday, 2003), is the large computational power required for a classical analyses. If a grouping mechanism is available for the dataset, then one solution is to summarise the data within known groups by a set of summary statistics, termed 'symbols', and then perform an analysis on the smaller dimensional 'symbolic' dataset. Diday (1989) first proposed the notion of Symbolic Data Analysis (SDA), whereby these summary level observations are analysed instead of the much larger underlying classical dataset, allowing for huge potential savings in computation, storage and transmission if enough information is retained in the symbolic dataset. Examples of commonly recorded symbols are intervals, histograms, distributions and lists (Billard and Diday, 2003). For the above examples, the underlying classical data (also termed 'microdata') is divided into disjoint groups using some known criteria (age, gender, nationality, weight), and then aggregated into a set of symbolic observations, one per group for a total of B groups. The choice of groups is dependent on the problem at hand, with different configurations appropriate for different forms of analysis. For example, in an analysis of medical data, it might make sense to group individual patient records by gender, age and weight, as you would expect similar observations within each group. SDA methods are also useful for situations where the privacy of the individuals is required to be maintained, and so the data is only available in the form of group level summaries, thus protecting the privacy of individual level observations. For example, when performing an analysis on the salaries of all the employees in a company, individual records may not be available, in order to protect the sensitive financial information. Instead, the data for each division of the company might be aggregated into a set of intervals or histograms (one per variable), from which the practitioner needs to extract a meaningful analysis.

When analysing symbolic data in the form of intervals, a simple approach is to perform the classical analysis on the midpoints of the intervals/subintervals. Comparable approaches are available for histogram-valued data, whereby the midpoints of the histogram bins, weighted

by their respective counts, are analysed in place of the underlying microdata. Approaches such as these have been utilised to derive estima tes for the sample variances and covariances of a symbolic dataset (Billard and Diday, 2003, Oliveira et al., 2018). While this approach is computationally efficient and easy to implement, it ignores the internal variation of the microdata within each symbol. An assumption commonly utilised in SDA to address this problem is that of within-symbol uniformity. That is, we assume the distribution of an underlying classical observation, given the symbol (interval, histogram subinterval), is uniform (Billard and Diday, 2006). This assumption has been used to derive expressions for the sample symbolic mean and variance for univariate symbolic datasets by Bertrand and Goupil (2000), Billard (2007), and was then extended to estimating the sample symbolic correlation for multidimensional symbolic datasets by Billard and Diday (2003), Billard (2007). Oliveira et al. (2018) explored how these expressions have an underlying microdata interpretation, and are in fact comparable to their classical equivalents, if the uniform within-symbol distribution is correctly specified.

When the within-symbol uniformity assumption is violated, SDA results are often significantly different to that of a comparable analysis performed on the complete underlying classical dataset (Beranger et al., 2018). When analysis at the group level is all that is required, this difference is meaningless as the group differences are of interest, not the individual differences. However, if an analysis with an underlying classical interpretation is desired, then a parametric assumption is often utilised to remove the within-symbol uniformity assumption. Heitjan (1989) examined the case whereby data arrives in a grouped continuous form (histograms), and a normal distribution is assumed for the underlying microdata. A grouped likelihood function is then constructed using the difference of cumulative distribution functions (cdf's) for each histogram bin, taken to the power of the count for that bin. The results from an analysis of the underlying classical data, the grouped likelihood function and a classical analysis of the midpoints with sheppards correction are then compared, with the latter two performing comparably. Heitjan and Rubin (1991) then investigated the effects of the coarsening mechanism for histogram data, and determined that if the coarsening occurs according to a Coarsening at Random (CaR) process, then the likelihood of the coarsening can be ignored. Le Rademacher and Billard (2011) derive likelihood functions for intervals and histograms with group-level interpretations, whereby parametric functions are assumed for the parameters of interest. Zhang et al. (2019) construct interval symbolic likelihood functions for parametric models, whereby parametric models are fit to the interval-valued observations, with the results comparable to that of the classical analysis performed on the microdata if enough information about the underlying classical data is retained in the aggregation process. Beranger et al. (2018) then proposed a generalised symbolic likelihood function for symbolic datasets for which an underlying parametric model can be assumed for the microdata, and propose examples of new types of symbolic observations for which this method of analysis can be easily applied to. The parametric frameworks for intervals and histograms presented in Heitjan (1989) and Zhang et al. (2019) respectively are then special cases of this construction. The generalised symbolic likelihood is shown to reduce to the classical likelihood function for a certain level of data aggregation, with the subsequent results allowing inferences to be made at both the group and individual levels.

If a parametric form of the underlying microdata is unable to be assumed, then non-parametric techniques for symbolic data can be utilised to obtain inferences with an underlying classical interpretation. For data arriving in the form of a histogram, Scott and Sheather (1985) and Hall (1996) examine the errors associated with a kernel density estimation on the midpoints of each histogram bin, weighted by the bin's count. Analogous methods to the classical case can be utilised to determine the appropriate kernel bandwidth, and subsequent underlying classical densities. Minnotte (1996) and Koo and Kooperberg (2000) fit splines to the observed histogram proportions, with reasonable results if there aren't any non-empty bins (excluding bins on the edge). Histogram density estimates are also constructed using underlying interval/histogram datasets by Billard and Diday (2003), whereby the density at any given point within the domain of the microdata is estimated as the probability that an underlying observation would lie in the subinterval that point falls in (given the symbolic dataset), weighted by the size of the subinterval.

The thesis proceeds as follows. In Chapter 2 we undertake a thorough literature review of the currently developed SDA methodologies. Basic descriptive statistics for common symbolic observations are described, along with existing methodologies for parametric and non-parametric analyses of symbolic data. Particular attention is paid to the methods available within SDA whereby the results of the analysis have an underlying classical interpretation, and are comparable to the results that would have been obtained from a complete analysis of the underlying classical dataset.

Often a likelihood analysis requires the evaluation of large-dimensional, intractable functions. Composite likelihoods (Cox and Reid, 2004, Lindsay, 1982, Varin, 2008) are used as a solution to this issue, whereby large dimensional intractable likelihoods are replaced by the product of smaller-dimensional, unbiased likelihoods of marginal or conditional events. The resultant analysis however typically has a large number of terms (especially if the number of variables D is large), and so for datasets with a large number of observations, a composite likelihood analysis is computationally very intensive. A composite likelihood analysis of summary level symbols, instead of the complete classical dataset, is a potential method of reducing this computational burden. In Chapter 3 we extend the likelihood framework of Beranger et al. (2018) to the composite likelihood setting, whereby large-dimensional datasets are aggregated into sets of lower-dimensional histograms, on which a likelihood analysis is undertaken. The estimators obtained via this symbolic composite likelihood function are shown to be asymptotically consistent with the classical estimator with an increasing amount of information retention in the aggregation process. The derived symbolic composite likelihood function is then utilised in the analysis of a large temperature dataset as an example of its applicability in the analysis of max-stable processes. An interesting result that arises from this analysis is that for classical data that has been aggregated into a set of bivariate histograms, while the parameter estimates obtained from a symbolic composite approach are asymptotically consistent with an increasing number of bivariate bins, their variances require the number of histogram replications to increase in order to converge towards the classical results. Variances for the parameter estimates are then obtained via the Godambe information matrix (Godambe, 1960, White, 1982), allowing

standard parametric inferences to be obtained.

For large dimensional data, the symbolic likelihood framework of Beranger et al. (2018) requires the solution of a large dimensional integral for histogram valued data. For many known distributions, such as the bivariate max-stable models investigated in Chapter 3, these integrals have closed form solutions, and so are computationally simple to evaluate. However, for many parametric models, such as logistic regression, these integrals have no closed form solution for multidimensional data, and thus require numerical methods to evaluate, increasing the complexity of the computation. Due to the curse of dimensionality, the complexity of the integral also grows with the dimension, meaning for datasets with many (more than 2 or 3) predictor variables, a symbolic analysis becomes computationally infeasible. Furthermore, a composite likelihood approach isn't available, as lower dimensional unbiased likelihoods do not exist for the logistic regression model. In Chapter 4 we derive closed form expressions for the approximate composite likelihood of a multi-class logistic regression model, whereby the predictor variables arrive in the form of univariate histograms, one per class. A distributional assumption is utilised for the predictor variables, leading to the reduction of an intractable large dimensional integral (and subsequent likelihood) to a computational simple univariate function. If this distributional assumption is reasonable, predictions obtained from the approximate univariate likelihood are shown to be comparable to that of a classical analysis on the microdata. If the distributional assumption is violated, normalisation procedures are readily available to improve the assumption. A consequence of this is that when a logistic regression analysis is desired, a continuous classical underlying dataset can be stored, transferred and analyse in the form of a set of univariate histograms, leading to massive gains in computation.

In the above analyses, the uniformity-within-intervals assumption problem previously described is fixed using a parametric assumption for the underlying microdata. However, if no parametric assumption is available (due to lack of information, desire for flexibility, etc...) then non-parametric methods can be utilised in the analysis of symbolic data. The methodology of Beranger et al. (2018) requires a parametric assumption of the model to be fitted, however, and so new methods need to be developed for symbolic data that fit within a non-parametric framework for cases where the uniformity-within-intervals assumption is known to be violated. In Chapter 5 we extend the generalised symbolic likelihood presented by Beranger et al. (2018) to this non-parametric framework, whereby general forms of symbolic statistics with underlying classical interpretations (for which the uniform assumption estimates are a special case) are presented, and examined as solutions to classical Estimating Equations (EE's). Methods of estimating each within-symbol distribution for intervals and histograms without the need for an underlying parametric or uniform assumption are presented that produce closer estimates to the classical case than estimates obtained via the uniformity-within-intervals assumption for certain settings. A symbolic empirical likelihood approach is then derived for the estimation of the variances of the symbolic EE estimates, with statistics such as means, variances, correlations and quantiles used in the demonstration of its applicability.

For privacy reasons, often count data is rounded to a known degree, such that the underlying classical datapoint from which a rounded observation arose is known to belong to a given subset

of count value, for which the analysis of is not trivial. In Chapter 6 we apply the previously derived new methodologies on the applied analysis of rounded count data. Symbolic equivalents are developed for several types of Generalised Linear Models (GLMs), such as Poisson, Binomial and Ordinal Logistic regression, which take into account the variation associated with rounded count data, along with the incorporation of extra marginal information. It is shown through simulations that the parameter estimates obtained from a symbolic GLM analysis are closer to those of the analysis of the complete (unknown) classical microdata than an analysis of the rounded data, with the results improving as the degree of rounding decreases. Furthermore, the utilisation of additional information that wasn't originally included in the models in this symbolic analysis further improves the results in terms of biases, variances and Mean Squared Error (MSE). As an example, for a Poisson analysis the number of women for a given group is treated as the response variable, with classically observed covariates. Due to privacy reasons, the response is only reported rounded to the nearest 5, and so the symbolic approaches described above can be used to improve on a classical analysis of the rounded data. The incorporation of the additionally known rounded observations for the total number of staff and men then allows a better specification of the distribution of a given underlying classical datapoint, given its rounded observation.

In Chapter 7 we conclude with a summary of the work developed within this thesis, along with a discussion of the potential impacts. We then comment on the potential future directions of SDA thesis, and discuss some open ended questions. Appendices and extra information for the work undertaken in this thesis are then given after the references and following the conclusion.

Chapter 2

Literature review

2.1 Introduction

The evolving capabilities of modern computation has led to increasingly large and complex datasets, the analysis of which is not trivial (Bock and Diday, 2000). These datasets are often aggregated into smaller dimensional summary level datasets, with the aim of reducing the burden associated with computation, storage and transmission, whilst still retaining enough information within the summary level datasets to produce meaningful and informative analyses. Diday (1989) first introduced the notion of Symbolic Data Analysis (SDA), whereby the underlying classical data is aggregated into a set of summary level (symbolic) statistics, where each symbol represents the data from a known subset of the microdata, termed a 'class'. Classes can be chosen using a broad range of criteria, such as temporal factors (e.g. each class represents a different month), known categorical factors (e.g. species, gender), or even completely randomly. Intervals, histograms, lists and distributions are examples of commonly used symbolic constructions, used to summarise underlying microdata Billard (2011). The statistical analyses performed on the symbolic observations then consider each symbol as an individual in the classical setting. As an example, a dataset with one observation per day could be aggregated into a dataset with one symbol per year.

n	City	Type	Age	Gender	Sys. pressure (mm Hg)	Dias. pressure (mm Hg)
1	Boston	Medical	24	M	120	79
2	Boston	Medical	56	M	130	90
3	Chicago	Dental	48	M	126	82
4	El Paso	Medical	47	F	121	86
5	Byron	Dental	79	F	150	88
6	Concord	Medical	12	M	126	85
7	Atlanta	Medical	67	F	134	89
8	Boston	Optical	73	F	121	81
...	...	...	...	...	...	...
...	...	...	...	...	...	...

Table 2.1: Classical data table for sample medical records

n	Type × gender	Age	Sys. pressure (mm Hg)	Dias. pressure (mm Hg)
1	Dental Males	[17,76]	[113,126]	[72,88]
2	Dental Females	[20,70]	[116,150]	[78,97]
3	Medical Males	[6,84]	[108,132]	[74,98]
4	Medical Females	[11,87]	[114,135]	[72,96]
5	Optical Males	[57,86]	[114,114]	[72,78]
6	Optical Females	[73,79]	[106,121]	[78,81]

Table 2.2: Symbolic data table for sample medical records

We now present a simple example of a classical and resultant symbolic big data table, taken as subsets of tables presented in Billard and Diday (2006), to demonstrate how underlying classical microdata can be aggregated into a set of symbols for each class. In Table 2.1, we have a classical data table, where each row represents a classical individual medical record observation, and the city, type of medical service used (Medical, Dental or Optical) and Gender (Male, Female) are recorded as categorical observations, with continous observations rounded to integer values given for Age, Systolic pressure and Diastolic pressure. If we then aggregate the continuous observations in this dataset into a set of univariate intervals, whereby each symbol represents the data from one unique combination of Type and Gender (class), we obtain the symbolic data table in Table 2.2. Note that there are only 6 rows, as that is the total number of unique combinations of gender and type, and thus the size of the dataset has been greatly reduced. In this example, a subsequent analysis of the smaller symbolic data table would be computationally superior to a classical analysis of the full underlying classical data.

2.2 Symbolic Descriptive Statistics

We now define the notation we will use throughout the rest of this chapter. Let $X = (X_{[1]}, ..., X_{[D]}) \in D_X$ be a $D-$dimensional random variable with domain $D_X = D_{X_{[1]}} \times ... \times D_{X_{[D]}}$, and suppose $\mathbf{X} = (X_1, ..., X_N)$ represent N *i.i.d.* replications of X, with realisations given as $x_1, ..., x_N$. Denote $\mathbf{X}^{(b)} = (X_1^{(b)}, ..., X_{c_b}^{(b)})$ as a subset of $\mathbf{X}$, such that $\mathbf{X} = \mathbf{X}^{(1)} \cup \cdots \cup \mathbf{X}^{(B)}$, with realisation given by $\boldsymbol{x}^{(b)}$. c_b therefore represents the size of the b^{th} subset, such that $\sum_{b=1}^{B} c_b = N$. Suppose our dataset $X_1, ..., X_N$ is aggregated into a set of symbols $\mathbf{S} = (S_1, ..., S_B)$, where each S_b represents a set of summary level statistics for all the underlying microdata from a given pre-specified class $\mathbf{X}^{(b)}$, with realisation given by s_b, $b = 1, ..., B$. Examples of how classes might be constructed are given in Section 2.1.

2.2.1 Interval-Valued Symbols

Suppose that $D_X = \mathbb{R}^D$ and that for each class, $b = 1, ..., B$, the data $X_1, ..., X_N$ is aggregated into a set of intervals $s_{bd} = (l_{bd}, u_{bd})$, where l_{bd} and u_{bd} represent respectively the minimum and maximum values for the underlying classical data for the b^{th} class and d^{th} variable, such that $s_b = (s_{b1}, ..., s_{bD})$, $b = 1, ..., B$, $d = 1, ..., D$. An equivalent paramerisation with a bijective mapping to (l_{bd}, u_{bd}) was proposed by Brito and Silva (2012) and is given as $s_{bd} = (m_{bd}, r_{bd})$,

where m_{bd} and r_{bd} represent respectively the midpoint and range of the interval. Using an assumption of uniformity for the distribution of the microdata within each interval, Bertrand and Goupil (2000) defined the empirical cumulative distribution and density functions for a set of B interval observations $\mathbf{s} = (s_1, ..., s_B)$ as follows.

Definition 2.2.1. If a uniform distribution is assumed for the microdata within each interval-valued symbol, then the cumulative distribution function (cdf) of a random variable for the d^{th} dimension, given its observed interval dataset is given as

$$F_{\mathbf{S}}(x_{[d]}) = \frac{1}{B}\sum_{b=1}^{B} F_{S_{bd}}(x_{[d]}),$$

where

$$F_{S_{bd}}(x_{[d]}) = \begin{cases} 0 & \text{if } x_{[d]} \leq l_{bd} \\ \frac{x_{[d]}-l_{bd}}{u_{bd}-l_{bd}} & \text{if } x_{[d]} \in [l_{bd}, u_{bd}] \\ 1 & \text{otherwise.} \end{cases}$$

Definition 2.2.2. Similarly, the density function for the observed interval dataset for the d^{th} dimension is given as

$$f_{\mathbf{S}}(x_{[d]}) = \frac{1}{B}\sum_{b=1}^{B} f_{S_{bd}}(x_{[d]}), \tag{2.1}$$

where

$$f_{S_{bd}}(x_{[d]}) = \begin{cases} \frac{1}{u_{bd}-l_{bd}} & \text{if } x_{[d]} \in [l_{bd}, u_{bd}] \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

Bertrand and Goupil (2000) then derive expressions for the sample mean and variance of the symbolic dataset using the within-symbol uniformity assumption, whereby the mean and variance of a classical observation within the bd^{th} interval are given respectively as $\mu_{bd} = \frac{l_{bd}+u_{bd}}{2}$ and $\sigma^2_{bd} = \frac{(u_{bd}-l_{bd})^2}{12}$, which are clearly the mean and variance of a random variable distributed uniformly on (l_{bd}, u_{bd}).

Definition 2.2.3. The symbolic sample mean and variance for the d^{th} dimension of an interval-valued dataset are given respectively as

$$\mu_d = \frac{1}{B}\sum_{b=1}^{B}\frac{l_{bd}+u_{bd}}{2} = \frac{1}{B}\sum_{b=1}^{B}\mu_{bd} \tag{2.3}$$

$$\sigma^2_d = \frac{1}{3B}\sum_{b=1}^{B}(u^2_{bd} + l_{bd}u_{bd} + l^2_{bd}) - \frac{1}{4B^2}\sum_{b=1}^{B}(u_{bd}+l_{bd})^2. \tag{2.4}$$

Billard (2007) rearranges Equation (2.4) to obtain an expression for the symbolic sample variance that has an intuitive interpretation, given as

$$\sigma_d^2 = \frac{1}{B}\sum_{b=1}^{B}(\mu_{bd} - \mu_d)^2 + \frac{1}{B}\sum_{b=1}^{B}\frac{(u_{bd} - l_{bd})^2}{12}. \tag{2.5}$$

We see that the first part of Equation (2.5) is the sample variance of the symbolic means $\mu_{1d}, ..., \mu_{Bd}$ of the interval observations, i.e. the between sum of squares (SSB) divided by the sample size, and the second part of Equation (2.5) is the average variance of a datapoint within each interval, i.e. the within sum of squares (SSW). Billard (2007) therefore show that the symbolic sample variance can be expressed as the summation of the within and between sum of squares, i.e.

$$\sigma_d^2 = \frac{1}{B}\left(SSW + SSB\right). \tag{2.6}$$

Equation (2.6) can be shown to be equivalent to the sample variance for the underlying microdata, if the microdata if truly uniformly distributed within each interval. This expression now seems intuitive, as the total symbolic sample variance is now given as the sum of the variance of the interval means and the within-interval variance of the microdata for each symbol, and in fact in Chapter 5 we explore the implications of this interpretation, with the consequence that the symbolic sample variance is close to that of the complete underlying microdata if the within-interval uniformity assumption is reasonable, but significantly different if this assumption is violated. Billard and Diday (2003) note that the above expressions for the distribution and density functions and the sample means and variances weight the contributions of each symbol equally, i.e. each term in each summation is equally weighted by $\frac{1}{B}$. In the construction of the symbols however this may not be valid if different numbers of observations contributed to each interval. The above expressions may therefore be extended to the case whereby each interval observation is assigned a weight p_b, resulting in the following expressions.

Definition 2.2.4. For weights $p_1, ..., p_B$, $\sum_{b=1}^{B} p_b = 1$, the cumulative distribution and density functions for the d^{th} dimension of an interval dataset are given as

$$F_{\mathbf{S}}(x_{[d]}) = \sum_{b=1}^{B} p_b F_{S_{bd}}(x_{[d]}). \tag{2.7}$$

Similarly, the density function of the interval dataset for the d^{th} dimension is given as

$$f_{\mathbf{S}}(x_{[d]}) = \sum_{b=1}^{B} p_b f_{S_{bd}}(x_{[d]}). \tag{2.8}$$

Definition 2.2.5. For weights $p_1, ..., p_B$, $\sum_{b=1}^{B} p_b = 1$, the symbolic sample mean and variance

of an interval dataset are given as

$$\mu_d = \sum_{b=1}^{B} p_b \frac{l_{bd} + u_{bd}}{2} = \sum_{b=1}^{B} p_b \mu_{bd} \tag{2.9}$$

$$\sigma_d^2 = \sum_{b=1}^{B} p_b (\mu_{bd} - \mu_d)^2 + \sum_{b=1}^{B} p_b \frac{(u_{bd} - l_{bd})^2}{12}. \tag{2.10}$$

For aggregation processes where different amounts of microdata contribute to each interval, the intuitive choice is $p_b = \frac{c_b}{N}$, where c_b represents the number of underlying datapoints that contributed to the b^{th} interval (i.e. the number of individuals in the b^{th} class), and N represents the total number of observations, such that $\sum_{b=1}^{B} c_b = N$. Billard (2003) then extend the univariate constructions of Bertrand and Goupil (2000) to the bivariate setting, deriving expressions for the bivariate cumulative distribution and density functions of a set of bivariate intervals (rectangles). We present the general case with p_b here, from which the specific equally weighted case with $p_b = \frac{1}{B}$ easily follows.

Definition 2.2.6. Assuming a uniform distribution within each bivariate interval-valued symbol, the cumulative distribution function (cdf) of the interval dataset with respect to variables d and e is given as

$$F_{\mathbf{S}}(x_{[d]}, x_{[e]}) = \sum_{b=1}^{B} p_b F_{S_{bde}}(x_{[d]}, x_{[e]}), \tag{2.11}$$

where

$$F_{S_{bde}}(x_{[d]}) = \begin{cases} \frac{(x_{[d]} - l_{bd})(x_{[e]} - l_{be})}{(u_{bd} - l_{bd})(u_{be} - l_{be})} & \text{if } (x_{[d]}, x_{[e]}) \in (l_{bd}, u_{bd}) \times (l_{be}, u_{be}) \\ 0 & \text{otherwise.} \end{cases} \tag{2.12}$$

Similarly, the density function of the interval dataset is given as

$$f_{\mathbf{S}}(x_{[d]}, x_{[e]}) = \sum_{b=1}^{B} p_b f_{S_{bde}}(x_{[d]}, x_{[e]}), \tag{2.13}$$

where

$$f_{S_{bde}}(x_{[d]}, x_{[e]}) = \begin{cases} \frac{1}{(u_{bd} - l_{bd})(u_{be} - l_{be})} & \text{if } (x_{[d]}, x_{[e]}) \in (l_{bd}, u_{bd}) \times (l_{be}, u_{be}) \\ 0 & \text{otherwise.} \end{cases} \tag{2.14}$$

Note that $f_{S_{bde}}(x_{[d]}, x_{[e]})$ is simply the density function of a variable that is uniformly distributed over the rectangle $(l_{bd}, u_{bd}) \times (l_{be}, u_{be})$. Bivariate expressions for the sample mean and variance are equivalent to that of the univariate case in Equations (2.9) and (2.10), due to the independence between margins for the estimation of those parameters. Two different definitions of the symbolic sample covariance are proposed in Billard (2007) and Billard (2008), given as follows.

Definition 2.2.7. For an interval dataset as described above, Billard (2007) defines the symbolic

sample covariance for margins d and e as

$$\text{Cov}(x_{[d]}, x_{[e]}) = \sigma_{de} = \sum_{b=1}^{B} c_b \mu_{bd} \mu_{be} - \mu_d \mu_e. \tag{2.15}$$

In contrast, Billard (2008) defines the symbolic sample covariance function $\text{Cov}(x_{[d]}, x_{[e]})$ as

$$\sigma_{de} = \sum_{b=1}^{B} \frac{c_b}{N} \frac{(u_{bd} - l_{bd})(u_{be} - l_{be})}{12} + \sum_{b=1}^{B} \frac{c_b}{N} (\mu_{bd} - \mu_d)(\mu_{be} - \mu_e). \tag{2.16}$$

Covariance matrices Σ, where $\Sigma_{c,d} = \sigma_{de}$ if $c \neq d$ and $\Sigma_{dd} = \sigma_d^2$ can then easily be obtained, along with the sample correlation function

$$\rho(x_{[d]}, x_{[e]}) = \frac{\sigma_{de}}{\sigma_d \sigma_e}.$$

An explanation for the differences in symbolic covariance definitions is provided in Oliveira et al. (2018), and involves the underlying classical data assumptions for each expression. Each covariance definition can in fact be considered the sample covariance of the underlying classical data, provided certain within-symbol distributional assumptions are met. For example, we see that Equation (2.15) has a simple interpretation as the covariance between the intervals means, meaning that if there is no within-symbol variance (i.e. every underlying classical datapoint occurs at an interval midpoint), this expression is equivalent to the covariance of the underlying microdata. Similarly to the sample symbolic variance, equation (2.16) reduces to the summation of the within sum of products and between sum of products for uniformly distributed bivariate variables, meaning that if the microdata follow independent uniform distributions within each interval, this definition of the sample symbolic covariance will be comparable to that of the full underlying classical dataset. The microdata conditions considered are fairly restrictive however, and not often applicable to real observed data. We provide more general definitions of the symbolic covariance in Chapter 5, along with better methods of estimating the within-symbol distributions/correlations.

Symbolic estimates for quantities such as means, variances and covariances/correlations are crucial in many types of statistical analyses. In particular, symbolic versions of Principal Component Analysis (PCA) have been developed for intervals by Billard and Le Rademacher (2012) and Gioia (2006), and for histograms by Le Rademacher and Billard (2013) in which definitions of symbolic means, variances and correlations are required. Other methods of symbolic PCA utilise the locations of the centers and ranges (Lauro and Palumbo, 2000, Douzal-Chouakria et al., 2011). A good overview of all of these symbolic PCA methods can be found in Le Rademacher (2008).

2.2.2 Histogram-Valued Symbols

Suppose now that the data for each class is aggregated into a set of univariate histograms $s_{bd} = (c_{bd1}; (y^d_{b0}, y^d_{b1}], ..., c_{bB^d_b}; (y^d_{b(B^d_b-1)}, y^d_{bB^d_b}])$, with B^d_b subintervals for the univariate histogram

for the b^{th} class and d^{th} variable, $y^d_{bc} > y^d_{b(c-1)}$, $b = 1, ..., B$, $c = 1, ..., B^d_b$ and

$$c_{bdc} = \sum_{n=1}^{N} \mathbb{I}(x_{nd} \in (y^d_{b(c-1)}, y^d_{bc}] \cap x_n \in \boldsymbol{x}^{(b)}).$$

Note that c_{bdc} therefore represents the number of observations for the b^{th} class whose d^{th} margin falls in the c^{th} bin of that respective histogram, such that $\sum_{b=1}^{B} \sum_{c=1}^{B^d_b} c_{bdc} = N$, $d = 1, ..., D$. Billard and Diday (2003) assume a uniform distribution for the microdata within each subinterval, allowing them to derive the following expression for the density function for histogram-valued data.

Definition 2.2.8. If a uniform distribution is assumed for the microdata within each subinterval for each univariate histogram as described above, then the density function (cdf) of the histogram dataset is given as

$$f_{\mathbf{S}}(x_{[d]}) = \sum_{b=1}^{B} \sum_{c=1}^{B_{bd}} \frac{c_{bdc}}{N} f_{S_{bdc}}(x_{[d]}), \tag{2.17}$$

where

$$f_{S_{bdc}}(x_{[d]}) = \begin{cases} \frac{1}{y^d_{bc} - y^d_{b(c-1)}} & \text{if } x_{[d]} \in (y^d_{b(c-1)}, y^d_{bc}] \\ 0 & \text{otherwise.} \end{cases} \tag{2.18}$$

Note that Equation (2.17) is equivalent to the density of a set of univariate intervals $(l_{bdc}, u_{bdc}]$ with proportions $p_{bdc} = \frac{c_{bdc}}{N}$, where $l_{bdc} = y^d_{b(c-1)}$ and $u_{bdc} = y^d_{bc}$.

Similarly to the interval case, expressions for sample means and variances can be obtained via the same uniformity within each subinterval assumption, meaning that the means and variances for each subinterval are given respectively as $\mu_{bdc} = \frac{y^d_{b(c-1)} + y^d_{bc}}{2}$ and $\sigma^2_{bdc} = \frac{(y^d_{bc} - y^d_{b(c-1)})^2}{12}$, $b = 1, ..., B$, $d = 1, ..., D$, $c = 1, ..., B_{bd}$.

Definition 2.2.9. For a histogram dataset as described above, the symbolic sample mean and variance of the d^{th} dimension are given as

$$\mu_d = \sum_{b=1}^{B} \sum_{c=1}^{B_{bd}} \frac{c_{bdc}}{N} \frac{y^d_{b(c-1)} + y^d_{bc}}{2} = \sum_{b=1}^{B} \sum_{c=1}^{B_{bd}} \frac{c_{bdc}}{N} \mu_{bdc}$$

$$\sigma^2_d = \sum_{b=1}^{B} \sum_{c=1}^{B_{bd}} \frac{c_{bdc}}{N} (\mu_{bdc} - \mu_d)^2 + \sum_{b=1}^{B} \sum_{c=1}^{B_{bd}} \frac{c_{bdc}}{N} \frac{(y^d_{bc} - y^d_{b(c-1)})^2}{12}.$$

Billard (2008) use an analogous derivation to the interval case to obtain an expression for the symbolic covariance for histogram-valued observations, again assuming uniformity within each histogram subinterval. As with the interval case, this expression is shown as the sum of the within observations sum of products and the between observations sum of products.

Definition 2.2.10. For a histogram dataset as described above, the symbolic sample covariance for dimensions d and e is given as

$$\sigma_{de} = \sum_{b=1}^{B}\sum_{c_1=1}^{B_{bd}}\sum_{c_2=1}^{B_{be}} \frac{c_{bdc_1}c_{bec_2}}{N}\frac{\delta_{bdc_1}\delta_{bec_2}}{12} + \sum_{b=1}^{B}\left\{\sum_{c_1=1}^{B_{bd}} \frac{c_{bdc_1}}{N}\left(\mu_{bdc_1}-\mu_d\right)\right\}\left\{\sum_{c_2=1}^{B_{be}} \frac{c_{bec_2}}{N}\left(\mu_{bec_2}-\mu_e\right)\right\},$$

where $\delta_{bdc} = y_{bdc} - y_{bd(c-1)}$, $b = 1, ..., B$, $c = 1, ..., B_{bd}$, $d = 1, ..., D$.

As with the interval case, these expressions are comparable to the classical equivalents, if the within-symbol uniformity assumption is correct. However, when estimates with an underlying classical interpretation are desired and this uniformity assumption is violated, different methods are needed to obtain better estimates.

2.3 Parametric models for symbolic datasets

When the within-symbol uniformity assumption is likely to be violated, a parametric approach can be used to assume some internal non-uniform structure within each symbol. Heitjan (1989) explores the use of Sheppard's correction (Sheppard, 1897) in the estimation of moments from an equally spaced histogram-valued dataset, assuming a reasonable normal structure to the underlying microdata. In the construction of the histograms, they assume the aggregation is done via a 'coarsening' of the data, where each bin count represents the number of underlying classical datapoints that are closest to that bin midpoint.

Definition 2.3.1. Suppose θ_{kbd} is the observed k^{th} moment of the observed midpoints data (weighted by their respective counts) for the b^{th} histogram and d^{th} variable, i.e. the observed moment of the set of histogram midpoints $\mu_{bdc} = \frac{y^d_{b(c-1)}+y^d_{bc}}{2}$, each observed c_{bdc} times, $b = 1, ..., B$, $c = 1, ..., B_{bd}$, $d = 1, ..., D$. Let $\delta_{bd} = y^d_{bc} - y^d_{b(c-1)}$ represent the standard bin width for each histogram. Then closer moment estimates (to that of the latent unrounded data) for the b^{th} histogram and d^{th} variable, which we denote as $\hat{\theta}_{kbd}$, can be obtained via the following correction formulae for $k = 1, 2, 3, 4$.

$$\begin{aligned}
\hat{\theta}_{1bd} &= \theta_{1bd} \\
\hat{\theta}_{2bd} &= \theta_{2bd} - \frac{\delta^2_{bd}}{12} \\
\hat{\theta}_{3bd} &= \theta_{3bd} \\
\hat{\theta}_{4bd} &= \theta_{4bd} - \frac{\delta^2_{bd}\theta_{2bd}}{2} + \frac{7\delta^4_{bd}}{240}.
\end{aligned}$$

Heitjan (1989) compares this approach with the raw midpoint estimates and estimates obtained using a weighted product of bin probabilities for each subinterval (which we will later see is a specific case of the general symbolic likelihood approach), where the bin probabilities are estimated using a normal parametric assumption. It is shown that if the normality assumption is reasonable, Sheppard's correction and the parametric method perform comparably and better

than the naive classic analysis of the observed midpoints, although Sheppard's correction is computationally simpler. Heitjan and Rubin (1991) explore the impact of the coarsening mechanism on the subsequent parametric analysis. Suppose we have a random vector $X \sim f_X(x;\theta)$, distributed according to a parametric distribution f with parameter θ. Suppose also that instead of observing X, we observe some aggregated version Y, a coarsened version of X.

Definition 2.3.2. Suppose the coarsening mechanism has a stochastic nature, in that the likelihood of Y being observed, given it arose from an underlying latent classical variable X, is denoted as $g(y|x;\theta)$. The likelihood of the observed coarsened variables is then given as

$$L(\theta;y) \propto \int_{D_{X|X\to Y}} g(y|x;\theta) f_X(x;\theta) dx, \tag{2.19}$$

where $D_{X|X\to Y}$ represents the domain of the underlying microdata, given the observed histograms. If the data are considered Coarsened at Random (CAR), then the density function $g(y|x;\theta)$ is uniformative and can be ignored in the parametric analysis of the data.

For interval and histogram-valued symbolic data, Le Rademacher and Billard (2011) proposed a likelihood based approach whereby the estimates obtained from the analysis are interpretable at the symbolic level. A set of vectors $\theta_1, ..., \theta_B$ is created whereby each vector uniquely defines a distinct symbol, i.e. a bijective mapping exists between S_b and θ_b, $b = 1, ..., B$. A parametric model is then specified for each symbol, from which parameter estimates can be obtained via Maximum Likelihood Estimation.

Example. As an example, suppose have a set of $D-$dimensional intervals (with equal counts) $s_b = (s_{b1}, ..., s_{bD})$, $s_{bd} = (l_{bd}, u_{bd})$, $b = 1, ..., B$, $d = 1, ..., D$, as described in Section 2.2.1. Define $\theta_{bd} = (\theta_{bd1}, \theta_{bd2})$ as the vector arising from a bijective mapping from s_{bd}, where θ_{bd1} and θ_{bd2} represent the interval mean and variance respectively, $b = 1, ..., B$, $d = 1, ..., D$. We now specify a normal $N(\mu_d, \sigma_d^2)$ distribution with mean μ_d and variance σ_d^2 for θ_{bd1}, and an exponential $\text{Exp}(\beta_d)$ distribution with mean β_d for θ_{bd2}, $d = 1, ..., D$, such that an assumption of uniformity within each symbol results in

$$\begin{aligned} \frac{l_{bd} + u_{bd}}{2} &\sim N(\mu_d, \sigma_d^2) \\ \frac{(u_{bd} - l_{bd})^2}{12} &\sim \text{Exp}(\beta_d). \end{aligned}$$

Note that we are assuming independence between the interval means and variances, and also between margins. The likelihood for the d^{th} margin is then given as

$$L(\mu_d, \sigma_d^2, \beta_d; \beta_{1d}, ..., \beta_{Bd}) = \prod_{b=1}^{B} g(\theta_{bd1}; \mu_d, \sigma_d^2) h(\theta_{bd2}; \beta_d), \tag{2.20}$$

where

$$g(\theta_{bd1}; \mu_d, \sigma_d^2) = \frac{1}{\sqrt{2\pi\sigma_d^2}} \exp\left(-\frac{(\frac{l_{bd}+u_{bd}}{2} - \mu_d)^2}{2\sigma_d^2} \right)$$

and

$$h(\theta_{bd2};\beta_d) = \frac{1}{\beta_d}\exp\left(\frac{(u_{bd}-l_{bd})^2}{12\beta_d}\right)$$

are the density functions for the normal and exponential distributions respectively. Differentiating the subsequent log-likelihood with respect to each parameter, and then solving for zero provides the following Maximum Likelihood Estimators (MLE's).

$$\hat{\mu}_d = \frac{1}{B}\sum_{b=1}^{B}\frac{l_{bd}+u_{bd}}{2}$$
$$\hat{\sigma}_d^2 = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{l_{bd}+u_{bd}}{2}-\mu_d\right)^2$$
$$\hat{\beta} = \frac{1}{B}\sum_{b=1}^{B}\frac{(u_{bd}-l_{bd})^2}{12}.$$

More complex examples involving multivariate and dependent data, as well as histogram equivalents are presented in Le Rademacher and Billard (2011). Brito and Silva (2012) utilise a similar approach by fitting normal and skew-normal distributions to the means and log ranges of interval-valued datasets, assuming again uniformity within each symbol. Lin et al. (2017) utilise a Bayesian approach in fitting distributions to the means and log ranges of interval in a meta analysis of interval-valued count data for species. These approaches are useful for obtaining symbolic-level parametric inferences for interval and histogram data, and the output of such an analyses has an intuitive symbolic explanation, however is insufficient if the practitioner wishes to obtain inferences that are interpretable at the level of the underlying latent classical data.

Beranger et al. (2018) proposed a generalised likelihood construction for symbolic data, from which inferences with a classical interpretation can be obtained, provided enough information about the underlying latent microdata is retained. This approach allows the generalised parametric modelling of symbolic datasets, from which analyses such as Zhang et al. (2019), Koo and Kooperberg (2000), Cadez et al. (2002) and Heitjan (1989) can be considered special cases for specific types of symbolic observations. This symbolic likelihood framework can be easily adapted to new symbol designs, as highlighted by its implementation for different forms of interval constructions in Beranger et al. (2018). Suppose $X \sim f_X(x;\theta)$ is distributed according to a parametric distribution f with parameter vector θ, and denote N *i.i.d.* realisations of X as $x_1, ..., x_N$. Suppose now that instead of observing $\boldsymbol{x} = (x_1, ...x_N)$, we observe a set of symbols $\mathbf{s} = (s_1, ..., s_B)$, whereby c_b underlying classical observations were aggregated into the b^{th} symbol and each observation contributed to exactly one symbol.

Definition 2.3.3. The likelihood of each symbolic observation s_b, given the parameter vector and information about the aggregation process Q, is given as

$$L(s_b;\theta,Q) = \int_{D_{\mathbf{X}}} L(\boldsymbol{x};\theta) f_{s_b|\mathbf{X}}(s_b|\boldsymbol{x},Q)d\boldsymbol{x}, \tag{2.21}$$

where $f_{s_b|\mathbf{X}}(s_b|\boldsymbol{x},Q)$ is the density of the observed symbol s_b, given the underlying classical

observations $\boldsymbol{x}$ from which it arose. The likelihood and log-likelihood of the symbolic dataset $\mathbf{s} = (s_1, ..., s_B)$ are therefore given as

$$L(\mathbf{s}; \theta, Q) = \prod_{b=1}^{B} L(s_b; \theta, Q)^{c_b} \tag{2.22}$$

$$\rightarrow \log L(\mathbf{s}; \theta, Q) = \sum_{b=1}^{B} c_b \log L(s_b; \theta, Q). \tag{2.23}$$

Example. Suppose $D = 1$ and each s_b represents a univariate interval-valued observation with count c_b, $b = 1, ..., B$. Using equations (2.22) and (2.23) and specifying a parametric model for the underlying classical data allows the construction of

$$L(s; \theta, Q) = \prod_{b=1}^{B} P_b(\theta)^{c_b}$$

$$\rightarrow \log L(s; \theta, Q) = \sum_{b=1}^{B} c_b \log P_b(\theta)),$$

as the respective likelihood and log-likelihood for the symbolic dataset $\mathbf{s} = (s_1, ..., s_B)$, where $P_b(\theta) = \int_{l_b}^{u_b} f_X(x; \theta) dx$.

For intervals, it is shown in Zhang et al. (2019) that MLE's obtained via the maximisation of the above log-likelihood are asympotically consistent with that of a likelihood analysis performed on the underlying classical microdata, with decreasing interval widths. As each interval gets smaller, more information is retained about the underlying microdata from which that symbol arose from, leading to more accurate results.

Parametric analyses such as those outlined above are often sufficient in the analysis of low-dimensional symbolic datasets. However, an increase in data dimension often leads to intractable high-dimensional likelihood functions, in which high-dimensional integrals are often required to be evaluated numerically. Furthermore, the curse of dimensionality means that the computation required to evaluate said integrals grows exponentially with the number of variables. This leads to large computational costs, for even for a moderate number of variables D, often making a parametric symbolic analysis infeasible. If a model with an underlying classical interpretation is desired, new methods need to be developed that reduce the dimension of the symbolic likelihoods, and therefore reduce the computational burden. In Chapter 3 we extend the likelihood framework of (Beranger et al., 2018) to the composite likelihood setting, allowing high-dimensional datasets to be reduced to low-dimensional marginal histograms, and subsequently analysed. In Chapter 4 we develop a new composite marginal approach to logistic regression modelling to similarly reduce the dimension and subsequent computational burden of the symbolic likelihood.

2.4 Non-Parametric models for SDA

When a parametric model can be assumed for the underlying classical data, it follows that each symbolic observation is assigned an internal structure, given the parameter vector. However, when we are unable to specific a parametric model for the data, due to reasons such as a lack of information, a desire for flexibility or no known parametric family seems to model underlying data at least reasonably well, different methods are needed to specify the internal structure of each symbol, beyond reverting back to the within-symbol uniformity assumption. An intuitive starting point is to estimate the underlying non-parametric density using the observed symbols.

Scott (1985) investigated the theoretical properties of the frequency polygon, and showed that the density estimate obtained from this construction is far superior in terms of MSE than that obtained from the original histogram. The frequency polygon is a computationally simple construction of a density estimate from histogram-valued data, and is constructed by connecting the midpoints of adjacent histogram bins at heights equal to the proportion of data within each bin. Jones (1998) proposed a similar estimator, denoted as the edge frequency polygon, in which the edges of each histogram bin are joined by straight lines, with their weights calculated as the average of weights of the adjacent histogram bins. Billard and Diday (2003) explored the use of a histogram constructed from a set of intervals as a means of representing the data. The constructed histograms can then be used to obtain estimates for quantities such as the symbolic means and variances, or covariances in the multidimensional setting. We now present their constructed histograms for observed interval datasets, adjusted for the setting whereby the counts for each symbol are potentially unequal. A set of $C+1$ break points $(y_0, ..., y_C)$ are specified, such that $\delta_c = y_c - y_{c-1}$, $c = 1, ..., C$ and there are C subintervals in total.

Definition 2.4.1. For univariate, unequally weighted interval-valued data $s_b = (c_b, (l_b, u_b])$, $b = 1, ..., B$, with $\sum_{b=1}^{B} c_b = N$, the Billard and Diday (2003) histogram density estimate for a point $x \in D_X$ is given as

$$f(x) = \frac{1}{N} \sum_{c=1}^{C} \frac{p_c}{N} \mathbb{I}(x \in (y_{c-1}, y_c]),$$

where

$$p_c = \sum_{b=1}^{B} \frac{c_b}{N} \frac{|(y_{c-1}, y_c) \cap (l_b, u_b)|}{|(l_b, u_b)|}$$

and $|A|$ represents the length of the region A.

Methods that have been used to estimate a non-parametric curve from non-standard data include spline methods (Gu, 1993, Smith et al., 2004, Rizzi et al., 2016) and a bootstrap kernel method (Wang and Wertelecki, 2013). Hall (1982) investigated the bias that occurs when performing a Kernel Density Estimation (KDE) analysis (Parzen, 1962) on the midpoints of binned data, and showed that as the bin width of each histogram decreases, the binned KDE estimator approaches the density estimate obtained from classical KDE analysis on the underlying microdata. Scott and Sheather (1985) examined the errors associated with performing a classical kernel density analysis on a rounded sample, and provide formulas for the errors

associated with performing a classic KDE analysis on histogram bin midpoints, as compared to an analysis of the underlying latent data. This extends to formulas that allow the evaluation of the 'optimal' smoothness parameter for the binned KDE analysis, which has equivalent forms in the classial setting. Hall (1996) then provided further results, whereby the mean squared error of the binned KDE estimator can be compared to that of the classical case, and also the true unknown underlying classical density. This allows the development of formulas that allow the practitioner to determine the minimum bin width required for each histogram, in order to achieve a specified level of accuracy.

Definition 2.4.2. Suppose $X_1, ..., X_N$ are N *i.i.d.* univariate random variables, distributed according to an unknown distribution f. The kernel density estimator, as described by Parzen (1962), is denoted as

$$f_N(x; h) = \frac{1}{Nh} \sum_{n=1}^{N} K\left(\frac{x - x_n}{h}\right), \tag{2.24}$$

where h is a specified bandwidth parameter and K is a given kernel function. Common examples of kernel functions K include the standard normal and uniform distributions.

Definition 2.4.3. Suppose now that $\boldsymbol{x} = (x_1, ..., x_N)$ is binned using C known bins $y_0 < ... < y_C$, where $c_c = \sum_{n=1}^{N} \mathbb{I}(x_n \in (y_{c-1}, y_c])$, $\delta = \delta_c = y_c - y_{c-1}$ and $m_c = \frac{y_{c-1}+y_c}{2}$, $c = 1, ..., C$. Hall (1982) and Scott and Sheather (1985) define the binned kernel density estimator as

$$f_N(x; h, s) = \frac{1}{Nh} \sum_{c=1}^{C} c_c K\left(\frac{x - m_c}{h}\right). \tag{2.25}$$

Equations for the optimal value of the bandwidth parameter h can then be derived using similar methodologies to their counterparts in the classical setting.

Some interesting similarities exist between the fields of missing data analysis, measurement error analysis and SDA, in that in some settings the missing data/measured-with-error variables can themselves be considered symbolic observations. Wang and Pepe (2000) proposed an expected estimating equations approach to accomodate the fact that some variables/observations are measured with error. In this approach, the expectation of the classical estimating equations for the measured-with-error observations are calculated using the data that is fully observed, and arrives in classical pointwise form. While this method isn't applicable to the field of SDA, given in SDA often there are no classical observations available, it does use a similar rationale to the estimators we derive in Chapter 5, whereby the internal non-parametric structure of each symbol is estimated according to its expectation given the rest of the symbolic dataset. Wang et al. (2008) extend this expected estimating equations methodology to the missing data, covariate measurement error and missclassification settings. For covariate measurement error, the contributions of missing covariates to the estimating equation are again replaced by their expectations, given some prior specification. In the examples they provided, this prior specification involved a joint parametric assumption for the underlying missing microdata and the errors, which can lead to issues if the parametric assumption is violated or cannot be made. For

missing and missclassified data, the framework is largely similar. The contributions of the missing/missclassified observations to the estimating equations are replaced by their expectations with respect to a parametric model dependent on available fully observed data. Elashoff and Ryan (2004) propose the use of the Expectation-Maximisation (EM) algorithm to estimate the expectation of missing data estimating equations, again using classically observed data. Zhou et al. (2008) also address the missing data problem for non-parametric estimating equations by using kernels to impute their contributions. This approach removes the need for a parametric assumption, but is still not generally applicable in the field of SDA, as the imputed kernels are still constructed using available fully observed classical data.

Approaches such as those described above involving the evaluation of the expectation of estimating equations, given some non-standard data, have been extended to the generalised linear model (GLM) setting by Lipsitz et al. (2004), who derive likelihood functions for GLMs for data where some discrete covariates are rounded to a known degree. A distribution is defined for the coarsened covariate, given the fully observed data. Given the discrete nature of the covariate data, the likelihood function has a closed form analytical expression involving simple summations over the possible classical estimating equations, and thus its optimisation is computationally reasonable. In chapter 6 we develop a similar method to analyse rounded discrete data, whereby the distribution of the coarsened covariate is defined using other rounded observations, as well as additional information not originally utilised in the model. Johnson (2006) extend the methodology of Lipsitz et al. (2004) to the case of continuous rounded variables, and discuss the computational issues associated with the evaluation of the likelihood, given an intractable integral. Some of these computational issues can be addressed using the methodology presented in Chapter 4, albeit only for logistic regression. A similar method employing a Bayesian approach is then employed in Johnson and Wiest (2014), whereby a parametric framework is assumed for the parameters of the model, leading to computationally easier analyses.

2.5 Conclusion

Symbolic data can arise in a number of different forms from the aggregation of an underlying classical dataset. Methods of analysis and inference for symbolic data are currently well developed if interpretations at the symbolic level are desired. However if the practitioner wishes to obtain results with an underlying classical interpretation, then new methods of analysis need to be developed. Furthermore, methods are needed that not only possess a classical interpretation, but also produce comparable results to an equivalent classical analysis performed on the underlying microdata for a certain level of aggregation. If the within-symbol uniformity assumption is reasonable, then current SDA methodologies, such as the formulas described above and derived by Bertrand and Goupil (2000), Billard and Diday (2003) and Billard (2007), are sufficient in providing comparable results to the classical case. However, this assumption is often violated, as the underlying classical data is often continuous and generated from a non-uniform process. When this within-symbol uniformity assumption is violated, different forms of parametric analyses are available whereby an underlying classical parametric model is fit to a set of symbolic

data.

When the parametric assumption described above is violated, or the practitioner does not wish to be restricted to a parametric framework, then non-parametric statistics provides methods of obtaining inferences in both the classical and symbolic setting. In the symbolic setting, kernel estimation methods exist for various types of symbolic data, revolving largely around performing a classical analysis on interval midpoints. New methods are needed that are able to better estimate the internal structure of a set of symbols, that don't rely on a parametric assumption and takes into account the internal variation. A consequence of better estimation of the underlying non-parametric density/dataset is that better estimates for internal parameters of a symbol can be obtained, such as means and variances, which have many applications including hypothesis testing and PCA.

The contributions of this thesis are as follows: In Chapter 3, we extend the generalised symbolic likelihood construction of Beranger et al. (2018) to the composite likelihood framework, demonstrating its application in an analysis of high-dimensional max-stable data. In Chapter 4 we then utilise the symbolic likelihood framework to develop logistic regression models for classification data with histogram-valued covariates. In Chapter 5 we develop a non-parametric framework for symbolic estimating equations and derive methods that allow the better estimation of within-symbol and sample quantities, such as means, variances, etc., with variances obtained through a derived symbolic empirical likelihood method. In Chapter 6 we apply the symbolic framework to GLMs with rounded, discrete data, with specific examples given for poisson, binomial and ordinal logistic regression, and utilise this methodology in the real data GLM analysis of count data. A discussion then follows in Chaper 7, along with references and appendices.

Chapter 3

A composite likelihood approach for histogram-valued random variables

3.1 Introduction

Continuing advances in measurement technology and information storage are leading to the creation of increasingly large and complex datasets. This inevitably brings new inferential challenges. Symbolic data analysis (SDA), a relatively new field in statistics, has been developed as one way of addressing these issues (e.g. Diday, 1989, Bock and Diday, 2000). In essence, SDA argues that many important questions can be answered without needing to observe data at the micro-level, and that higher-level, group-based information may be sufficient. As a result, SDA methodology aggregates the micro-data into a much smaller number of distributional summaries, such as random rectangles, random histograms and categorical multi-valued variables, each summarising a portion of the larger dataset (Dias and Brito, 2015, Le Rademacher and Billard, 2013, Billard and Diday, 2006). These new data "points" (i.e. distributions) are then analysed directly, without any further reference to the micro-data. See e.g. Billard (2011), Bertrand and Goupil (2000) and Billard and Diday (2003) for an exposition of these ideas.

SDA methods have found wide application, and have been developed for a range of inferential procedures, including regression models (Dias and Brito, 2015), principle component analysis (Kosmelj and Billard, 2014), time series analysis (Wang et al., 2016), clustering (Brito et al., 2015), discriminant analysis (Silva and Brito, 2015) and Bayesian hierarchical modelling (Lin et al., 2017). Likelihood-based methods for distributional data were introduced by Le Rademacher and Billard (2011) for direct modelling at the level of the distributional summary.

More recently, Zhang et al. (2019) and Beranger et al. (2018) developed likelihood functions for observed random rectangles and histograms that directly accounts for the process of constructing the symbols from the underlying micro-data. By explicitly considering the full generative process – from micro-data generation to constructing the resulting distributional summary – the resulting symbolic likelihood allows the fitting of the standard micro-data likelihood, but while only observing the distributional-based data summaries. The symbolic likelihood reduces

to the standard micro-data likelihood as the observed symbols reduce to the underlying micro-data (e.g. as the number of histogram bins gets large, and the size of each histogram bin gets small). Beranger et al. (2018) demonstrate a 14× computational speed up for the symbolic analysis over the standard micro-data analysis for computing the maximum likelihood estimates of a hierarchical skew-normal model.

While attractive, a limitation of this approach is that grid-based multivariate histograms become highly inefficient as data summaries as the dimension of the data increases. This means that the histogram-based approach in Beranger et al. (2018), where the computational overhead is proportional to the number and dimension of histogram bins, is practically limited to lower-dimensional data analyses.

In this paper we address this problem by extending the likelihood-based approach of Beranger et al. (2018) to the composite-likelihood setting. Focusing on histogram-based distributional summaries, the components of the composite likelihood are constructed based on low-dimensional marginal histograms derived from the full K-dimensional histogram. We demonstrate consistency of the resulting symbolic composite maximum likelihood estimator, and show that for a certain level of data aggregation, the symbolic composite likelihood function provides a useful and more computationally efficient substitute for the standard micro-data analysis. We obtain results that describe the reduction in information that occurs when aggregating the micro-data into histograms, and how this reduction is dependent on the number of observed histograms. These results also provide insights on the efficiency of standard composite likelihood techniques when the micro-data are grouped into blocks, but where the location of data within each block is not known.

While the above techniques are general, throughout we are motivated by the need to develop computationally viable statistical techniques for fitting max-stable process models for spatial extremes. This becomes particularly challenging when both the number of spatial dimensions K (the number of physical recording stations) and the number of observations over time (N) become large, as is the case with millennial scale climate simulations (Huang et al., 2016). While composite-likelihood techniques (Padoan et al., 2010, Blanchet and Davison, 2011, Varin et al., 2011, Lee et al., 2013, Castruccio et al., 2016, Beranger et al., 2019) provide one way to approach the issue of spatial dimensions, they are not able to cope with large amounts of observed data at each spatial location. By developing composite likelihood techniques for the analysis of K-dimensional histogram-valued random variables, we are able to directly and efficiently fit max-stable processe models to very large temporal datasets.

This article is structured as follows: In Section 3.2 we describe the ideas behind the symbolic likelihood framework of Beranger et al. (2018), with a focus on histogram-valued random variables, extend this approach to the case of a marginal histogram, and briefly present relevant background on composite likelihood methods.

In Section 3.3 we extend the histogram-based symbolic likelihood function to the composite likelihood setting. We demonstrate that increasing the number of bins (and reducing their size) in each histogram yields composite maximum likelihood estimators (MLEs) that are asymptotically consistent with those of the classical (micro-data) setting, but at a potentially much

cheaper computational cost. While these composite MLEs retain this asymptotic consistency regardless of the method of histogram construction (as long as the volume of each bin approaches zero as the number of bins approaches infinity) and how many random histograms are used, their variances depend heavily on the amount of temporal information retained during the data aggregation process. Accordingly we show that increasing the number of random histograms leads to an overall decrease in the variance of the composite MLE. In Section 3.4 we explore the performance of the histogram-based composite likelihood function through simulation studies using max-stable processes, and in Section 3.5 we analyse real and future-simulated datasets comprising daily maxima temperature data from 105 locations across Australia. We conclude with a Discussion.

3.2 Symbolic and composite likelihoods

We first provide a brief overview of likelihood-based methods for *symbolic* random variables, in particular focusing on histogram-valued random variables and the approach of Beranger et al. (2018). Motivated by a desire to reduce computational overheads as the dimension of the histogram K increases, we extend this setup to the case of a *marginal*-histogram (i.e. a lower-dimensional margin of an original histogram). We then briefly review the ideas behind composite likelihoods in a general setting.

3.2.1 Generative symbolic likelihoods

In simple terms, symbolic random variables are distributional-valued random variables that are constructed by the aggregation of standard, classical random variables into a distributional summary form, such as a random interval or random histogram. Symbolic data analysis is the study and analysis of symbolic random variables (Billard, 2011, Billard and Diday, 2003, Bock and Diday, 2000). Within this field, two main likelihood-based techniques have been developed for the analysis of symbolic data; one based on analysing the symbols directly (Le Rademacher and Billard, 2011, Brito and Silva, 2012, Lin et al., 2017) and one based on also modelling the construction of the symbols from the generating process of the classical random variables (Beranger et al., 2018, Zhang et al., 2019). This latter technique allows for the use of symbolic data analysis methods as a means to expedite standard data analyses for large and complex datasets. We adopt both this approach and motivation here.

The general construction of Beranger et al. (2018) is given as follows. Denote by $\boldsymbol{X} = (X_1, \ldots, X_N)$ a vector of i.i.d. classical random variables, which takes values in some space $D_{\boldsymbol{X}}$ and has density $g_{\boldsymbol{X}}(\cdot\,;\theta)$ with unknown parameter vector θ. Each X_i takes values in D_X and has density $g_X(\cdot\,;\theta) = \int g_{\boldsymbol{X}}(\cdot\,;\theta) d\boldsymbol{X}_{-i}$ where $\boldsymbol{X}_{-i} = \boldsymbol{X}/X_i$, so that $D_{\boldsymbol{X}} = (D_X)^N$. The observed values $\boldsymbol{x}$ of $\boldsymbol{X}$ can then be aggregated into a distribution-valued symbol s, itself a realisation of some symbolic random variable $S \in D_S$, according to a known function $f_{S|\boldsymbol{X}=\boldsymbol{x}}(s|\boldsymbol{x}, \phi)$. The likelihood associated with the process of generating and constructing the observed symbol s is

then given by

$$L(s;\theta,\phi) \propto \int_{D_{\boldsymbol{X}}} f_{S|\boldsymbol{X}=\boldsymbol{x}}(s|\boldsymbol{x},\phi) g_{\boldsymbol{X}}(\boldsymbol{x};\theta) d\boldsymbol{x}. \tag{3.1}$$

That is, $L(s;\theta,\phi)$ is the expectation of the classical data likelihood $g_{\boldsymbol{X}}(\boldsymbol{x};\theta)$ over all possible classical datasets $\boldsymbol{x}$ that could have produced the observed symbol s.

Beranger et al. (2018) considered several forms for $f_{S|\boldsymbol{X}=\boldsymbol{x}}(s|\boldsymbol{x},\phi)$ that allowed for different types of symbol (e.g. random intervals, hyper-rectangles and different forms of random histogram) and accordingly different resulting forms of symbolic likelihood function. Here we focus on the fixed-bin, random-counts histogram, although extension of the results in this article to other symbolic likelihood forms is possible.

Suppose that $X_1,\ldots,X_N$ are K-dimensional random vectors with $D_X = \mathbb{R}^K$. The collection of N classical data observations $\boldsymbol{x} \in \mathbb{R}^{N\times K}$ may be aggregated into a K-dimensional histogram on D_X, where the k-th margin of D_X is partitioned into $B^k \in \mathbb{N}$ bins, so that $B^1 \times \cdots \times B^K$ bins are created in D_X through the K-dimensional intersections of each marginal bin. Indexing each bin $\boldsymbol{b} = (b_1,\ldots,b_K)$, $b_k = 1,\ldots,B^k$, as the vector of marginal bin indices, bin $\boldsymbol{b}$ may be constructed over the space $\Upsilon_{\boldsymbol{b}} = \Upsilon^1_{\boldsymbol{b}} \times \cdots \times \Upsilon^K_{\boldsymbol{b}}$, where $\Upsilon^k_{\boldsymbol{b}} = (y^k_{b_k-1}, y^k_{b_k}] \subset \mathbb{R}$, and where, for each margin k, $-\infty < y^k_0 < y^k_1 < \ldots < y^k_{B^k} < \infty$ are fixed points that define the change from one bin to the next. That is, $\boldsymbol{b}$ describes the coordinates of a bin within the K-dimensional histogram and $\Upsilon_{\boldsymbol{b}} \subseteq \mathbb{R}^K$ defines the space that it covers.

Now let $S_{\boldsymbol{b}}$ denote the random number of observed data points $X_1,\ldots,X_N$ that fall in bin $\boldsymbol{b}$. Then $\boldsymbol{S} = (S_{\boldsymbol{1}},\ldots,S_{\boldsymbol{B}})$ is the vector of counts from the first bin $\boldsymbol{1} = (1,\ldots,1)$ to the last bin $\boldsymbol{B} = (B^1,\ldots,B^K)$, of length $B^1 \times \cdots \times B^K$, and which satisfies $\sum_{\boldsymbol{b}} S_{\boldsymbol{b}} = N$. That is, $\boldsymbol{S}$ is a random histogram with N observations. Following Beranger et al. (2018), and assuming that $g_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^N g_X(x_i;\theta)$, the resulting symbolic likelihood function (3.1) then becomes

$$L(\boldsymbol{s};\theta) \propto \frac{N!}{s_{\boldsymbol{1}}!\ldots s_{\boldsymbol{B}}!} \prod_{\boldsymbol{b}=\boldsymbol{1}}^{\boldsymbol{B}} P_{\boldsymbol{b}}(\theta)^{s_{\boldsymbol{b}}}, \tag{3.2}$$

where $\boldsymbol{s} = (s_{\boldsymbol{1}},\ldots,s_{\boldsymbol{B}})$ is the observed value of $\boldsymbol{S}$, and where $P_{\boldsymbol{b}}(\theta) = \int_{\Upsilon_{\boldsymbol{b}}} g_{\boldsymbol{X}}(\boldsymbol{z};\theta) d\boldsymbol{z}$ is the probability of observing a datapoint in bin $\Upsilon_{\boldsymbol{b}}$ under the model $g_{\boldsymbol{X}}(\boldsymbol{x};\theta)$. (The ϕ parameter in (3.1), which controls quantities relevant to constructing the symbol, is fixed in this setting, and so we omit it from subsequent notation.) This multinomial form of likelihood makes intuitive sense in that maximising this likelihood amounts to choosing parameters θ that optimally match the empirical bin proportions with the corresponding bin probabilities under the model $g_{\boldsymbol{X}}(\boldsymbol{x};\theta)$.

Looking ahead to Section 3.3 where we will be constructing composite symbolic likelihood functions, suppose that we are only interested in a subset of the K dimensions, represented by some index set $\boldsymbol{i} = (i_1,\ldots,i_I) \subseteq \{1,\ldots,K\}$, where for convenience $i_1 < \ldots < i_I$. We may then construct the associated I-dimensional marginal histogram, defining $\boldsymbol{b^i}$ as the subvector of $\boldsymbol{b}$ containing those elements corresponding to the index set $\boldsymbol{i}$. (We use this notation more generally, so that a vector with superscript $\boldsymbol{i}$ means the subvector containing those elements corresponding to the index set $\boldsymbol{i}$.) Then if $S^{\boldsymbol{i}}_{\boldsymbol{b^i}}$ is the random number of observed data points $X^{\boldsymbol{i}}_1,\ldots,X^{\boldsymbol{i}}_N$ that fall in bin $\boldsymbol{b^i}$, we may construct an I-dimensional random *marginal* histogram

$\boldsymbol{S^i} = (S^i_{\mathbf{1}^i}, \ldots, S^i_{\boldsymbol{B}^i})$ as the associated vector of random counts from the first bin $\mathbf{1}^i = (1, \ldots, 1)$ to the last bin $\boldsymbol{B}^i = (B^{i_1}, \ldots, B^{i_I})$. The vector $\boldsymbol{S^i}$ has length $B^{i_1} \times \ldots \times B^{i_I}$ and satisfies $\sum_{\boldsymbol{b}^i} S^i_{\boldsymbol{b}^i} = N$.

Note that we can write $\boldsymbol{S}^i_{\boldsymbol{b}^i} = \sum_{\tilde{\boldsymbol{b}}:\tilde{\boldsymbol{b}}^i=\boldsymbol{b}^i} \boldsymbol{S}_{\tilde{\boldsymbol{b}}}$ so that we are effectively marginalising out the non-indexed set $-\boldsymbol{i} = \{1, \ldots, K\}/\boldsymbol{i}$ dimensions of the histogram $\boldsymbol{S}$. Hence, $\boldsymbol{S^i}$ is truly a marginal histogram of $\boldsymbol{S}$ in the usual sense of the term.

Similarly to (3.2), the resulting symbolic likelihood function for the marginal histogram $\boldsymbol{S^i}$ is then given by

$$L(\boldsymbol{S^i}; \theta) \propto \frac{N!}{s^i_{\mathbf{1}^i}! \cdots s^i_{\boldsymbol{B}^i}!} \prod_{\boldsymbol{b}^i=\mathbf{1}^i}^{\boldsymbol{B}^i} P_{\boldsymbol{b}^i}(\theta)^{s^i_{\boldsymbol{b}^i}}, \tag{3.3}$$

where $\boldsymbol{s^i} = (s^i_{\mathbf{1}^i}, \ldots, s^i_{\boldsymbol{B}^i})$ denotes the observed value of $\boldsymbol{S^i}$ and $P_{\boldsymbol{b}^i}(\theta) = \int_{\Upsilon_{\boldsymbol{b}^i}} g^i_{\boldsymbol{X}^i}(\boldsymbol{z^i}; \theta) d\boldsymbol{z^i}$ is the probability of observing a datapoint in the I-dimensional marginal bin $\Upsilon_{\boldsymbol{b}^i}$ under the marginal model $g^i_{\boldsymbol{X}^i}(\boldsymbol{x^i}; \theta) = \int g_{\boldsymbol{X}}(\boldsymbol{z}; \theta) d\boldsymbol{z}^{-\boldsymbol{i}}$, where $\boldsymbol{z}^{-\boldsymbol{i}}$ is the vector of elements of $\boldsymbol{z}$ that are not in $\boldsymbol{z^i}$. In the case where $I = \{1, \ldots, K\}$ then (3.3) is equal to (3.2).

Following similar arguments to Beranger et al. (2018), the symbolic likelihood $L(\boldsymbol{S^i}; \theta)$ approaches the equivalent classical data likelihood $L(\boldsymbol{X^i}; \theta) = g^i_{\boldsymbol{X}^i}(\boldsymbol{X}^i; \theta)$ as the number of bins in the marginal histogram approaches infinity and the volume of each bin approaches zero. In particular, suppose for simplicity that the length $|\Upsilon^k_{\boldsymbol{b}}| = y^k_{b_k} - y^k_{b_k-1}$ of each univariate marginal bin $\Upsilon^k_{\boldsymbol{b}} = (y^k_{b_k-1}, y^k_{b_k}]$ is equal for each margin $k = 1, \ldots, K$, with fixed endpoints y^k_0 and $y^k_{B^k}$. Then as $B^k \to \infty$ the number of equally spaced bins grows, but their length $|\Upsilon^k_{\boldsymbol{b}}| \to 0$. Then

$$\lim_{\substack{B^k \to \infty \\ k=1,\ldots,K}} L(\boldsymbol{S^i}; \theta) = L(\boldsymbol{X^i}; \theta).$$

Intuitively in this setting, as the number of bins gets large and their volume reduces, in the limit almost all bins will be empty, with each observed datapoint x^i being contained in exactly one bin. For the symbolic likelihood (3.3), this means that empty bins ($\boldsymbol{s}^i_{\boldsymbol{b}^i} = 0$) will not contribute to the likelihood, and the N non-empty bins ($\boldsymbol{s}^i_{\boldsymbol{b}^i} = 1$) will contribute the term $g^i_{\boldsymbol{X}^i}(\boldsymbol{x^i}; \theta) = g_{X^i}(x^i; \theta)$, which is the equivalent term contributed to the classical likelihood function $L(\boldsymbol{X^i}; \theta)$.

As a result, this means that taking more bins will allow $L(\boldsymbol{S^i}; \theta)$, taken as an approximation to $L(\boldsymbol{X^i}; \theta)$, to approximate the classical data likelihood arbitrarily well. The difference is that the symbolic likelihood contains $B^1 \times \ldots \times B^K$ terms, which may be considerably less than the N terms of the classical data likelihood $L(\boldsymbol{X^i}; \theta) = \prod_{k=1}^N g_{X^i}(x^i_k; \theta)$ for large datasets. In this setting, the tradeoff of improved computational efficiency for some, perhaps small, approximation error may be attractive.

In particular, we may construct the log-likelihood function of a bivariate random marginal histogram $\boldsymbol{S}^{\boldsymbol{i}_2}$ by specifying the indices $\boldsymbol{i}_2 = (i_1, i_2)$, marginal bin indices $\boldsymbol{b}_2 = (b_{i_1}, b_{i_2})$ and number of bins $B^{i_1} \times B^{i_2}$, giving

$$\ell(\boldsymbol{S}^{\boldsymbol{i}_2}; \theta) \propto \sum_{b_{i_1}=1}^{B^{i_1}} \sum_{b_{i_2}=1}^{B^{i_2}} s^{\boldsymbol{i}_2}_{(b_{i_1}, b_{i_2})} \log P_{(b_{i_1}, b_{i_2})}(\theta). \tag{3.4}$$

Similarly, specifying $\boldsymbol{i} = (i_1, i_2, i_3)$ leads to the log-likelihood function of a trivariate random marginal histogram $\boldsymbol{S}^{\boldsymbol{i}_3}$ with $B^{i_1} \times B^{i_2} \times B^{i_3}$ bins indexed by $\boldsymbol{b}_3 = (b_{i_1}, b_{i_2}, b_{i_3})$, given by

$$\ell(\boldsymbol{S}^{\boldsymbol{i}_3}; \theta) \propto \sum_{b_{i_1}=1}^{B^{i_1}} \sum_{b_{i_2}=1}^{B^{i_2}} \sum_{b_{i_3}=1}^{B^{i_3}} s^{\boldsymbol{i}_3}_{(b_{i_1}, b_{i_2}, b_{i_3})} \log P_{(b_{i_1}, b_{i_2}, b_{i_3})}(\theta). \tag{3.5}$$

Clearly the number of terms in the full symbolic likelihood (3.2), $B^1 \times \ldots \times B^K$, increases exponentially as the dimension of the histogram, K, increases. This is further compounded since larger B^k, $k = 1, \ldots, K$, will produce a closer likelihood approximation $L(\boldsymbol{S}; \theta) \approx L(\boldsymbol{X}; \theta)$, which may be desirable. Similarly, the complexity of efficiently computing the K-dimensional integral $P_{\boldsymbol{b}}(\theta) = \int_{\Upsilon_{\boldsymbol{b}}} g_{\boldsymbol{X}}(\boldsymbol{z}; \theta) d\boldsymbol{z}$ also increases with K. Together this means that it may rapidly become practically infeasible to directly use the symbolic likelihood of Beranger et al. (2018) in more than, say, $K = 5$ or 6 dimensions, which reduces the applicability of this approach. However, the computational overheads of the bivariate and trivariate marginal histogram log-likelihoods (3.4) and (3.5) will be much lower. This motivates the use of composite likelihood techniques, constructed from marginal histograms $\boldsymbol{S}^{\boldsymbol{i}}$ of $\boldsymbol{S}$, which we now describe within the symbolic likelihood setting.

3.2.2 Composite likelihoods

Composite likelihoods, part of the family of pseudo-likelihood functions, are one practical technique for constructing asymptotically consistent likelihood-based parameter estimates when the standard likelihood function is computationally intractable (Lindsay, 1988, Varin et al., 2011). Such intractability can occur in many common modelling scenarios (Varin and Vidoni, 2005, Sisson et al., 2018). In particular, in Section 3.4 we examine max-stable process models for spatial extremes (Davison et al., 2012, Padoan et al., 2010), for which closed-form densities are available for models with $K = 2$ or 3 spatial locations, but not for the larger K required in practical applications, typically measured in the hundreds. See Section 3.4 for further details. Composite likelihoods are defined as the weighted product of conditional or marginal events of a process, each of which may be described by e.g. an ordinary likelihood function (Lindsay, 1988). If we assume all weights are equal for simplicity, a composite likelihood function can be expressed as $L_{CL}(\boldsymbol{x}; \theta) \propto \prod_{i=1}^{m} L_i(\boldsymbol{x}; \theta)$, where $L_i(\boldsymbol{x}; \theta)$ is the likelihood function of a conditional or marginal event of $\boldsymbol{x}$ for a given parameter vector θ.

A special case of the composite likelihood function is the j-wise composite likelihood function, comprising all j-dimensional marginal events. Using the same notation as in Section 3.2.1, and defining $\mathcal{I}_j = \{\boldsymbol{i} : \boldsymbol{i} \subseteq \{1, \ldots, K\}, |\boldsymbol{i}| = j\}$ to be the set of all j-dimensional subsets of $\{1, \ldots, K\}$, the j-wise composite likelihood function can be written as

$$L^{(j)}_{CL}(\boldsymbol{x}; \theta) \propto \prod_{\boldsymbol{i} \in \mathcal{I}_j} g^{\boldsymbol{i}}_{\boldsymbol{X}^{\boldsymbol{i}}}(\boldsymbol{x}^{\boldsymbol{i}}; \theta), \tag{3.6}$$

where, as before, $g^{\boldsymbol{i}}$ represents the j-dimensional (marginal) density associated with the j-wise

event $\boldsymbol{i} \in \mathcal{I}_j$. In analogy with (3.4) and (3.5), when $j = 2$ the pairwise composite log-likelihood function, $\ell_{CL}^{(2)}$, is given by

$$\ell_{CL}^{(2)}(\boldsymbol{x};\theta) \propto \sum_{i_1=1}^{K-1} \sum_{i_2=i_1+1}^{K} \log g_{\boldsymbol{X}^{i_1},\boldsymbol{X}^{i_2}}(\boldsymbol{x}^{i_1},\boldsymbol{x}^{i_2};\theta), \tag{3.7}$$

and similarly for $j = 3$, the triple-wise composite log-likelihood, $\ell_{CL}^{(3)}$, is given by

$$\ell_{CL}^{(3)}(\boldsymbol{x};\theta) \propto \sum_{i_1=1}^{K-2} \sum_{i_2=i_1+1}^{K-1} \sum_{i_3=i_2+1}^{K} \log g_{\boldsymbol{X}^{i_1},\boldsymbol{X}^{i_2},\boldsymbol{X}^{i_3}}(\boldsymbol{x}^{i_1},\boldsymbol{x}^{i_2},\boldsymbol{x}^{i_3};\theta).$$

Taking first order partial derivatives of $\ell_{CL}^{(j)}(\boldsymbol{x};\theta)$ with respect to θ yields the composite score function $\nabla\ell_{CL}^{(j)}(\theta;\boldsymbol{x})$, and taking second order partial derivatives gives the Hessian matrix $\nabla^2\ell_{CL}^{(j)}(\theta;\boldsymbol{x})$. Lindsay (1988) showed that the resulting maximum j-wise composite likelihood estimator, $\hat{\theta}_{CL}^{(j)}$, is asymptotically consistent and distributed as

$$\sqrt{N}\left(\hat{\theta}_{CL}^{(j)} - \theta\right) \to N\left(0, G^{(j)}(\theta)^{-1}\right),$$

where $G^{(j)}$ is the (j-wise) Godambe information matrix (Godambe, 1960) defined by $G^{(j)}(\theta) = H^{(j)}(\theta)J^{(j)}(\theta)^{-1}H^{(j)}(\theta)$, where $H^{(j)}(\theta) = -\mathbb{E}_g(\nabla^2\ell_{CL}^{(j)}(\theta;\boldsymbol{x}))$ and $J^{(j)}(\theta) = \mathbb{V}_g(\nabla\ell_{CL}^{(j)}(\theta;\boldsymbol{x}))$ are respectively the sensitivity and variability matrices. For standard likelihoods we have $j = K$ and $\mathcal{I} = \{(1,\ldots,K)\}$, and so dropping the superscripts, $H(\theta) = J(\theta)$ and the Godambe information matrix reduces to $G(\theta) = H(\theta) = I(\theta)$, where $I(\theta)$ is the Fisher information matrix. The above result shows that the composite MLE is asymptotically unbiased, however it is worth noting that $G(\theta)^{-1}$ often does not attain the Cramer-Rao lower bound and subsequently there is a decrease in efficiency when the the composite MLE is used in the place of the standard MLE (Varin et al., 2011).

3.3 Composite likelihood functions for histogram-valued data

In this Section we introduce a composite likelihood function for random histograms that is constructed using sets of marginal histograms. We will first present the main result, before examining the consistency and variability of the symbolic composite MLE in turn, as the form of each of these has interesting implications for statistical inference using random histograms.

3.3.1 Composite likelihood function

Suppose that we observe T independent replicates, $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_T$, of the random variable $\boldsymbol{X} = (X_1,\ldots,X_N) \in \mathbb{R}^{K\times N}$ over some index variable $t = 1,\ldots,T$ (e.g. time), and denote the realised values as $\boldsymbol{x}_t$. For each $\boldsymbol{X}_t$, $t = 1,\ldots,T$, we may construct a K-dimensional random histogram $\boldsymbol{S}_t$ over the set of bins $\{\boldsymbol{1},\ldots,\boldsymbol{B}\}$. A j-dimensional marginal histogram of $\boldsymbol{S}_t$ may then be constructed as $\boldsymbol{S}_t^{\boldsymbol{i}}$, where $\boldsymbol{i} \in \mathcal{I}_j$. For a given model $g_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^{N} g_X(x_i;\theta)$ for the

micro-data $\boldsymbol{X}_t$, the likelihood of the marginal histogram $\boldsymbol{S}_t^{\boldsymbol{i}}$ is then given by $L(\boldsymbol{S}_t^{\boldsymbol{i}};\theta)$ in (3.3). We can now define the j-wise symbolic composite likelihood for all j-dimensional marginal histograms $\boldsymbol{S}_t^{\boldsymbol{i}}$ of $\boldsymbol{S}_t$, $\boldsymbol{i} \in \mathcal{I}_j$, $t = 1, \ldots, T$ as follows.

Proposition 3.3.1. *Writing $\boldsymbol{S}_{1:T} = (\boldsymbol{S}_1, \ldots, \boldsymbol{S}_T)$ as the collection of K-dimensional histograms, the j-wise symbolic composite likelihood for $\boldsymbol{S}_{1:T}$ is given by*

$$L_{SCL}^{(j)}(\boldsymbol{S}_{1:T};\theta) = \prod_{t=1}^{T} \prod_{\boldsymbol{i}\in\mathcal{I}_j} L(\boldsymbol{S}_t^{\boldsymbol{i}};\theta), \tag{3.8}$$

where $L(\boldsymbol{S}_t^{\boldsymbol{i}};\theta)$ is defined in (3.3). Defining the maximum j-wise symbolic composite likelihood estimator as $\hat{\theta}_{SCL}^{(j)} = \arg\max_\theta L_{SCL}^{(j)}(\boldsymbol{S}_{1:T};\theta)$, following standard composite likelihood construction arguments (Lindsay, 1988) we have

$$\sqrt{N}\left(\hat{\theta}_{SCL}^{(j)} - \theta\right) \to \mathcal{N}\left(0, G^{(j)}(\theta)^{-1}\right),$$

as $N \to \infty$ where $G^{(j)}(\theta) = H^{(j)}(\theta)J^{(j)}(\theta)^{-1}H^{(j)}(\theta)$, and where estimates of the sensitivity and variability matrices are given by

$$\hat{H}(\hat{\theta}_{SCL}^{(j)}) = -\sum_{t=1}^{T}\sum_{\boldsymbol{i}\in\mathcal{I}_j} \nabla^2 \ell(\boldsymbol{S}_t^{\boldsymbol{i}};\theta) = -\sum_{t=1}^{T}\sum_{\boldsymbol{i}\in\mathcal{I}_j}\sum_{\boldsymbol{b}^{\boldsymbol{i}}=\boldsymbol{1}^{\boldsymbol{i}}}^{\boldsymbol{B}^{\boldsymbol{i}}} s_{t,\boldsymbol{b}^{\boldsymbol{i}}}^{\boldsymbol{i}} \nabla^2 \log P_{t,\boldsymbol{b}^{\boldsymbol{i}}}(\hat{\theta}_{SCL}^{(j)}) \tag{3.9}$$

$$\begin{aligned}\hat{J}(\hat{\theta}_{SCL}^{(j)}) &= \sum_{t=1}^{T}\left(\sum_{\boldsymbol{i}\in\mathcal{I}_j}\nabla\ell(\boldsymbol{S}_t^{\boldsymbol{i}};\theta)\right)\left(\sum_{\boldsymbol{i}\in\mathcal{I}_j}\nabla\ell(\boldsymbol{S}_t^{\boldsymbol{i}};\theta)\right)^\top \\ &= \sum_{t=1}^{T}\left(\sum_{\boldsymbol{i}\in\mathcal{I}_j}\sum_{\boldsymbol{b}^{\boldsymbol{i}}=\boldsymbol{1}^{\boldsymbol{i}}}^{\boldsymbol{B}^{\boldsymbol{i}}} s_{t,\boldsymbol{b}^{\boldsymbol{i}}}^{\boldsymbol{i}} \nabla \log P_{t,\boldsymbol{b}^{\boldsymbol{i}}}(\hat{\theta}_{SCL}^{(j)})\right)\left(\sum_{\boldsymbol{i}\in\mathcal{I}_j}\sum_{\boldsymbol{b}^{\boldsymbol{i}}=\boldsymbol{1}^{\boldsymbol{i}}}^{\boldsymbol{B}^{\boldsymbol{i}}} s_{t,\boldsymbol{b}^{\boldsymbol{i}}}^{\boldsymbol{i}} \nabla \log P_{t,\boldsymbol{b}^{\boldsymbol{i}}}(\hat{\theta}_{SCL}^{(j)})\right)^\top,\end{aligned} \tag{3.10}$$

where t subscripts indicate dependence on $\boldsymbol{S}_t$.

For example, the pairwise ($j = 2$) symbolic composite log-likelihood function is given by

$$\ell_{SCL}^{(2)}(\boldsymbol{S}_{1:T};\theta) = \sum_{t=1}^{T}\sum_{i_1=1}^{K-1}\sum_{i_2=i_1+1}^{K} \ell(\boldsymbol{S}_t^{(i_1,i_2)};\theta) \tag{3.11}$$

where $\ell(\boldsymbol{S}_t^{(i_1,i_2)};\theta)$ is given by (3.4), and the triple-wise ($j = 3$) symbolic composite log-likelihood function is given by

$$\ell_{SCL}^{(3)}(\boldsymbol{S}_{1:T};\theta) = \sum_{t=1}^{T}\sum_{i_1=1}^{K-2}\sum_{i_2=i_1+1}^{K-1}\sum_{i_3=i_2+1}^{K} \ell(\boldsymbol{S}_t^{(i_1,i_2,i_3)};\theta) \tag{3.12}$$

where $\ell(\boldsymbol{S}_t^{(i_1,i_2,i_3)};\theta)$ is given by (3.5).

3.3.2 Symbolic composite maximum likelihood estimator consistency

It is straightforward to show that the j-wise symbolic composite likelihood estimator $\hat{\theta}^{(j)}_{SCL}$ that maximises (3.8) is consistent with the equivalent composite likelihood estimator $\hat{\theta}^{(j)}_{CL}$ that maximises $L^{(j)}_{CL}(\boldsymbol{X}_{1:T};\theta) = \prod_{t=1}^{T} L^{(j)}_{CL}(\boldsymbol{X}_t;\theta)$ where $L^{(j)}_{CL}(\boldsymbol{X}_t;\theta)$ is given by (3.6) as the number of bins in each marginal histogram approaches infinity and the volume of each bin approaches zero.

We show this by extending the univariate proof described by Zhang (2017) to (w.l.o.g) the bivariate ($j = 2$) setting, from which the extension to the K-dimensional case is immediate.

Consider the pairwise composite log likelihood given in (3.11). In this case, for $\boldsymbol{i} = (i_1, i_2) \in \mathcal{I}_2$, and for any $t = 1, \ldots, T$ (although dropping the subscript t for clarity), the probability that a bivariate micro-data observation $X^{\boldsymbol{i}} \in \mathbb{R}^2$ falls in marginal bin $\boldsymbol{b^i} = (b_{i_1}, b_{i_2})$ over the region $(y^{i_1}_{b_{i_1}-1}, y^{i_1}_{b_{i_1}}] \times (y^{i_2}_{b_{i_2}-1}, y^{i_2}_{b_{i_2}}]$ is

$$P_{\boldsymbol{b^i}}(\theta) = G^{\boldsymbol{i}}_{X^{\boldsymbol{i}}}(y^{i_1}_{b_{i_1}}, y^{i_2}_{b_{i_2}};\theta) - G^{\boldsymbol{i}}_{X^{\boldsymbol{i}}}(y^{i_1}_{b_{i_1}-1}, y^{i_2}_{b_{i_2}};\theta) - G^{\boldsymbol{i}}_{X^{\boldsymbol{i}}}(y^{i_1}_{b_{i_1}}, y^{i_2}_{b_{i_2}-1};\theta) + G^{\boldsymbol{i}}_{X^{\boldsymbol{i}}}(y^{i_1}_{b_{i_1}-1}, y^{i_2}_{b_{i_2}-1};\theta),$$

where $G_X(x;\theta)$ is the distribution function of $g_X(x;\theta)$.

Fixing the i_2 margin, by the mean value theorem there exists a $\tilde{x}_{b_{i_1}} \in (y^{i_1}_{b_{i_1}-1}, y^{i_1}_{b_{i_1}}]$ such that

$$P_{\boldsymbol{b^i}}(\theta) = (y^{i_1}_{b_{i_1}} - y^{i_1}_{b_{i_1}-1})\frac{d}{dx_1}G^{\boldsymbol{i}}_{X^{\boldsymbol{i}}}(\tilde{x}_{b_{i_1}}, y^{i_2}_{b_{i_2}};\theta) - (y^{i_1}_{b_{i_1}} - y^{i_1}_{b_{i_1}-1})\frac{d}{dx_1}G^{\boldsymbol{i}}_{X^{\boldsymbol{i}}}(\tilde{x}_{b_{i_1}}, y^{i_2}_{b_{i_2}-1};\theta),$$

where $\frac{d}{dx_k}G_X$ denotes differentiation with respect to the k-th component of G_X. Similarly fixing the i_1 margin, again by the mean value theorem there exists a $\tilde{x}_{b_{i_2}} \in (y^{i_2}_{b_{i_2}-1}, y^{i_2}_{b_{i_2}}]$ such that

$$\begin{aligned} P_{\boldsymbol{b^i}}(\theta) &= (y^{i_1}_{b_{i_1}} - y^{i_1}_{b_{i_1}-1})\frac{d}{dx_1}[G_X(\tilde{x}_{b_{i_1}}, y^{i_2}_{b_{i_2}};\theta) - G_X(\tilde{x}_{i_{b_1}}, y^{i_2}_{b_{i_2}-1};\theta)] \\ &= (y^{i_1}_{b_{i_1}} - y^{i_1}_{b_{i_1}-1})(y^{i_2}_{b_{i_2}} - y^{i_2}_{b_{i_2}-1})\frac{d}{dx_1}\frac{d}{dx_2}G_X(\tilde{x}_{b_{i_1}}, \tilde{x}_{b_{i_2}};\theta) \\ &\propto \frac{d}{dx_1}\frac{d}{dx_2}G_X(\tilde{x}_{b_{i_1}}, \tilde{x}_{b_{i_2}};\theta) = g_X(\tilde{x}_{b_{i_1}}, \tilde{x}_{b_{i_2}};\theta). \end{aligned}$$

This allows the pairwise symbolic composite log likelihood to be written as

$$\ell^{(2)}_{SCL}(\boldsymbol{S}_{1:T};\theta) \propto \sum_{t=1}^{T}\sum_{i_1=1}^{K-1}\sum_{i_2=i_1+1}^{K}\sum_{b_{i_1}=1}^{B^{i_1}}\sum_{b_{i_2}=1}^{B^{i_2}} \boldsymbol{s^i}_{(b_{i_1},b_{i_2})} \log g^{\boldsymbol{i}}_{X^{\boldsymbol{i}}}(\tilde{x}_{b_{i_1}}, \tilde{x}_{b_{i_2}};\theta).$$

Now, letting the number of bins $B^{i_1}, B^{i_2} \to \infty$ such that each bin's volume $\to 0$ means that in the limit each bin will either contain zero ($s^{\boldsymbol{i}}_{(b_{i_1},b_{i_2})} = 0$) or, assuming continuous data, exactly one observation ($s^{\boldsymbol{i}}_{(b_{i_1},b_{i_2})} = 1$). In the case where a bin contains exactly one observation, the m-th observed classical datapoint (x_{m,i_1}, x_{m,i_2}), we have $(y^{i_1}_{b_{i_1}-1}, y^{i_1}_{b_{i_1}}] \times (y^{i_2}_{b_{i_2}-1}, y^{i_2}_{b_{i_2}}] \to (x_{m,i_1}, x_{m,i_2})$.

Hence $(\tilde{x}_{b_{i_1}}, \tilde{x}_{b_{i_2}}) \to (x_{m,i_1}, x_{m,i_2})$ and so

$$\ell^{(2)}_{SCL}(\boldsymbol{S}_{1:T};\theta) \;\to\; \sum_{t=1}^{T}\sum_{i_1=1}^{K-1}\sum_{i_2=i_1+1}^{K}\sum_{m=1}^{N} \log g^{\boldsymbol{i}}_{X^{\boldsymbol{i}}}(x_{m,b_{i_1}}, x_{m,b_{i_2}};\theta),$$

which is has a maximum at $\hat{\theta}^{(2)}_{CL}$. This argument straightforwardly extends to the j-wise symbolic composite likelihood by iterated use of the mean value theorem.

This result means that the symbolic composite likelihood can be considered an asymptotically consistent approximation of the standard composite likelihood, which yields an approximation of the standard MLE. As a consequence, this approach can be considered 'an approximation of an approximation', with the limiting case with increasing B (and reduction in bin volume) resulting in the standard composite MLE, and this, simultaneously with increasing N yields consistent estimates to the true value θ. The approximation can be arbitrarily close to the classical composite equivalent (though at the cost of increasing computational overheads) as the number of bins increases.

There are a number of specifications under which the T random histograms may be constructed from the underlying micro-data (and the details of these are encoded in the parameter ϕ in (3.1)). These specifications control the location and sizes of the bins in each random histogram, and the number of random histograms, T, itself. While we do not discuss the merits of particular constructions here, we note that the above asymptotic consistency result for the symbolic composite log likelihood holds regardless of the method of bin construction in each histogram (as long as the volume of each bin approaches zero as the number of bins approaches infinity), and regardless of the number of random histograms, T (as long as the underlying micro-data $X_1, \ldots, X_N$ are stationary). Consistency also holds for different numbers of micro-data encoded in each random histogram $\boldsymbol{S}_t$ as long as there is sufficient data in enough unique bins that $\ell(\boldsymbol{S}^{\boldsymbol{i}}_t;\theta)$ is well defined and satisfies the usual regularity conditions.

In particular, if each random histogram has exactly the same bins, so that $y^k_{t,b_k} = y^k_{b_k}$ for all $t = 1, \ldots, T$, then the choice of T has no effect on the symbolic composite maximum likelihood estimator. That is, $\hat{\theta}_{SCL}$ takes the same value independently of the number of random histograms T. This is easily seen as

$$\sum_{t=1}^{T} s^{\boldsymbol{i}}_{t,\boldsymbol{b}^{\boldsymbol{i}}} = s^{\boldsymbol{i}}_{\boldsymbol{b}^{\boldsymbol{i}}},\ \forall \boldsymbol{b}, \boldsymbol{i}, \tag{3.13}$$

where $s^{\boldsymbol{i}}_{\boldsymbol{b}^{\boldsymbol{i}}}$ is the count of all micro-data falling in (marginal) bin $\boldsymbol{b}^{\boldsymbol{i}}$ when all data are allocated to a single ($T = 1$) histogram. As a result, we then have

$$\sum_{t=1}^{T}\sum_{\boldsymbol{i}\in\mathcal{I}_j}\sum_{\boldsymbol{b}^{\boldsymbol{i}}=\boldsymbol{1}^{\boldsymbol{i}}}^{\boldsymbol{B}^{\boldsymbol{i}}} s^{\boldsymbol{i}}_{t,\boldsymbol{b}^{\boldsymbol{i}}} \log P_{t,\boldsymbol{b}^{\boldsymbol{i}}}(\theta) = \sum_{\boldsymbol{i}\in\mathcal{I}_j}\sum_{\boldsymbol{b}^{\boldsymbol{i}}=\boldsymbol{1}^{\boldsymbol{i}}}^{\boldsymbol{B}^{\boldsymbol{i}}} s^{\boldsymbol{i}}_{\boldsymbol{b}^{\boldsymbol{i}}} \log P_{\boldsymbol{b}^{\boldsymbol{i}}}(\theta),$$

and so the resulting symbolic composite maximum likelihood estimators are equivalent. As a result, if primary interest of an analysis is of fast computation of $\hat{\theta}_{SCL}$, then the optimal choice is by constructing $T = 1$ random histograms, as this will allow for the fastest optimisation of

$\ell^{(j)}_{SCL}(\boldsymbol{S}_{1:T};\theta)$. (Note that if all bins are equal, then this single histogram can be created by simply summing the counts in each bin, following (3.13).) However, $T = 1$ will not be the optimal choice if interest is also in computing $\text{Var}(\hat{\theta}^{(j)}_{SCL})$ – see the following Section.

3.3.3 Variance consistency

We now show the conditions under which the symbolic Godambe information matrix $G(\hat{\theta}^{(j)}_{SCL})$ converges to the standard Godambe matrix $G(\hat{\theta}^{(j)}_{CL})$. In particular, we will show that as the number of equally spaced histogram bins becomes large (so that $B^k \to \infty$ for $k = 1, \ldots, K$) while the volume of each bin approaches zero ($|\Upsilon_{\boldsymbol{b}}| \to 0$, $\forall \boldsymbol{b}$), and as the number of histograms $T \to N$ so that each histogram contains exactly one micro-data observation, then

$$\lim_{T\to N} \lim_{\substack{B^k\to\infty \\ k=1,\ldots,K}} \text{Var}(\hat{\theta}^{(j)}_{SCL}) = \text{Var}(\hat{\theta}^{(j)}_{CL}).$$

Following the same arguments as in Section 3.3.2 it is straightforward to show that

$$\lim_{\substack{B^k\to\infty \\ k=1,\ldots,K}} \hat{H}(\hat{\theta}^{(j)}_{SCL}) = \hat{H}(\hat{\theta}^{(j)}_{CL}),$$

so that the symbolic Hessian matrix converges to the standard composite likelihood Hessian matrix, regardless of the number of histograms, T, due to the additive form of (3.9). Numerical estimates of $\hat{H}(\hat{\theta}^{(j)}_{SCL})$ can be obtained through numerical methods during maximum likelihood estimation (e.g. using the `optim` function in `R`).

The natural estimator for the variability matrix is the empirical variance estimator (3.10). With increasing T, the sum of the counts in each histogram $\boldsymbol{S}_t$ decreases in magnitude until there is exactly 1 non-empty bin with count 1 in each of $T = N$ marginal histograms. At this point

$$\sum_{\boldsymbol{b}^{\boldsymbol{i}}=\boldsymbol{1}^{\boldsymbol{i}}}^{\boldsymbol{B}^{\boldsymbol{i}}} s^{\boldsymbol{i}}_{t,\boldsymbol{b}^{\boldsymbol{i}}} = 1, \qquad \forall \boldsymbol{i} \in \mathcal{I}_j,\ t = 1, \ldots, N.$$

As a result, the limit of the symbolic composite log-likelihood function, as $T \to N$, is

$$\lim_{T\to N} \ell^{(j)}_{SCL}(\boldsymbol{S}_{1:T};\theta) \propto \lim_{T\to N} \sum_{t=1}^{T} \sum_{\boldsymbol{i}\in\mathcal{I}_j} \sum_{\boldsymbol{b}^{\boldsymbol{i}}=\boldsymbol{1}^{\boldsymbol{i}}}^{\boldsymbol{B}^{\boldsymbol{i}}} \boldsymbol{s}^{\boldsymbol{i}}_{t,\boldsymbol{b}^{\boldsymbol{i}}} \log P_{t,\boldsymbol{b}^{\boldsymbol{i}}}(\theta) = \sum_{t=1}^{N} \sum_{\boldsymbol{i}\in\mathcal{I}_j} \log P_{t,\boldsymbol{b}^{(t)\boldsymbol{i}}}(\theta),$$

where $\boldsymbol{b}^{(t)}$ denotes the bin which contains the single micro-data observation x_t in histogram $\boldsymbol{S}_t$. Because

$$\lim_{\substack{B^k\to\infty \\ k=1,\ldots,K}} \log P_{t,\boldsymbol{b}^{(t)\boldsymbol{i}}}(\theta) = \log g^{\boldsymbol{i}}_{X^{\boldsymbol{i}}}(x^{\boldsymbol{i}}_t;\theta)$$

reduces to the standard composite likelihood marginal event component as the histogram bins

reduce in size, then $\lim_{\substack{B^k\to\infty\\k=1,\dots,K}} \hat{\theta}^{(j)}_{SCL} = \hat{\theta}^{(j)}_{CL}$. It then follows that from (3.10)

$$\begin{aligned}
\lim_{T\to N}\lim_{\substack{B^k\to\infty\\k=1,\dots,K}} \hat{J}(\hat{\theta}^{(j)}_{SCL}) &= \lim_{\substack{B^k\to\infty\\k=1,\dots,K}} \sum_{t=1}^{N}\left(\sum_{\boldsymbol{i}\in\mathcal{I}_j}\nabla P_{t,\boldsymbol{b}^{(t)}\boldsymbol{i}}(\hat{\theta}^{(j)}_{SCL})\right)\left(\sum_{\boldsymbol{i}\in\mathcal{I}_j}\nabla P_{t,\boldsymbol{b}^{(t)}\boldsymbol{i}}(\hat{\theta}^{(j)}_{SCL})\right)^{\top}\\
&= \lim_{\substack{B^k\to\infty\\k=1,\dots,K}} \sum_{t=1}^{N}\left(\sum_{\boldsymbol{i}\in\mathcal{I}_j}\nabla g^{\boldsymbol{i}}_{X^{\boldsymbol{i}}}(x^{\boldsymbol{i}}_t;\hat{\theta}^{(j)}_{CL})\right)\left(\sum_{\boldsymbol{i}\in\mathcal{I}_j}\nabla g^{\boldsymbol{i}}_{X^{\boldsymbol{i}}}(x^{\boldsymbol{i}}_t;\hat{\theta}^{(j)}_{CL})\right)^{\top}\\
&= \hat{J}(\hat{\theta}^{(j)}_{CL}).
\end{aligned}$$

Convergence of the symbolic Godambe information matrix $G(\hat{\theta}^{(j)}_{SCL})$ to the standard Godambe matrix $G(\hat{\theta}^{(j)}_{CL})$ then follows under these limit conditions.

While the above result confirms that the limiting behaviour of $\hat{\theta}^{(j)}_{SCL}$ is the same as $\hat{\theta}^{(j)}_{CL}$, in particular as $T \to N$, in practice we may prefer to have less than N random histograms for a given analysis, particularly if N is very large. In this setting, for a fixed $T < N$ we then have

$$\lim_{\substack{B^k\to\infty\\k=1,\dots,K}} \hat{J}(\hat{\theta}^{(j)}_{SCL}) = \sum_{t=1}^{T}\left(\sum_{\boldsymbol{i}\in\mathcal{I}_j}\nabla g^{\boldsymbol{i}}_{\boldsymbol{X}^{\boldsymbol{i}}}(\boldsymbol{x}^{\boldsymbol{i}}_t;\hat{\theta}^{(j)}_{SCL})\right)\left(\sum_{\boldsymbol{i}\in\mathcal{I}_j}\nabla g^{\boldsymbol{i}}_{\boldsymbol{X}^{\boldsymbol{i}}}(\boldsymbol{x}^{\boldsymbol{i}}_t;\hat{\theta}^{(j)}_{SCL})\right)^{\top} \tag{3.14}$$

using similar arguments to the above.

Compared to the standard composite likelihood sensitivity matrix $\hat{J}(\hat{\theta}^{(j)}_{CL})$, (3.14) can be interpreted as the sensitivity matrix for a classical (micro-data) dataset where some temporal information is lost. That is, we know which time block (histogram) $t = 1, ..., T$ each observation came from, but not specifically when each observation occurred within that block. As a result the variability of $\hat{\theta}^{(j)}_{SCL}$ will always be larger for a smaller number of time blocks. As T increases, more temporal information is retained as each time block then decreases in size. This leads to more precise knowledge about when each data point may have been observed, and accordingly leading to a reduction in the variance of $\hat{\theta}^{(j)}_{SCL}$. The standard composite likelihood case is recovered for $T = N$ when the time of each datapoint is known exactly.

Equation (3.14) thereby characterises the loss in precision for the standard composite MLE as temporal information is lost. It also characterises the limiting performance (in the sense of $B^k \to \infty$, $\forall k$) of the symbolic composite MLE. (This relationship is explored explicitly in Section 3.4.2.) However the advantage of working with $\hat{\theta}^{(j)}_{SCL}$ is that the likelihood function is typically more computationally efficient to evaluate for large N. As such, estimating $\text{Var}(\hat{\theta}^{(j)}_{SCL})$ represents a trade-off between greater precision (larger T) and greater computational and data storage efficiency (smaller T).

In practice, the analyst would choose T as small as possible such that the inferential goals (perhaps depending on confidence intervals of model parameters) are still viable, in order to maximise overall analysis efficiency. Recall that, as discussed in Section 3.3.2, if all histogram bins are equal, computation of the symbolic composite MLE itself can be achieved at low cost by combining all histograms into a single histogram ($T = 1$). So the main impact of the number

of histograms is on the variability of the symbolic composite MLE. If the underlying data is available, then $T = N$ can be used to determine the lowest possible variance of the obtained estimates, as the computational cost of evaluating the variance, given the estimates, is only a small proportion of the total computation when compared to the cost of evaluating the estimates via the optimisation of the likelihood.

Increasing values of B can be investigated sequentially in order to determine the point at which comparable estimates and standard errors to the classical composite likelihood function are obtained. For the value B at which the change in results compared to the previously investigated value is negligible, the practitioner can be confident that further increasing the number of bins will not significantly improve the analysis, although it will increase the computational burden. Although this approach requires the optimisation of multiple symbolic composite likelihood estimators $\hat{\theta}^{(j)}_{SCL}$ for varying values of B, the simulations in the following section will demonstrate that this is still more computationally efficient than the existing classical analysis for large datasets, due to the large computational gains associated with employing a symbolic approach. This simple approach is utilised in both the simulation studies in Section 3.4, and the real data analysis in Section 3.5.

3.4 Simulation studies

We now examine the performance of the symbolic composite maximum likelihood estimator within the context of our motivating application – modelling spatial extremes using max-stable processes. We first briefly introduce these, before comparing $\hat{\theta}^{(j)}_{SCL}$ to standard composite likelihoods in accuracy, precision and efficiency under a range of modelling scenarios.

3.4.1 Max-stable process models

Jenkinson (1955) first proposed a limiting distribution for modelling datasets comprising of block maxima. Suppose $X_1, \ldots, X_n \in D$, in some continuous space D, are i.i.d. univariate random variables with distribution function F, and $M_n = \max\{X_1, \ldots, X_n\}$. If there exist constants $a_n > 0$, $b_n \in \mathbb{R}$ such that

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x),$$

is non-degenerate, for all $x \in D$, then G is a member of the generalised extreme value (GEV) family whose distribution function is given by $G(x; \mu, \sigma, \xi) = \exp\{-v(x; \mu, \sigma, \xi)\}$, where $\mu \in \mathbb{R}$, $\sigma > 0$, $\xi \in \mathbb{R}$, $v(y; \mu, \sigma, \xi) = \left(1+\xi\frac{y-\mu}{\sigma}\right)_+^{-\frac{1}{\xi}}$ when $\xi \neq 0$ and $e^{-\frac{y-\mu}{\sigma}}$ otherwise, and $a_+ = \min\{0, a\}$.

Max-stable processes (de Haan, 1984, Resnick, 1987, de Haan and Ferreira, 2006) are a popular tool to model spatial extremes. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. copies of a stochastic process $\{X(t) : t \in \mathcal{T}\}$ over some space $\mathcal{T}$. If continuous functions $a_n(t) > 0$,

$b_n(t) \in \mathbb{R}$ exist such that

$$\lim_{n\to\infty} \frac{\max_{i=1,\ldots,n} X_i(t) - b_n(t)}{a_n(t)} = Y(t)$$

is non-degenerate, then $Y(t)$ is a max-stable process. Spectral representations (de Haan, 1984, Schlather, 2002) allow to define max-stable models for $Y(t)$ such as the flexible extremal skew-t (Beranger et al., 2017) and its particular cases. Here we select the Gaussian max-stable process (Smith, 1990), one of the simplest parametric models. Genton et al. (2011) derived the joint distribution function of this model for $K \geq 2$ spatial locations with coordinates $t_k \in \mathcal{T} = \mathbb{R}^d$, $k = 1, \ldots, K$, where $K \leq d + 1$. Let $\tilde{T} = (t_1, \ldots, t_K) \in \mathbb{R}^{d\times K}$ be the matrix of coordinates for the locations, and $\tilde{T}_{-k}$ be the matrix $\tilde{T}$ without the k^{th} column, $k = 1, \ldots, K$. Also let $\mathbf{v} = (v_1, \ldots, v_K)^\top \in \mathbb{R}_+^K$ and $c^{(j)}(\mathbf{v}) = \left(c_1^{(j)}(\mathbf{v}), \ldots, c_{j-1}^{(j)}(\mathbf{v}), c_{j+1}^{(j)}(\mathbf{v}), \ldots, c_K^{(j)}(\mathbf{v})\right)^\top \in \mathbb{R}^{K-1}$, where, for $k = 1, \ldots, K$, $v_k = v(y_k; \mu, \sigma, \xi)^{-1}$ and $c_k^{(j)}(\mathbf{v}) = (t_j - t_k)^\top \Sigma^{-1}(t_j - t_k)/2 - \log\left(\frac{v_j}{v_k}\right)$. Then, writing $\Sigma^{(j)} = (t_j 1_{K-1}^\top - \tilde{T}_{-j})^\top \Sigma^{-1}(t_j 1_{K-1}^\top - \tilde{T}_{-j})$, where $1_d = (1, \ldots, 1)^\top \in \mathbb{R}^d$, the distribution function of the Gaussian max-stable process model can be written as

$$P(Y_1(t) \leq y_1, \ldots, Y_K(t) \leq y_K) = \exp\left\{-\sum_{j=1}^{K} \frac{1}{v_j} \Phi_{K-1}\left(c^{(j)}(\mathbf{v}); \Sigma^{(j)}\right)\right\}, \tag{3.15}$$

where $\Phi_d(\,\cdot\,; \Sigma)$ is the $d-$dimensional zero-mean Gaussian distribution function with covariance matrix Σ. Each univariate margin of this process is a GEV distribution. The parameters for this model are the spatial covariance matrix $\Sigma = [\sigma_{ij}]$ and the marginal GEV parameters μ, σ, ξ.

For typical spatial problems the number of spatial locations K is in the order of hundreds. We use $K \sim 100$ in some of the below simulations and the future-simulation climate data analysis in Section 3.5. However, for a $d = 2$ dimensional surface, (3.15) is only valid for $K = 2$ or 3 locations, and for other constructions of max-stable models the distribution function becomes rapidly intractable for more than a handful of spatial locations. For this reason, composite likelihood techniques are attractive in practice.

In the following we compare the performance of both symbolic composite and standard composite likelihood MLEs ($\hat{\theta}_{SCL}^{(j)}$ and $\hat{\theta}_{CL}^{(j)}$ respectively) in scenarios following those in Padoan et al. (2010) and Genton et al. (2011), where $\theta = (\sigma_{11}, \sigma_{12}, \sigma_{22}, \mu, \sigma, \xi)$.

For each experiment, K locations are generated uniformly over the space $\mathcal{T} = [0, 40] \times [0, 40]$ ($d = 2$). For each location, N realisations are generated from the Gaussian max-stable model using the `R` package `SpatialExtremes` (Ribatet, 2015) with standard Gumbel margins (i.e. $(\mu, \sigma, \xi) = (0, 1, 0)$).

3.4.2 Comparisons with composite likelihoods

Varying the number of bins, B

We generate $N = 1\,000$ realisations for $K = 15$ locations and 5 different configurations of the covariance matrix Σ, with true values given in Table 3.1, which represent a range of dependence

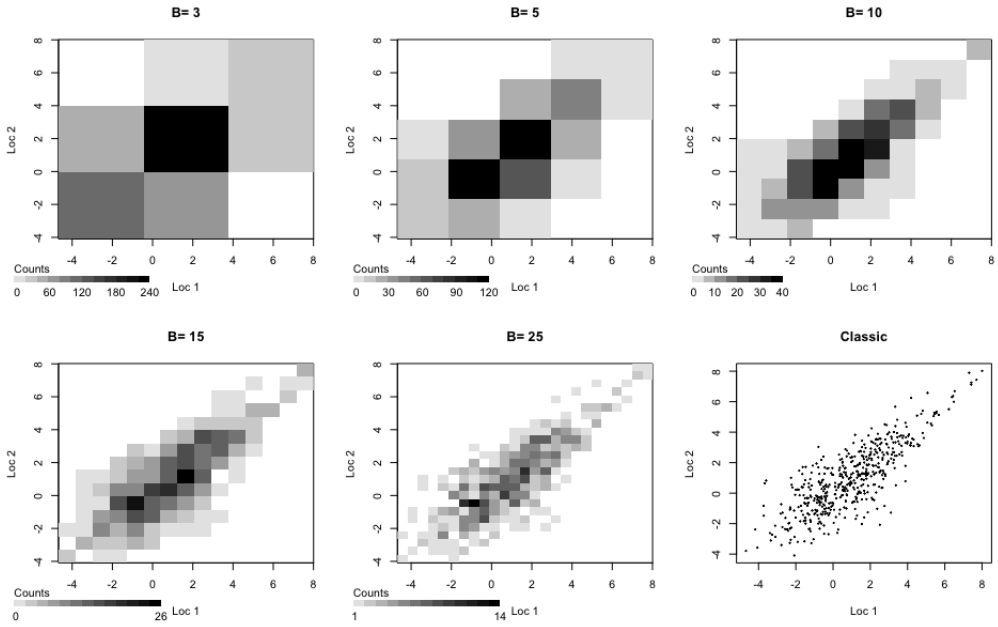

Figure 3.1: $B \times B$ bivariate histograms for different values of B for the same classical dataset (bottom right panel) of size $N = 1\,000$, generated at two spatial locations under the Gaussian max-stable model with $\Sigma = \Sigma_3$ (Table 3.1).

Model	σ_{11}	σ_{12}	σ_{22}
Σ_1	300	0	300
Σ_2	300	150	300
Σ_3	300	150	200
Σ_4	3 000	1 500	3 000
Σ_5	30	15	30

Table 3.1: Spatial dependence parameter specifications for the Gaussian max-stable model, following Padoan et al. (2010).

scenarios. For each dataset a single histogram $\boldsymbol{S}$ ($T = 1$) is 'constructed', although in practice we only construct all histograms $\boldsymbol{S}^{\boldsymbol{i}}, \boldsymbol{i} \in \mathcal{I}_2$ for each pair of spatial locations. The number of bins is constant in each dimension $B^k = B$, $k = 1, \ldots, K$, and we specify $B = 2, 3, 5, 10, 15$ and 25. Figure 3.1 shows the resulting bivariate histograms for two locations with $\Sigma = \Sigma_3$.

Table 3.2 reports the resulting mean symbolic composite and composite MLEs, $\hat{\theta}^{(2)}_{SCL}$ and $\hat{\theta}^{(2)}_{CL}$, with standard errors in parentheses, based on 1 000 replicate analyses, for different values of B. While for low B there is high variability in the estimates, as B increases the mean MLEs and standard errors approach the same quantities obtained under the classical data analysis, even in cases of very strong (Σ_4) or very weak (Σ_5) dependence.

In this case, comparable estimates to the composite MLEs are available for $B = 25$, however practically viable estimates (with larger variances) can be obtained for much smaller values ($B \approx 10$).

Model	B	σ_{11}	σ_{12}	σ_{22}	μ	σ	ξ
Σ_1	2	335.5 (585.5)	5.7 (232.2)	317.2 (125.1)	0.0383 (0.1639)	0.8687 (0.0061)	-0.0194 (0.0301)
	3	301.0 (34.5)	-0.1 (16.9)	301.9 (33.5)	0.0812 (0.0550)	0.9195 (0.0342)	0.0182 (0.0210)
	5	299.1 (23.1)	-0.9 (13.2)	299.9 (24.1)	0.0067 (0.0295)	0.9666 (0.0285)	0.0136 (0.0194)
	10	299.8 (20.2)	-0.5 (11.1)	300.0 (20.9)	-0.0015 (0.0276)	0.9898 (0.0186)	0.0039 (0.0120)
	15	299.8 (18.9)	-0.3 (10.4)	300.0 (19.5)	-0.0017 (0.0272)	0.9929 (0.0179)	0.0027 (0.0110)
	25	299.7 (18.0)	-0.3 (10.0)	300.2 (18.9)	-0.0016 (0.0272)	0.9954 (0.0179)	0.0013 (0.0102)
	Classic	**300.76 (17.1)**	**-0.4 (9.7)**	**301.02 (18.1)**	**-0.0019 (0.0262)**	**0.9986 (0.0173)**	**0.0007 (0.0084)**
Σ_2	2	316.59 (149.1)	165.1 (246.8)	332.9 (153.5)	0.3763 (0.1448)	0.8671 (0.0632)	-0.0163 (0.0284)
	3	299.6 (35.0)	149.7 (24.9)	300.8 (33.7)	0.0755 (0.0439)	0.9258 (0.0284)	0.0151 (0.0192)
	5	298.9 (23.4)	149.2 (16.7)	299.9 (23.4)	0.0077 (0.0280)	0.9705 (0.0266)	0.0114 (0.0182)
	10	299.3 (20.2)	149.6 (13.9)	300.3 (19.9)	0.0002 (0.0267)	0.9912 (0.0182)	0.0023 (0.0118)
	15	299.4 (19.2)	149.7 (13.2)	300.5 (19.0)	-0.0001 (0.0265)	0.9941 (0.0179)	0.0021 (0.0108)
	25	299.7 (18.3)	149.9 (12.5)	300.5 (18.1)	0.0001 (0.0265)	0.9964 (0.0176)	0.0009 (0.0100)
	Classic	**300.7 (17.0)**	**150.4 (11.6)**	**301.53 (17.0)**	**-0.0002 (0.0258)**	**0.9997 (0.0172)**	**0.0004 (0.0081)**
Σ_3	2	321.6 (360.0)	162.3 (210.6)	210.8 (131.2)	0.3596 (0.1310)	0.8671 (0.0586)	-0.0150 (0.0271)
	3	296.1 (30.6)	147.4 (20.1)	197.9 (19.9)	0.0723 (0.0422)	0.9302 (0.0280)	0.0113 (0.0174)
	5	298.8 (23.3)	149.4 (15.3)	199.6 (15.4)	0.0065 (0.0263)	0.9713 (0.0237)	0.0102 (0.0170)
	10	299.0 (19.3)	149.6 (12.3)	199.7 (12.9)	-0.0001 (0.0252)	0.9908 (0.0174)	0.0031 (0.0114)
	15	299.5 (18.7)	149.8 (11.6)	199.8 (12.1)	-0.0009 (0.0249)	0.9942 (0.0170)	0.0021 (0.0105)
	25	299.7 (17.8)	150.0 (11.2)	200.0 (11.8)	-0.0009 (0.0251)	0.9963 (0.0168)	0.0009 (0.0096)
	Classic	**300.7 (16.4)**	**150.6 (10.2)**	**200.6 (10.9)**	**-0.0013 (0.0243)**	**0.9993 (0.0164)**	**0.0004 (0.0079)**
Σ_4	2	3554 (2071)	1848 (1319)	3473 (1839)	0.4337 (0.2211)	0.8691 (0.0847)	-0.0393 (0.0342)
	3	2954 (435)	1453 (294)	2952 (405)	0.0857 (0.0729)	0.9132 (0.0418)	0.0202 (0.0250)
	5	3003 (345)	1500 (244)	2996 (337)	0.0071 (0.0355)	0.9626 (0.0366)	0.0156 (0.0258)
	10	3002 (249)	1506 (169)	2997 (239)	-0.0004 (0.0323)	0.9891 (0.0233)	0.0030 (0.0172)
	15	2992 (217)	1498 (148)	2988 (211)	-0.0009 (0.0318)	0.9930 (0.0224)	0.0009 (0.0147)
	25	2992 (199)	1499 (136)	2991 (200)	-0.0010 (0.0318)	0.9953 (0.0222)	-0.0001 (0.0128)
	Classic	**3002 (190)**	**1503 (124)**	**2999 (189)**	**-0.0001 (0.0308)**	**0.9988 (0.0217)**	**-0.0025 (0.0113)**
Σ_5	2	30.97 (3.57)	15.53 (2.81)	30.98 (3.86)	0.3356 (0.1003)	0.8662 (0.0456)	-0.0002 (0.0093)
	3	29.83 (2.04)	14.89 (1.58)	29.82 (2.18)	0.0633 (0.0246)	0.9452 (0.0184)	0.0032 (0.0099)
	5	29.86 (1.54)	14.85 (1.17)	29.82 (1.71)	0.0071 (0.0157)	0.9821 (0.0140)	0.0021 (0.0076)
	10	29.93 (1.27)	14.92 (0.95)	29.91 (1.45)	0.0012 (0.0149)	0.9928 (0.0111)	0.0009 (0.0046)
	15	29.96 (1.20)	14.93 (0.91)	29.91 (1.33)	0.0004 (0.0146)	0.9952 (0.0108)	0.0007 (0.0038)
	25	29.97 (1.13)	14.95 (0.86)	29.94 (1.28)	0.0001 (0.0145)	0.9970 (0.0106)	0.0003 (0.0031)
	Classic	**30.10 (0.94)**	**15.06 (0.66)**	**30.06 (1.03)**	**-0.0004 (0.0144)**	**0.9997 (0.0104)**	**0.0000 (0.0004)**

Table 3.2: Mean (and standard errors) of the symbolic composite MLE $\hat{\theta}^{(2)}_{SCL}$ and composite MLE $\hat{\theta}^{(2)}_{CL}$ (Classic) from 1000 replications of the Gaussian max-stable process model, for $B \times B$ histograms for varying values of B. Results based on $N = 1\,000$ observations at $K = 15$ spatial locations and $T = 1$ random histogram.

Varying the number of bins and marginal histogram dimension

We generate $N = 10^6$ realisations for $K = 10$ locations using the covariance parameter specification $\Sigma = \Sigma_3$. Both pairwise ($B_2 \times B_2$ marginal histograms) and triplewise ($B_3 \times B_3 \times B_3$ marginal histograms) symbolic composite MLEs, $\hat{\theta}^{(2)}_{SCL}$ and $\hat{\theta}^{(3)}_{SCL}$, were computed and compared for varying values of B_2 and B_3, constructed from a single ($T = 1$) random histogram.

Table 3.3 reports the resulting means and standard errors of $\hat{\theta}^{(2)}_{SCL}$ and $\hat{\theta}^{(3)}_{SCL}$ obtained over 200 replicate analyses. Each row represents marginal pairwise and triplewise histograms with approximately equal numbers of bins (i.e. $B_2^2 \approx B_3^3$) representing approximately equivalent computational overheads. As before, both symbolic composite MLEs converge as the number of bins increases.

When the number of bins are comparable (i.e. $B_2^2 \approx B_3^3$) the pairwise estimates invariably have smaller standard errors than the triplewise estimates. This can be attributed to the direct tradeoff between a lower resolution histogram in higher dimensions compared to a higher resolution histogram in lower dimensions, when keeping the number of histogram bins comparable. In this case, the extra lower-dimensional precision is more informative for the model parameters than higher-dimensional information, and so the pairwise estimator is more efficient. However, when the number of bins in each margin is the same ($B_2 = B_3$), so that the resolution in each dimension is the same, but where the triplewise estimator uses higher-dimensional information (using more bins), then the triplewise composite MLE is naturally the most efficient.

$B_2^2 \vert B_3^2$	σ_{11}		σ_{12}		σ_{22}	
	Pair	Triple	Pair	Triple	Pair	Triple
$3^2 \vert 2^3$	300.62 (2.80)	298.98 (8.45)	150.35 (1.94)	149.36 (5.76)	200.14 (1.74)	199.68 (5.46)
$5^2 \vert 3^3$	300.55 (0.95)	300.23 (2.44)	150.40 (0.66)	150.09 (1.66)	200.26 (0.55)	200.02 (1.50)
$8^2 \vert 4^3$	300.45 (0.80)	300.21 (1.28)	150.31 (0.54)	150.16 (0.86)	200.20 (0.50)	200.07 (0.82)
$11^2 \vert 5^3$	300.57 (0.72)	300.42 (0.91)	150.39 (0.46)	150.30 (0.62)	200.22 (0.38)	200.19 (0.56)

$B_2^2 \vert B_3^2$	μ		σ		ξ	
	Pair	Triple	Pair	Triple	Pair	Triple
$3^2 \vert 2^3$	0.0426 (0.0217)	0.1515 (0.0494)	0.9803 (0.0094)	0.9718 (0.0112)	0.0039 (0.0023)	0.0004 (0.0055)
$5^2 \vert 3^3$	0.0016 (0.0025)	0.0411 (0.0209)	0.9978 (0.0033)	0.9807 (0.0092)	0.0008 (0.0013)	0.0037 (0.0023)
$8^2 \vert 4^3$	0.0001 (0.0007)	0.0093 (0.0079)	0.9999 (0.0007)	0.9926 (0.0056)	0.0001 (0.0001)	0.0020 (0.0016)
$11^2 \vert 5^3$	0.0000 (0.0008)	0.0015 (0.0023)	0.9999 (0.0001)	0.9978 (0.0029)	0.0000 (0.0001)	0.0008 (0.0011)

Table 3.3: Mean (and standard errors) of the pairwise ($\hat{\theta}^{(2)}_{SCL}$) and triplewise ($\hat{\theta}^{(3)}_{SCL}$) symbolic composite MLEs from 200 replications of the Gaussian max-stable process model for $B_2 \times B_2$ (pairwise) and $B_3 \times B_3 \times B_3$ (triplewise) histograms, with varying B_2, B_3. Rows correspond to $B_2^2 \approx B_3^3$ to compare approximately equal numbers of histogram bins. Results based on $N = 10^6$ observations at $K = 10$ spatial locations, $T = 1$ random histogram and $\Sigma = \Sigma_3$.

Varying the number of spatial locations, K

We generate $N = 10^6$ realisations at K locations (for varying K) using the covariance parameter specification $\Sigma = \Sigma_3$. The random locations for smaller K are a subset of those for larger K. Both pairwise and triplewise symbolic composite MLEs, $\hat{\theta}^{(2)}_{SCL}$ and $\hat{\theta}^{(3)}_{SCL}$, are computed, using $B_2 \times B_2$ and $B_3 \times B_3 \times B_3$ random marginal histograms, where $B_2 = 8$ and $B_3 = 4$ so that each marginal histogram has 64 bins.

Table 3.4 reports the resulting means and standard errors of $\hat{\theta}^{(2)}_{SCL}$ and $\hat{\theta}^{(3)}_{SCL}$ for different values of K, based on 200 replicate analyses. As expected, as K increases both composite MLEs become increasingly accurate, particularly the dependence parameters $(\sigma_{11}, \sigma_{12}, \sigma_{22})$, as the amount of spatial information increases, with the pairwise composite MLEs producing more accurate estimates for an equivalent number of bins. These results are consistent with those for standard pairwise and triplewise composite MLEs seen in e.g. Padoan et al. (2010) and Genton et al. (2011).

Varying the number of underlying observations, N

One of the motivations for aggregating micro-data into random histograms before an analysis is that the analysis, while losing some information in the data, will be much faster. We generate $N = 10^3, \ldots, 10^7$ realisations for $K = 10$ locations using the covariance parameter specification $\Sigma = \Sigma_3$. We compute standard pairwise composite ($\hat{\theta}^{(2)}_{CL}$) and symbolic pairwise composite ($\hat{\theta}^{(2)}_{SCL}$) MLEs, with $B_2 = 25$ and $T = 1$.

Table 3.5 reports the resulting means and standard errors of $\hat{\theta}^{(2)}_{CL}$ and $\hat{\theta}^{(2)}_{SCL}$ for different values of N, based on 100 replicate analyses. As expected, as N increases the composite MLEs become increasingly accurate, with the standard composite MLEs outperforming the symbolic composite MLEs, although the difference here is relatively minor as we are using 25×25 histogram bins in each pairwise comparison. However, it was not computationally viable to compute $\hat{\theta}^{(2)}_{CL}$ for $N \geq 10^6$. To explore this in more detail, these simulations were repeated for $K = 20, 50, 100$

K	σ_{11}		σ_{12}		σ_{22}	
	Pair	Triple	Pair	Triple	Pair	Triple
3	300.44 (5.80)	299.24 (13.37)	150.30 (2.41)	150.02 (6.75)	201.55 (11.12)	200.12 (7.84)
5	300.35 (1.53)	299.95 (2.37)	150.28 (1.10)	150.02 (1.99)	200.22 (1.00)	199.98 (1.89)
10	300.21 (0.88)	299.95 (1.22)	150.15 (0.59)	149.99 (0.83)	200.10 (0.53)	199.94 (0.77)
15	300.19 (0.71)	299.93 (1.12)	150.12 (0.48)	150.00 (0.73)	200.06 (0.46)	200.00 (0.72)
20	300.20 (0.78)	299.99 (0.99)	150.14 (0.47)	150.02 (0.70)	200.08 (0.44)	199.99 (0.69)

K	μ		σ		ξ	
	Pair	Triple	Pair	Triple	Pair	Triple
3	-0.00003 (0.0011)	0.00727 (0.0102)	0.9999 (0.0009)	0.9947 (0.0069)	0.00006 (0.00062)	0.00121 (0.00193)
5	-0.00002 (0.0010)	0.00671 (0.0088)	0.9999 (0.0008)	0.9950 (0.0064)	0.00008 (0.00059)	0.00111 (0.00187)
10	-0.00006 (0.0009)	0.00595 (0.0068)	0.9999 (0.0007)	0.9956 (0.0047)	0.00009 (0.00048)	0.00093 (0.00133)
15	-0.00004 (0.0001)	0.00553 (0.0054)	0.9999 (0.0007)	0.9958 (0.0042)	0.00007 (0.00042)	0.00092 (0.00131)
20	-0.00005 (0.0001)	0.00524 (0.0053)	0.9999 (0.0007)	0.9961 (0.0039)	0.00006 (0.00048)	0.00080 (0.00121)

Table 3.4: Mean (and standard errors) of the pairwise ($\hat{\theta}^{(2)}_{SCL}$) and triplewise ($\hat{\theta}^{(3)}_{SCL}$) symbolic composite MLEs from 200 replications of the Gaussian max-stable process model for $B_2 \times B_2$ (pairwise) and $B_3 \times B_3 \times B_3$ (triplewise) histograms, with varying K. Results based on $N = 10^6$ observations in $T = 1$ random histogram with $B_2 = 8$ and $B_3 = 4$ (so that $B_2^2 = B_3^3$) and $\Sigma = \Sigma_3$.

spatial locations, and a slightly smaller range of observed data ($N = 1\,000$ to $500\,000$) to provide a better comparison with the standard composite MLEs.

Table 3.6 summarises the mean computation times (in seconds) for different stages involved in computing the composite MLEs, based on 10 replicate analyses. Simply in terms of optimising the respective likelihood functions, the symbolic composite likelihood (t_s) is much more efficient than the equivalent composite likelihood (t_c). The computational overheads of the former are essentially constant with respect to N, and so these are largely driven by the number of pairwise components ($K/(K-1)/2$) in the likelihood. The computational overheads of the composite likelihood are driven both by N and K, and so computing $\hat{\theta}^{(2)}_{CL}$ becomes largely impractical when either becomes moderately large. Clearly computation of $\hat{\theta}^{(2)}_{SCL}$ would take similar times to those in Table 3.6 for considerably larger N.

An additional step in computing $\hat{\theta}^{(2)}_{SCL}$ is construction of all bivariate marginal histograms $\boldsymbol{S}^{\boldsymbol{i}}, \boldsymbol{i} \in \mathcal{I}_2$. We constructed these in two alternative ways: using the `R` function `hist` (t_{histR}) and the `R` package `DeltaRho` (t_{histDR}) which provides an interface to `map-reduce` functionality whereby the histograms can be constructed in parallel on multiple processors and machines, and then combined.

For small values of N, using the simple `hist` function on a local machine is quicker than using `DeltaRho` and communicating between multiple machines. However, `DeltaRho` increasingly outperforms `hist` as the number of datapoints N increases. Our `DeltaRho` setup was modest with only 4 parallel machines; more expansive setups could drastically reduce histogram construction time for large N. Regardless of the histogram construction method adopted, it is clear that computing the symbolic composite MLE is considerably more efficient than the standard composite MLE.

N	σ_{11}		σ_{12}		σ_{22}	
	Classic	Pair	Classic	Pair	Classic	Pair
10^3	299.48 (17.09)	298.11 (17.24)	149.90 (10.37)	148.84 (11.05)	200.45 (11.05)	200.11 (11.69)
10^4	299.07 (5.76)	298.56 (6.07)	149.65 (3.26)	149.09 (3.63)	199.92 (3.32)	199.39 (3.70)
10^5	300.56 (1.56)	300.49 (2.07)	150.42 (0.98)	150.32 (1.27)	200.28 (1.14)	200.18 (1.49)
10^6	–	300.21 (0.61)	–	150.18 (0.45)	–	200.14 (0.43)
10^7	–	300.13 (0.23)	–	150.06 (0.17)	–	200.02 (0.18)

N	μ		σ		ξ	
	Classic	Pair	Classic	Pair	Classic	Pair
10^3	-0.0074 (0.0280)	-0.0077 (0.0286)	0.9972 (0.0169)	0.9964 (0.0170)	0.0016 (0.0115)	0.0024 (0.0123)
10^4	-0.0017 (0.0074)	-0.0013 (0.0076)	0.9989 (0.0051)	0.9988 (0.0052)	-0.0002 (0.0039)	-0.0002 (0.0040)
10^5	-0.0002 (0.0021)	-0.0002 (0.0025)	1.0000 (0.0014)	1.0000 (0.0015)	0.0001 (0.0010)	0.0001 (0.0013)
10^6	–	0.0000 (0.0007)	–	1.0000 (0.0004)	–	0.0000 (0.0004)
10^7	–	-0.0001 (0.0002)	–	1.0000 (0.0001)	–	0.0000 (0.0001)

Table 3.5: Mean (and standard errors) of the standard pairwise composite ($\hat{\theta}^{(2)}_{CL}$) and symbolic pairwise composite ($\hat{\theta}^{(2)}_{SCL}$) MLEs from 100 replications of the Gaussian max-stable process model with $B_2 \times B_2$ histograms with $B_2 = 25$. Results are based on $K = 10$ spatial locations, $T = 1$ random histogram and $\Sigma = \Sigma_3$.

Varying the number of histograms, T

Until now the N observed datapoints have been aggregated into a single histogram, $T = 1$ (or more precisely one low-dimensional marginal histogram per composite likelihood component). If each histogram $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_T$ has exactly the same bins then collapsing these to a single histogram, as discussed in Section 3.3.2, will produce the same symbolic composite MLE as if $T > 1$ histograms were used. However the number of random histograms T will affect the standard errors of $\hat{\theta}^{(j)}_{SCL}$, as discussed in Section 3.3.3. That is, by aggregating the spatially observed micro-data over multiple time points, there is a loss of information in knowing which observations at location t_i occurred at the same time as observations at location t_j within the same random histogram. This results in a loss of spatial information, which will impact the efficiency of the symbolic likelihood estimators.

To examine this we generate $N = 1\,000$ realisations for $K = 10$ spatial locations using the covariance parameter specification $\Sigma = \Sigma_3$. We compute the standard composite ($\hat{\theta}^{(2)}_{CL}$) and symbolic ($\hat{\theta}^{(2)}_{SCL}$) pairwise composite MLEs when aggregating the observations equally into $T = 4, 5, 10, 20, 40, 50, 100, 200$ and 1000 histograms $\boldsymbol{S}_t$ (so that for $T = 1\,000$ we have 1 observation per random histogram), with $B \times B = 25^2$ bins in each pairwise marginal histogram. The means of the Godambe standard errors for the composite MLEs for each value of T are reported in Table 3.7, based on 1 000 replicate analyses. This procedure is then repeated 100 times while varying the number of marginal histogram bins (B^2), with the results illustrated in Figure 3.2.

From Table 3.7, for a small number of histograms the estimated standard errors are large compared to the standard composite likelihood estimates due to the significant loss of temporal information. As T increases these standard errors reduce as more temporal information is recovered. With $T = N$ (and one data point per histogram) the standard errors become comparable, although the location of the single datapoint within each histogram for $T = N$ is still uncertain, and so unless the number of bins also increases, the standard errors of the symbolic composite MLE will be larger than those of the standard composite MLE, even for $T = N$.

N	$K = 10$				$K = 20$			
	t_c	t_s	t_{histDR}	t_{histR}	t_c	t_s	t_{histDR}	t_{histR}
1 000	71.9	22.5	0.8	0.1	383.4	79.6	1.8	0.4
5 000	291.8	19.0	0.8	0.3	1 578.2	99.3	2.1	1.0
10 000	591.7	23.8	0.9	0.5	3 125.4	103.2	2.4	1.8
50 000	2 626.8	24.2	1.7	2.1	20 459.4	107.3	4.5	7.6
100 000	5 610.7	25.4	2.4	4.2	–	115.0	6.9	14.9
500 000	31 083.1	23.2	7.5	20.6	–	96.1	26.6	73.5

N	$K = 50$				$K = 100$			
	t_c	t_s	t_{histDR}	t_{histR}	t_c	t_s	t_{histDR}	t_{histR}
1 000	7 333.9	528.5	9.3	3.0	–	2 238.0	78.8	12.0
5 000	27 616.5	665.1	10.6	7.7	–	2 650.2	81.7	30.9
10 000	–	696.3	12.4	13.5	–	2 356.6	85.8	54.1
50 000	–	744.8	24.8	59.0	–	2 300.6	131.6	237.0
100 000	–	768.1	41.3	115.7	–	2 766.9	188.2	461.8
500 000	–	802.9	156.1	561.3	–	3 111.5	627.1	2 243.5

Table 3.6: Mean computation times (seconds) for different components involved in computing $\hat{\theta}^{(2)}_{CL}$ and $\hat{\theta}^{(2)}_{SCL}$ for different classical dataset sizes N and number of spatial locations K, based on 10 replicate analyses. Columns t_c and t_s respectively show the time taken to optimise the standard composite and symbolic composite likelihood functions. Columns t_{histDR} and t_{histR} show the time taken to aggregate the data into histograms using `DeltaRho` and `R` function `hist` respectively. Results are based on $T = 1$ random histogram and $\Sigma = \Sigma_3$.

Figure 3.2 illustrates how the mean Godambe standard errors, for fixed T, approach the (square root of the) appropriate diagonal term of the limit (3.14) of the variability matrix $\hat{J}(\hat{\theta}^{(2)}_{SCL})$, as the number of histogram bins becomes large. As $T \to N$ this limit (horizontal dashed lines) approaches the equivalent standard errors under the standard composite likelihood (the lowest horizontal dashed line).

Of course, while standard error accuracy increases for larger T, computational overheads increase in proportion to T. Hence in practice, and with equal bins over all histograms, to compute the symbolic composite MLE $\hat{\theta}^{(j)}_{SCL}$ we would use $T = 1$, whereas to compute standard errors we would use as small a number of histograms as possible (to maximise computational efficiency) such that the scale of the standard errors is acceptable within the context of the given analysis.

T	σ_{11}	σ_{12}	σ_{22}	μ	σ	ξ
5	217.81	147.60	158.48	0.31	0.19	0.13
10	167.90	113.21	122.55	0.23	0.15	0.10
20	122.00	82.66	88.64	0.17	0.11	0.07
50	79.09	54.10	57.91	0.11	0.07	0.05
100	56.23	38.37	40.93	0.08	0.05	0.03
200	40.01	27.19	29.02	0.06	0.04	0.02
1000	17.94	12.28	13.07	0.03	0.02	0.01
Classic	**16.65**	**11.53**	**12.69**	**0.021**	**0.014**	**0.008**

Table 3.7: Means of the estimated Godambe standard errors of $\hat{\theta}^{(2)}_{SCL}$ and $\hat{\theta}^{(2)}_{CL}$ for different numbers of random histograms, T, based on 1 000 replicate analyses. Results are based on $N = 1\,000$ observations with $B = 25$ and $\Sigma = \Sigma_3$.

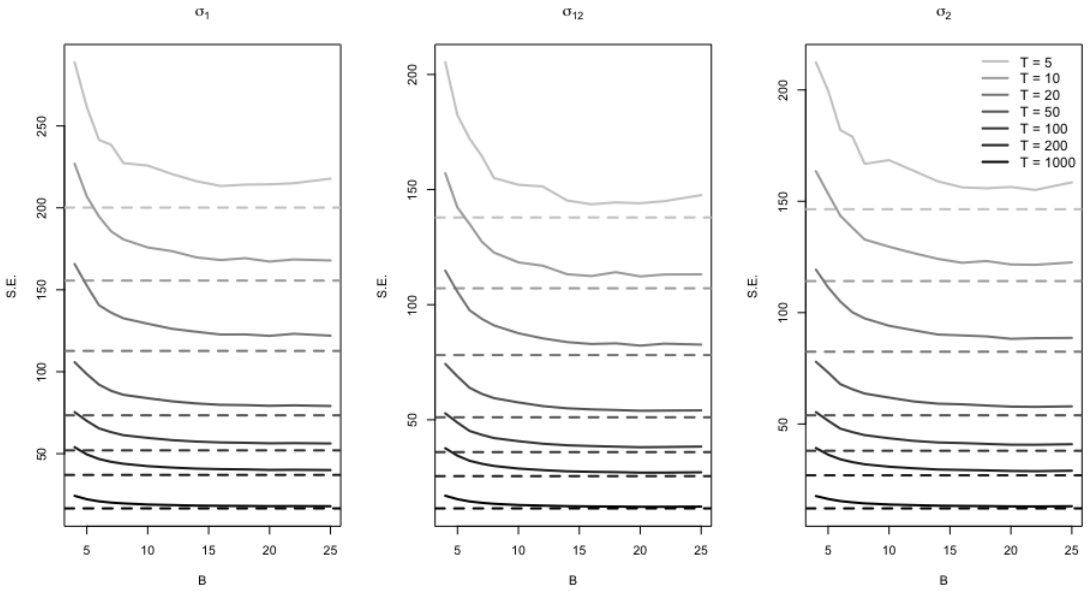

Figure 3.2: Godambe standard errors (solid lines) for the dependance parameters $(\sigma_{11}, \sigma_{12}, \sigma_{22})$ of $\hat{\theta}^{(2)}_{SCL}$ for varying number of random histograms T, and number of marginal histogram bins B^2. Dashed horizontal lines denote the appropriate term of the limit (3.14) of the variability matrix $\hat{J}(\hat{\theta}^{(2)}_{SCL})$. Results are based on $N = 1\,000$ observations with $\Sigma = \Sigma_3$.

3.5 Analysis of millennial scale climate extremes

We consider daily maxima of historical temperature data (1850–2006) and future simulated temperature data (2006–2100) simulated using the CSIRO Mk3.6 climate model, for 105 grid locations (considered as the spatial co-ordinates) at the centre of $1.875° \times 1.875°$ grid cells over Australia (Figure 3.3). Two different scenarios (RCP4.5 and RCP8.5) are used to generate the future data, which represent two of the four greenhouse gas scenarios projected by the Intergovernmental Panel on Climate Change (IPCC) based on how much greenhouse gases are emitted in future years (Stocker et al., 2013). Due to seasonal periodicity, only data from 90 days across the summer months (December–February) are considered, to induce approximate stationarity of the process. Due to the temporal dependence evident in the RCP4.5 and RCP8.5 data the daily maximum temperatures at each spatial location were linearly detrended, so that the resulting block-maxima constitute the largest deviation above the mean temperature. Maxima are computed over 15-day blocks, resulting in 6 observations per year, and $N = 936$ and 570 total observations per location for the historical and climate model data respectively. Following Padoan et al. (2010) and Blanchet and Davison (2011) we fit the Gaussian max-stable process (Smith, 1990) model with spatially varying marginal parameters, in particular with

$$\mu(k) = \alpha_0 + \alpha_1 x(k) + \alpha_2 y(k), \qquad \sigma(k) = \beta_0 + \beta_1 x(k) + \beta_2 y(k), \qquad \xi(k) = \xi,$$

where $(x(k), y(k))$ are the spatial co-ordinates of the k-th location. Other co-variates (such as altitude) were not considered due to the reasonably flat nature of the topography across the

study region.

Table 3.8 lists the total number of terms in the standard pairwise composite likelihood, $\ell_{CL}^{(2)}(\theta)$, and the symbolic composite likelihood, $\ell_{CL}^{(2)}(\theta)$, for a single ($T = 1$) bivariate $B \times B$ histogram with $B = 15, 20, 25, 30$. While the number of terms in the symbolic likelihood is guaranteed to be lower than the standard likelihood if $B^2 < N$, in practice the number of non-empty histogram bins contributing to the likelihood (centre column, Table 3.8) can be considerably smaller, particularly for strongly dependent data. For the current analyses, the symbolic composite likelihood has significantly fewer terms, leading to substantially faster optimisation and lower computational costs than the standard composite likelihood. As discussed in Section 3.3, the symbolic composite MLE ($\hat{\theta}_{SCL}^{(2)}$) can be computed exactly with $T = 1$ random histogram, and so this optimisation (which evaluates the target function many times) can be very efficient. In contrast, $T = N$ histograms are required for the best variance estimates (see Table 3.7), and so the resulting computational overheads are comparable to that of the standard composite likelihood (though these are only a small proportion of total computation).

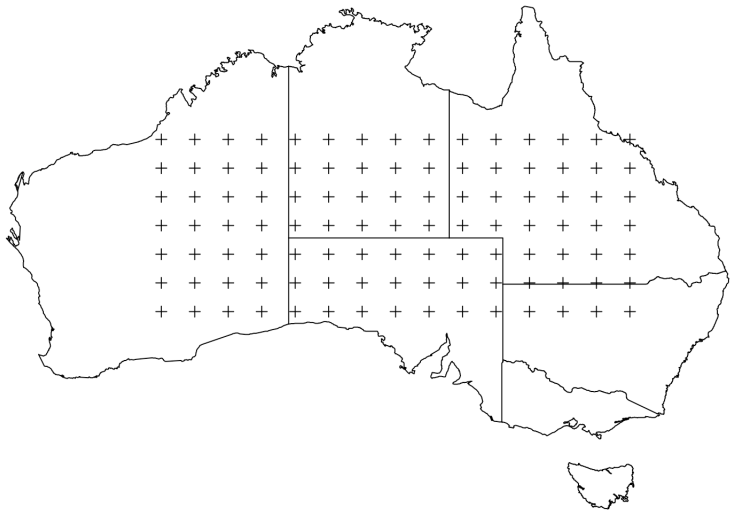

Figure 3.3: $K = 105$ spatial locations for the historical and future-simulated temperature data over Australia. Each cross represents the midpoint of a $1.875° \times 1.875°$ box in a spatial grid.

Table 3.9 displays the symbolic composite MLEs (and standard errors) of the three dependence parameters and the marginal shape parameter ξ for the Smith model, calculated using $B = 15, 20, 25, 30$. Comparable estimates are obtained for each value of B, with some clear convergence in both the point estimates and their standard errors as the resolution of each histogram increases. While the standard errors are naturally larger than those under the standard composite likelihood by construction, they are sufficiently small compared to the magnitude of the composite MLE in order to make meaningful inference.

Compared to the observed historical extremes, we can see a slight increase in spatial dependence for the RCP4.5 scenario data and a significant decrease in dependence for the RCP8.5

B	Historical ($N = 936$)	Actual RCP4.5/8.5 ($N = 570$)	Maximum RCP4.5/8.5 ($N = 570$)
15	642 898	529 584	1 228 500
20	960 403	774 060	2 184 000
25	1 286 714	1 016 565	3 412 500
30	1 609 923	1 247 465	4 914 000
Classic	**5 110 560**	**3 112 200**	**3 112 200**

Table 3.8: Total number of terms in each pairwise composite likelihood function for $N = 936, 570$ block maxima over $K = 105$ spatial locations. For standard composite likelihoods this corresponds to $NK(K-1)/2$ terms. For the symbolic composite likelihood constructed using a single ($T = 1$) $B \times B$ histogram, this corresponds to a maximum of $B^2K(K-1)/2$ terms. The actual number of symbolic composite likelihood terms corresponds to the number of non-empty histogram bins.

scenario.

The marginal shape parameter ξ is negative for all three datasets, with larger composite MLEs estimated for the future-simulated data compared to the historical data. This implies that the RCP4.5 and RCP8.5 data have higher upper bounds than that of the historical dataset, meaning larger deviations from the mean are expected for the future scenarios.

Figure 3.4 illustrates expected and observed (columns) 95-year return levels for each dataset (rows) for $B = 15, 30$. Higher expected (and observed) returns for the RCP4.5 and RCP8.5 scenarios compared to the historical setting are apparent.

Because extrapolation into and beyond the tails of observed data is sensitive to a model's parameter estimates, there are some differences in the return levels for the different values of B. This suggests that, for applications in spatial extremes at least, higher resolution histograms may be required, depending on the nature of inference required.

3.6 Discussion

In this article we have introduced a novel method for constructing composite likelihood functions for histogram-valued random variables. Working with random histograms as summaries of large datasets allows for computational efficiencies, as the histograms can efficiently represent large amounts of data in a concise form. The benefit of working with composite likelihoods in this setting is that the inefficiencies of working with histograms for higher dimensional data can largely be avoided.

Our theoretical results show that if the bins in each random histogram are the same, then the symbolic composite MLE can be computed exactly by combining the data into a single histogram (by summing the totals in each bin). As the majority of the computational time for an analysis is spent in optimising the likelihood, this is a particularly useful result that can lead to fast inference. The precision of the composite MLE, however, depends on the number of histograms: the more there are (assuming equal numbers of datapoints in each histogram), the lower the estimated variance of the composite MLE. This will either present hard limits on the possible level of inferential precision (if pre-made histograms are presented directly to

Historical Data

B	σ_{11}	σ_{12}	σ_{22}	ξ
15	176.4 (2.85)	-28.7 (0.32)	76.8 (3.29)	-0.266 (0.053)
20	164.2 (2.89)	-29.3 (0.30)	74.3 (4.69)	-0.264 (0.049)
25	162.4 (2.17)	-29.9 (0.33)	75.3 (2.84)	-0.264 (0.049)
30	161.6 (2.01)	-32.3 (0.29)	74.4 (2.34)	-0.264 (0.050)

RCP4.5 Data

B	σ_{11}	σ_{12}	σ_{22}	ξ
15	160.9 (9.42)	-34.1 (0.83)	79.0 (2.22)	-0.249 (0.074)
20	163.5 (5.95)	-41.1 (0.73)	77.6 (2.45)	-0.249 (0.076)
25	150.3 (3.49)	-33.1 (0.65)	70.7 (1.70)	-0.250 (0.073)
30	150.2 (1.50)	-31.6 (0.24)	70.7 (1.54)	-0.250 (0.069)

RCP8.5 Data

B	σ_{11}	σ_{12}	σ_{22}	ξ
15	128.7 (8.60)	-19.6 (0.92)	67.7 (3.92)	-0.232 (0.061)
20	128.0 (6.30)	-19.6 (1.29)	66.6 (3.32)	-0.231 (0.059)
25	136.0 (3.95)	-15.1 (0.93)	59.4 (3.17)	-0.234 (0.060)
30	129.9 (4.01)	-13.6 (0.83)	56.4 (2.94)	-0.233 (0.055)

Table 3.9: The mean and standard errors of the composite MLEs for Σ obtained for the 105 locations across Australia from the bivariate symbolic composite log-likelihood function for $B = 15, 20, 25, 30$.

the analyst), or allow a trade-off of precision for computation to be made. As computation of the Godambe information matrix is trivial compared to estimation of the composite MLE, if the full dataset is available, then a large number of histograms could be used for relatively low computational costs.

Our results have also shown the efficiency of standard composite likelihood techniques when the data are grouped into time blocks such that it is know which block any data point belongs to, but it is not known where the datapoint lies within each block.

We have not considered the question of how to best construct the random histograms. This was considered in the present context by Zhang et al. (2019) and Beranger et al. (2018). Possible approaches could follow standard nonparametric arguments of histogram binwidth selection (e.g. Scott and Sheather (1985), Wand (1997)) or more complex space-partitioning processes such as random trees, or alternatively be chosen to optimise pre-specified utility or loss functions. This is a current topic of active research. In terms of determining a sufficient number of bins B to ensure convergence to the classical composite MLE, a naive approach was utilised in the real data analysis in Section 3.5, in which increasing values of B were investigated until comparable results were obtained. While this approach does give an indication of where convergence to the classical composite MLE has occurred, ideally a method would be developed in which only one value of B needs to be investigated. Similarly to the optimal histogram construction previously described, the sufficient value for B could possibly be determined via a binwidth selection algorithm, or other loss-function based methods.

One of our motivations for analysing the extremes of very large climate datasets is that, while exceptions exist, it is not uncommon for statistical analysis to only occur independently

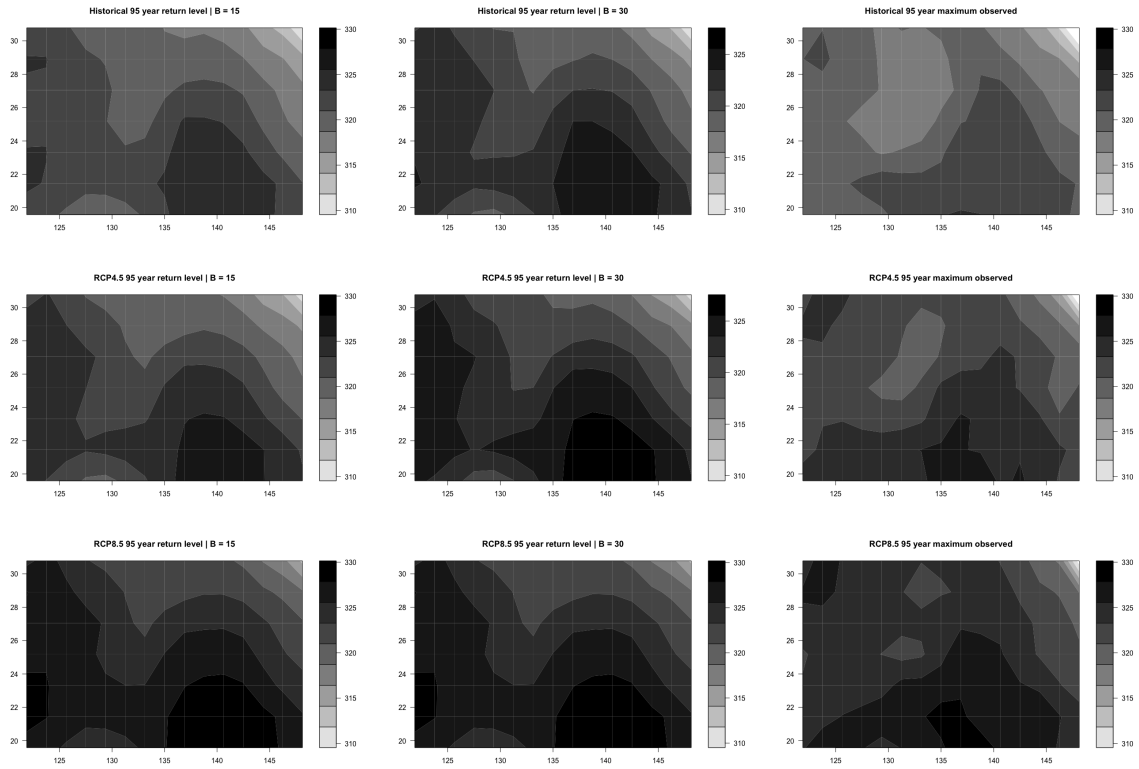

Figure 3.4: Predicted and observed 95-years return levels over Australia based on historical (top row), RCP4.5 (middle row) and RCP8.5 (bottom row) scenario data. Columns denote predictions based on $B^2 = 15 \times 15$ (left) and $B^2 = 30 \times 30$ (middle) histograms and interpolated observed maxima (right).

at each spatial location, with very little work done to analyse the spatial dependence (Huang et al., 2016). In Section 3.5, by fitting the Gaussian max-stable process to historical and future scenario Australian temperature data, we were able to explore changes in the spatial dependence structure that will accompany different levels of greenhouse gas emission levels in the coming years, and provide insight into the effects of these changes. It would be extremely challenging to perform these analyses, and others with even larger datasets, using standard techniques.

For the analysis of Australian temperature extremes, the data are presented as being located at the centre of a box within a grid. As such, the presented analysis ignores the fact that the data actually arose from the entire box, and not just this point location. One possible extension of the work in this article is to similarly treat the actual spatial locations of each datapoint within each grid box as unknown locations within a spatial histogram.

This would also allow datasets with extremely large numbers of locations (K) to be spatially aggregated into smaller datasets with spatial bins as the locations instead of pointwise coordinates, potentially drastically decreasing the computational cost and allowing the analysis of much higher dimensional data.

Acknowledgements

We are grateful to Dr Markus Donat (Climate Change Research Centre, UNSW Sydney) for providing the climate dataset used in Section 3.5. This research is supported by the Australian Research Council through the Australian Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS; CE140100049), and the Discovery Project scheme (FT170100079).

Chapter 4

Logistic regression models for aggregated data

4.1 Introduction

There are many well developed statistical methods for classification, such as logistic regression, discriminant analysis and clustering, which predict a categorical variable that can take one of K distinct values given an input vector of predictor variables (e.g. Hastie et al., 2008, Pampel, 2000). While these methods are effective for the analysis of standard data, when the data take non-standard forms, such as random interval- or random histogram-based predictors, existing methods are either underdeveloped or do not exist. Interval, histogram and other-distribution based data summaries can arise through measurement error, data quantisation, expert elicitation and, motivating this work, the desire to summarise large and complex datasets in an appropriate way so that they can be analysed more efficiently than the full dataset (e.g. Zhang et al., 2019). The field of Symbolic Data Analysis (SDA) was developed to analyse such distributional data (Diday, 1989, Billard and Diday, 2003, Billard, 2011, Beranger et al., 2018). However, with a few exceptions (discussed below), the parameters of existing SDA-based methods undesirably lose their interpretation as parameters of models of the underlying (standard) micro-data, which means that the resulting inferences are not directly comparable to the equivalent standard full-data analysis (Zhang et al., 2019, Beranger et al., 2018).

Logistic regression (e.g. Cox, 1958, Hosmer et al., 2013) is one method of performing regression for categorical response data that has been utilised extensively in many fields, including finance (Hauser and Booth, 2011, Hyunjoon and Zheng, 2010), epidemiology (Merlo et al., 2006), medicine (Min, 2013, Hosmer et al., 2013), diagnostics (Knottnerus, 1992) and modelling income (Pavlopoulos et al., 2010). Such models are typically fitted by numerical maximum likelihood estimation as there is no closed form estimator, given certain conditions on the response and predictors such that the maximum likelihood estimate (MLE) exists (Albert and Anderson, 1984). As a result, the computational overheads for determining the MLE can be high for very large datasets

With this motivation, Wang et al. (2018) developed a data-subsampling scheme for binary

logistic regression (for $K = 2$ classes) whereby each observation is assigned a weight according to a function that minimises the asymptotic mean squared error (MSE) of the MLE. A subsample is then taken using those weights, with parameter estimates then obtained by maximising the likelihood for the subsample. To calculate the observation weights, an initial parameter estimate is obtained using r_0 uniformly sampled observations. Wang et al. (2018) demonstrated good estimator performance from the 'optimal' subsample using $r_0 = 1\,000$.

In contrast, de Souza et al. (2011) presented SDA-motivated versions of logistic regression for interval-valued data (where the intervals are constructed from the minimum and maximum observed values of each predictor in the micro-data), whereby the regressions are constructed using the interval centres or endpoints as covariates. de Souza et al. (2008) developed a similar approach for multi-class classification with logistic regression using interval endpoints. While these SDA methods are computationally simple, because they are only based on random intervals (that is, two quantiles of the data) much of the distributional information in the predictor values is lost. In addition the implementations treat each interval equally, which can cause problems if there are unbalanced numbers of micro-data in each category. Finally, as discussed above, the fitted models cannot be compared to the equivalent models fitted to the full (non-summarised) dataset.

In other work, Tranmer and Steel (1997) investigated the effect of data aggregation on logistic regression when both the predictor and response variables are observed as the total sum of each variable for each group. Bhowmik et al. (2016) proposed an iterative algorithm for estimating Generalised Linear Models (GLMs) when the response variables are aggregated into histogram-valued random variables. The resulting training error was numerically shown to approach that of the full (non-aggregated) analysis as the number of histogram bins became large. Armstrong (1985) considered the case where a single covariate is measured with error, and the distribution of the coarsened (binned) value given the true latent observation is known. Estimates for GLM coefficients are then obtained through the utilisation of this known density. Johnson (2006) assumed Gaussian distributions for coarsened covariates and included their likelihood within the GLM framework. Lipsitz et al. (2004) proposed a method for implementing a GLM when one of the covariates is coarsened for only a subset of observations, whereby the resulting likelihood is the integral over the likelihood of all possible values that this variable could have taken, weighted by a density dependent on the uncoarsened, fully observed variables. Johnson and Wiest (2014) proposed a Bayesian approach for coarsened covariates in GLMs whereby a distribution is assumed for the coarsened covariates, given the uncoarsened data.

The above models work well if the distribution of the latent data given the observed data is well-specified, but run into problems if the distributional assumptions are violated. Further, if many covariates are provided in distributional form, these approaches require large computational overheads due to the curse of dimensionality. To the best of our knowledge, little progress has been made in developing logistic regression models for fully histogram-valued predictor variables, which provide far more insight into the shape of the underlying covariate data than the interval case. Here, our primary motivation is in histograms constructed from very large datasets by the analyst in order to facilitate increased computational speed of an analysis. In

this setting, data could arrive in the form of univariate histograms for each predictor for each of the K categories, instead of large $(N \times (K+1))$-dimensional tables (where N, the number of observed predictor and class label pairs, is very large), allowing savings in data transmission, storage and analysis.

In this article we develop methods for implementing logistic regression models with K response categories, where the covariate data for each response category is in the form of marginal or multivariate histograms (or random rectangles, as a special case of random histograms with a single bin). The basic component of our approach adopts the likelihood-based SDA construction of Beranger et al. (2018) (see also Zhang et al., 2019, Whitaker et al., 2019), which unlike other SDA-based methods, directly fits models for the underlying micro-data given the distributional-based data summaries.

We propose a mixed model in which the underlying data is analysed using a mixture of standard and SDA likelihoods: histogram bins with low counts are discarded and the residing micro-data are used directly to contribute standard likelihood terms, while bins with high counts contribute via the SDA framework. However, due to the computational difficulties associated with the potentially high-dimensional integrations required within the SDA likelihood, this mixed model is unsuitable for moderate numbers of predictor variables. For this reason, along with the benefits accompanying the potential additional savings in data storage and computational overheads, models based on lower-dimensional marginal histograms are developed.

We develop an approximate marginal composite likelihood approach to obtain estimates for the parameter vector of the complete (micro-data) model, using lower-dimensional marginal histograms of the observed covariate data. Univariate and bivariate histograms are considered for these models, with the predictions performed on test datasets shown to be comparable with the full standard likelihood models, but at a much lower computational cost. The resulting parameter estimates are directly comparable to the parameter estimates of the standard full-data analysis, enabling both micro-data and histogram-based predictions, as required.

This article is structured as follows. In Section 4.2 we outline two types of established logistic regression models for regular data. In Section 4.3 we briefly outline the general SDA-based likelihood construction of Beranger et al. (2018), and introduce the mixed standard/SDA likelihood model. We also develop the approximate composite likelihood approach based on lower-dimensional marginal histograms. In Section 4.4 we perform various simulation studies to demonstrate the effectiveness of each model. We show that each model is able to produce comparable prediction accuracy compared to the full model at a superior computational cost for a certain sample size, N. We also show that our method performs comparably with the recently developed optimal subsampling method for binary logistic regression by Wang et al. (2018), at a cheaper computational cost. In Section 4.5 we analyse satellite crop-prediction data from Queensland, Australia, and simulated particle collision dataset from the Machine Learning Repository (Dua and Graff, 2017), and show that our approach is highly competitive with much more computationally expensive standard statistical methods. We conclude with a discussion.

4.2 Logistic regression methods for classification

There are a variety of methods for classifying an instance into one of $K \geq 2$ classes. Let Y denote a discrete random variable taking a value in $\Omega = \{1, \ldots, K\}$ and $X \in \mathbb{R}^D$ an associated vector of explanatory variables. Using the information contained in the covariates X, the aim is to estimate the probability of the outcome of $Y = k$ for $k \in \Omega$.

Much work has been done on the comparison of the performance of various classification algorithms. Gladence et al. (2015) describe logistic regression along with various Bayes classification methods such as naive Bayes, multinomial naive Bayes, Bayes networks and updatable naive Bayes. The performance of these models are compared using five real datasets, and it is shown that logistic regression generally gives superior results to the different Bayes methods in terms of various metrics such as Precision, Recall, Mean Squared Error and Average Mean Absolute error. Kiang (2003) use synthetic examples to compare the performance of several well known statistical classification techniques, such as logistic regression, neural networks, kth nearest neighbour, discriminant analysis and decision trees, with a particular focus on the effect of violating several model assumptions, such as unimodality, non-collinearity, normality and low-correlations. The performance of each model is assessed using the misclassification rate on synthetic datasets that violate each of these model assumptions, with the authors determining that logistic regression and neural networks generally provide superior results compared to the other methods under most scenarios. It is worth noting that for multimodal data, neural networks significantly outperformed logistic regression (and the other methods).

The efficacy of logistic regression compared to other classification methods has similarly been investigated in the context of specific applications. For example, Cigsar and Unal (2019) compare the performance of logistic regression, Bayes networks, naive Bayes, random forests and multilayer perceptron models in the prediction of default by banking customers using a real dataset from the Turkish Statistical institute containing demographic and socioeconomic properties of individual customers, such as age, gender, revenue, health, bills, education, etc. The performance of each model is assessed using prediction accuracy, root mean squared error, ROC area, precision and recall. Logistic regression provided superior results for all these metrics with the exception of precision, for which it was slightly outperformed by Bayes networks and naive Bayes. This leads the authors to conclude that logistic regression is the best algorithm for analysing default risk within this dataset. Liu et al. (2011) compare the performance of logistic regression, regression trees and neural networks models in the prediction of violent reoffending using a UK dataset comprising of data from 1 225 UK male prisoners. Cross-validation was performed using four subsets of the data, with prediction accuracy and AUC used as the main indicators of performance success for each model. The authors concluded that for this dataset, neural networks slightly outperformed logistic regression and regression trees, however this increase in performance was deemed to be not significant. Given the benefits and drawbacks associated with each model, the authors recommend that the appropriateness of each model is specific to data characteristics and the aims of the study.

Here we focus on logistic regression, a widely used statistical modelling technique for bi-

nary ($K = 2$) dependent variables. Multinomial logistic regression is a generalisation of logistic regression to problems with possible outcomes taking values in Ω ($K \geq 2$). An alternative problem representation recasts multinomial classification as multiple binary classification problems. One-vs-Rest (OvR) logistic regression implements a separate binary logistic regression for each class $k \in \Omega$, assuming that each classification model is independent. To establish notation, we briefly describe both multinomial and OvR logistic regression methods below.

4.2.1 Multinomial logistic regression

For each outcome $k \in \Omega \backslash \{K\}$, the multinomial logistic regression model defines the log pairwise odds ratios through a linear model

$$\log\left(\frac{P_{\mathrm{M}}(Y = k|X)}{P_{\mathrm{M}}(Y = K|X)}\right) = \beta_{k0} + \beta_k^\top X, \tag{4.1}$$

where $P_{\mathrm{M}}(Y = k|X)$ denotes the probability that $Y = k$, under the multinomial model (M), when X is observed, $\beta_{k0} \in \mathbb{R}$ is an intercept and $\beta_k = (\beta_{k1}, \ldots, \beta_{kD})^\top \in \mathbb{R}^D$ represents the vector of regression coefficients associated with the D explanatory variables and the outcome k. The outcome K is referred to as the pivot or reference category and its corresponding parameter $(\beta_{K0}, \beta_K^\top)^\top$ is the zero vector. Thus (4.1) can be rearranged as

$$P_{\mathrm{M}}(Y = k|X) = \frac{e^{\beta_{k0} + \beta_k^\top X}}{1 + \sum_{j \in \Omega \backslash \{K\}} e^{\beta_{j0} + \beta_j^\top X}}$$

for all $k \in \Omega$. This implies that the odds of preferring one class over another do not depend on the presence or absence of other "irrelevant" alternatives.

Suppose that $\boldsymbol{X} = (X_1, ..., X_N)$ is a vector of D-dimensional random vectors and $Y = (Y_1, ..., Y_N)^\top$ is a vector of discrete random variables, with respective realisations $\boldsymbol{x} \in \mathbb{R}^{D \times N}$ and $y \in \Omega^N$. The likelihood function under the multinomial model is given by

$$L_{\mathrm{M}}(\boldsymbol{x}, y; \boldsymbol{\beta}) = \prod_{n=1}^{N} \prod_{k \in \Omega} P_{\mathrm{M}}(Y = k|X = x_n)^{\mathbb{1}\{y_n = k\}}, \tag{4.2}$$

where $\mathbb{1}\{\cdot\}$ represents the indicator function, and $\boldsymbol{\beta} = (\check{\beta}_1, \ldots, \check{\beta}_K) \in \mathbb{R}^{(D+1) \times K}$ with $\check{\beta}_k = (\beta_{k0}, \beta_k^\top)^\top \in \mathbb{R}^{D+1}$. We denote the maximun likelihood estimator for the multinomial logistic regression model as $\hat{\boldsymbol{\beta}}^{\mathrm{M}} = \operatorname{argmax}_{\boldsymbol{\beta}} \log L_{\mathrm{M}}(\boldsymbol{x}, y; \boldsymbol{\beta})$. Existence of the MLE can be examined through the concept of data separation.

Definition 4.2.1. (Multinomial model separation) There is complete separation of the data if for all $k \in \Omega$, a $\boldsymbol{b} = (b_1, \ldots, b_K)$, $b_k \in \mathbb{R}^D$, exists such that

$$(b_k - b_j)^\top x_n > 0 \text{ for all } n \text{ such that } y_n = k, j \neq k$$
$$(b_k - b_j)^\top x_n < 0 \text{ for all } n \text{ such that } y_n \neq k, j \neq k.$$

There is quasi-complete separation of the data if for all $k \in \Omega$, a $\boldsymbol{b} = (b_1, \ldots, b_K)$, $b_k \in \mathbb{R}^D$, exists such that

$$(b_k - b_j)^\top x_n \geq 0 \text{ for all } n \text{ such that } y_n = k, j \neq k$$
$$(b_k - b_j)^\top x_n \leq 0 \text{ for all } n \text{ such that } y_n \neq k, j \neq k,$$

with equality for at least one observation x_n in each class k.

Albert and Anderson (1984) proved that if there is neither complete nor quasi-complete separation in the data, then the MLE $\hat{\boldsymbol{\beta}}^{\mathrm{M}}$ exists for all classes $k \in \Omega$.

4.2.2 One-vs-Rest logistic regression

For each outcome $k \in \Omega$, the One-vs-Rest logistic regression model defines the log odds ratio through a linear model

$$\log\left(\frac{P_{\mathrm{O}}(Y = k|X)}{P_{\mathrm{O}}(Y \neq k|X)}\right) = \beta_{k0} + \beta_k^\top X,$$

where $P_{\mathrm{O}}(Y = k|X)$ denotes the probability that $Y = k$, under the OvR regression model (O), when X is observed and where β_{k0} and β_k are as defined previously. This ratio can be rearranged as

$$P_{\mathrm{O}}(Y = k|X) = \frac{e^{\beta_{k0} + \beta_k^\top X}}{1 + e^{\beta_{k0} + \beta_k^\top X}}$$

for all $k \in \Omega$. Note that here β_K is different from the zero vector as each individual binary model has an implied reference category and $\sum_{k\in\Omega} P_{\mathrm{O}}(Y = k|X) \neq 1$. The likelihood function can be written as

$$L_{\mathrm{O}}(\boldsymbol{x}, y; \boldsymbol{\beta}) = \prod_{n=1}^{N} \left(P_{\mathrm{O}}(Y = y_n|X = x_n) \prod_{k\in\Omega\setminus\{y_n\}} P_{\mathrm{O}}(Y \neq k|X = x_n) \right), \tag{4.3}$$

which is expressed as the product of K binary logistic regressions for each observation (x_n, y_n). The parameters of the multivariate and OvR regression models are not directly comparable, but the performance of each model can be assessed by evaluating their prediction accuracy on a training dataset (e.g. Eichelberger and Sheng, 2013).

As before, the MLE under the OvR model, $\hat{\boldsymbol{\beta}}^{\mathrm{O}} = \mathrm{argmax}_{\boldsymbol{\beta}} \log L_{\mathrm{O}}(\boldsymbol{x}, y; \boldsymbol{\beta})$, exists for all classes $k \in \Omega$ if there is neither complete nor quasi-complete separation of the data for each $k \in \Omega$, but under slightly modified definitions of separation compared to the multinomial model (Albert and Anderson, 1984).

Definition 4.2.2. (OvR model separation) There is complete separation of the data for the k^{th} class if a $b_k \in \mathbb{R}^D$ exists such that

$$b_k^\top x_n > 0 \text{ for all } n \text{ such that } y_n = k$$
$$b_k^\top x_n < 0 \text{ for all } n \text{ such that } y_n \neq k.$$

There is quasi-complete separation of the data for the k^{th} class if a $b_k \in \mathbb{R}^D$ exists such that

$$b_k^\top x_n \geq 0 \text{ for all } n \text{ such that } y_n = k$$
$$b_k^\top x_n \leq 0 \text{ for all } n \text{ such that } y_n \neq k,$$

with equality for at least one observation x_n in the k^{th} class, and at least one observation $x_{n'}$, $n' \neq n$, not in the k^{th} class.

4.2.3 Existing methods for large-sample logistic regression

If the practitioner wishes to use a Bayesian approach to fit a logistic regression model, much work has been done on the use of subsampling to deal with the large computational burden associated with huge datasets. MacLaurin and Adams (2014) present an algorithm that introduces a Bernoulli random variable for each observation that selects the observations to be included in the sample during each iteration of the Markov Chain Monte Carlo (MCMC) algorithm. This algorithm then only requires the evaluation of the likelihood of a subset of the complete dataset, but is still able to simulate from the exact posterior distribution. Logistic regression is used as an example that demonstrates the utility of the algorithm, with comparable results to the complete MCMC obtained at a cheaper computational cost. Quiroz et al. (2019) assume that the log-likelihood follows a Gaussian distribution, allowing the use of the Central Limit Theorem to obtain approximations of the log-likelihood. A bias-correction formula is then included to account for the bias exhibited by the resultant likelihood estimator. Logistic regression is used as one of the models for the synthetic examples, and the algorithm is shown to perform well compared to other recent MCMC subsampling methodologies. Gunawan et al. (2019) greatly speed up the sequential Monte Carlo algorithm by using data subsampling to approximate the target likelihood and show that this algorithm provides theoretical guarantees of fast convergence of the posterior distribution with increasing sample size. Logistic regression is used as an example in the simulation studies to demonstrate the utility of their method.

Rather than approximating the likelihood, Bardenet et al. (2014) propose to approximate the accept/reject step of each iteration of the Metropolis Hastings (MH) algorithm using a randomly sampled subset of the full dataset. By observing the acceptance ratio for a subsample of the data, upper and lower bounds can be obtained for the complete ratio. Guidelines are also given that allow the determination of a sufficient subsample size that ensures a given probability of correctness for the resultant accept/reject decision. Simulation studies are used to demonstrate that good results can be achieved at only a fraction of the computational cost, compared to the complete algorithm, with logistic regression used as a motivating example. Bardenet et al. (2017) improve upon the algorithm derived in Bardenet et al. (2014) by using a control variate approach to reduce the error associated with the bounds on the acceptance ratio. It is proven that this improvement leads to significant computational gains, with logistic regression again used as a demonstrative example in synthetic examples and a real data analysis. Bierkens et al. (2019) introduce a new Bayesian sampling algorithm (termed the Zig-Zag sampling algorithm)

that is easily estimated using sub-sampling, allowing for huge improvements in computation for large datasets. The strong regularity conditions for the model being investigate that the Zig-Zag sampling algorithm requires have been established for logistic regression, for which simulation studies are used to demonstrate its utility. Comparably low MSEs are obtained for the logistic regression model in various simulation studies, demonstrating the utility of this algorithm.

Subsampling has also been investigated as a method of dealing with the computational issues associated with fitting frequentist logistic regression models to large datasets. Owen (2007) derive asymptotic results for logistic regression models where one class is extremely rare, and show how under mild conditions this setting can lead to computational savings in the estimation of the model parameters. Fithian and Hasan (2014) investigate the problem of subsampling from large datasets in which there is a significant class imbalance, and derive a subsampling scheme in which the resultant subsample has a more balanced proportion of response classes. The general idea is to iteratively retain the observations for which the response y_n is considered surprising, given the covariate x_n, allowing for the most informative observations to comprise the subsample. Wang et al. (2018) derive a two-step subsampling scheme for logistic regression that minimises the asymptotic MSE of the resultant estimator. An initial parameter estimate is obtained from an initial uniformly sampled subsample, and then used to assign each observation a weight. A second subsample is then obtained using those weights, with the final parameter estimate obtained from a combination of the two samples. It was demonstrated via synthetic examples that for balanced datasets, this methodology obtained higher predictions and lower MSEs than that of Fithian and Hasan (2014), and as such is considered in the simulation studies in Section 4.4 as the ideal comparison for the models presented in this paper.

4.3 Classification for aggregated data

For each class $k \in \Omega$, define $\boldsymbol{X}^{(k)} = (X_n | Y_n = k, n = 1, \ldots, N) \in \mathbb{R}^{D \times N_k}$, where $N_k = \sum_{n=1}^{N} \mathbb{1}\{Y_n = k\}$, as the matrix of N_k covariate vectors associated with outcome k. In this way $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)}$ partition the full covariate set $\boldsymbol{X}$. When N is very large and $|\Omega| \ll N$ then it can be expected than one or more of the N_k will also be large. In this context, directly optimising likelihood functions for logistic regression models could be computationally prohibitive. As an alternative, it might be appealing to aggregate the information contained in $\boldsymbol{X}^{(k)}$ into distributional form (such as a histogram) and to implement a classification algorithm using these summaries only, which could be much more computationally efficient. The concept of performing statistical or inferential analyses on such distributional "datapoints" originated from Diday (1989), and has become known as symbolic data analysis, where the 'symbol' corresponds to the distributional summary (see also Billard and Diday, 2003, 2006, Bock and Diday, 2000).

A symbolic random variable $S_k \in \mathcal{D}_{S_k}$ can be viewed as the result of applying an aggregation function $\pi(\cdot)$ to $\boldsymbol{X}^{(k)} \in \mathcal{D}_{\boldsymbol{X}^{(k)}}$ (where $\mathcal{D}_{\boldsymbol{X}^{(k)}} = \mathbb{R}^{D \times N_k}$), i.e. $S_k = \pi(\boldsymbol{X}^{(k)}) : \mathcal{D}_{\boldsymbol{X}^{(k)}} \to \mathcal{D}_{S_k}$ so that $\boldsymbol{x}^{(k)} \mapsto s_k$. In the present context s_k corresponds to a vector of counts of the number of covariate vectors in $\boldsymbol{x}^{(k)}$ that reside in each histogram bin (see Section 4.3.1 below for more explicit detail). Various likelihood-based techniques for fitting statistical models given the information

in the distributional summaries have been developed (Le Rademacher and Billard, 2011, Brito and Silva, 2012, Lin et al., 2017, Beranger et al., 2018, Zhang et al., 2019). Here we follow the construction of Beranger et al. (2018) and Zhang et al. (2019) who fully model the construction of the symbols from the generating process of the standard random variables $\boldsymbol{X}^{(k)}$. Specifically, the likelihood of observing s_k is

$$L(s_k;\theta,\vartheta) \propto \int_{\mathcal{D}_{\boldsymbol{X}^{(k)}}} f_{S_k|\boldsymbol{X}^{(k)}=\boldsymbol{x}^{(k)}}\left(s_k|\boldsymbol{x}^{(k)},\vartheta\right) g_{\boldsymbol{X}^{(k)}}\left(\boldsymbol{x}^{(k)};\theta\right) \mathrm{d}\boldsymbol{x}^{(k)}, \tag{4.4}$$

where $f_{S_k|\boldsymbol{X}^{(k)}}(\,\cdot\,;\vartheta)$ is the conditional density of S_k given $\boldsymbol{X}^{(k)}$ relating to the aggregation of $\boldsymbol{x}^{(k)} \mapsto s_k$, $g_{\boldsymbol{X}^{(k)}}(\boldsymbol{x}^{(k)};\theta)$ is the standard likelihood function of the model at the data level with parameter of interest θ, and $\boldsymbol{x}^{(k)} = \left(x_1^{(k)},\ldots,x_{N_k}^{(k)}\right)$, with $x_n^{(k)} \in \mathbb{R}^D$ denoting the covariate vector of the n-th observation with outcome k. The likelihood (4.4) is a general expression, for which the density $f_{S_k|\boldsymbol{X}^{(k)}}(\,\cdot\,;\vartheta)$ takes different forms depending on the type of distributional summary considered (see Beranger et al., 2018, for several examples using random intervals/rectangles and random histograms). In the following we are interested in aggregating the covariates $\boldsymbol{X}^{(k)}$ that have the same outcome k into histograms (with fixed or random bins), S_k, and to fit logistic regression type models ($g_{\boldsymbol{X}^{(k)}}\left(\boldsymbol{x}^{(k)};\theta\right)$).

4.3.1 Logistic regressions using histogram-valued data

For each class $k \in \Omega$, suppose that the d-th margin of $\mathbb{R}^D$ is partitioned into B_k^d bins, so that $B_k^1 \times \ldots \times B_k^D$ bins are created in $\mathbb{R}^D$ through the D-dimensional intersections of each marginal bin. Index each bin by $\boldsymbol{b}_k = (b_{1_k},\ldots,b_{D_k})$, $b_{d_k} = 1,\ldots,B_k^d$, as the D-dimensional vector of co-ordinates of each bin in the histogram. The bin $\boldsymbol{b}_k$ is constructed over the space $\boldsymbol{\Upsilon}_{\boldsymbol{b}_k} = \Upsilon^1_{\boldsymbol{b}_k} \times \cdots \times \Upsilon^D_{\boldsymbol{b}_k} \subset \mathbb{R}^D$, where $\Upsilon^d_{\boldsymbol{b}_k} = (y^d_{b_d-1}, y^d_{b_d}] \subset \mathbb{R}$ is a univariate bin in the d-th margin, and where, for each margin d, $-\infty < y_0^d < y_1^d < \ldots < y_{B^k}^d < \infty$ are fixed points that define the change from one bin to the next. The index k has been omitted in the above bin delimitations in order not obscure notation any further, but it needs to be kept in mind that these are specific to the outcome $k \in \Omega$.

Let S_k represent a D-dimensional histogram constructed from $\boldsymbol{X}^{(k)}$ through the aggregation function π where

$$\begin{aligned} S_k = \pi(\boldsymbol{X}^{(k)}) : \mathbb{R}^{N_k \times D} &\to \{0,\ldots,N_k\}^{B_k^1 \times \ldots \times B_k^D} \\ \boldsymbol{x}^{(k)} &\mapsto s_k = \left(s_{\mathbf{1}_k} = \sum_{n=1}^{N_k} \mathbb{1}\{x_n^{(k)} \in \boldsymbol{\Upsilon}_{\mathbf{1}_k}\}, \ldots, s_{\boldsymbol{B}_k} = \sum_{n=1}^{N_k} \mathbb{1}\{x_n^{(k)} \in \boldsymbol{\Upsilon}_{\boldsymbol{B}_k}\} \right). \end{aligned} \tag{4.5}$$

The quantity $S_{\boldsymbol{b}_k}$ denotes the random number of observed data points $X_1^{(k)},\ldots,X_{N_k}^{(k)}$ that fall in the bin indexed by $\boldsymbol{b}_k$. Consequently, the histogram-valued random variable $S_k = (S_{\mathbf{1}_k},\ldots,S_{\boldsymbol{B}_k})$ represents the full $(B_k^1 \times \ldots \times B_k^D)$-dimensional vector of counts from the first bin $\mathbf{1}_k = (1,\ldots,1)$ to the last bin $\boldsymbol{B}_k = (B_k^1,\ldots,B_k^D)$. In this manner, we can construct the collection of histograms $\boldsymbol{S} = (S_1,\ldots,S_K)$, with one S_k for each outcome index $k \in \Omega$, that summarise the information

contained in $\boldsymbol{X} = (\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)})$.

Proposition 4.3.1. *Suppose that $\boldsymbol{X}^{(k)}$, the covariates associated with each outcome $k \in \Omega$, are aggregated via (4.5), and let $\boldsymbol{S} = (S_1, \ldots, S_K)$ denote the resulting collection of histograms. For this summarised data $\boldsymbol{S}$, using (4.4), the likelihood functions for the multinomial (4.2) and OvR (4.3) logistic regression models become*

$$L_{SM}(\boldsymbol{s};\boldsymbol{\beta}) \propto \prod_{k\in\Omega} \prod_{\boldsymbol{b}_k=\boldsymbol{1}_k}^{\boldsymbol{B}_k} \left(\int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}} P_M(Y = k|X = x)\mathrm{d}x \right)^{s_{\boldsymbol{b}_k}} \tag{4.6}$$

$$L_{SO}(\boldsymbol{s};\boldsymbol{\beta}) \propto \prod_{k\in\Omega} \prod_{\boldsymbol{b}_k=\boldsymbol{1}_k}^{\boldsymbol{B}_k} \left(\int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}} P_O(Y = k|X = x)\mathrm{d}x \prod_{k'\in\Omega\setminus\{k\}} \int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}} P_O(Y \neq k'|X = x)\mathrm{d}x \right)^{s_{\boldsymbol{b}_k}}, \tag{4.7}$$

where $\boldsymbol{s}$ is the observed value of $\boldsymbol{S}$, and $s_{\boldsymbol{b}_k}$ denotes the number of observations in bin $\boldsymbol{b}_k$ in histogram k. For a derivation see Appendix A.1.1.

We refer to these models as the symbolic multinomial (SM) and symbolic One-vs-Rest (SOvR) logistic models. In effect, the uncertainty of the location of each predictor X is averaged uniformly over its location in the histogram bin in which it resides. Note that the likelihood functions (4.6) and (4.7) only implicitly depend on the vector of outcomes y since for each possible outcome $k \in \Omega$ the covariates $\boldsymbol{X}^{(k)}$ are summarised in a histogram S_k, and so the product of the N (Y_n, X_n) observations in the standard likelihoods ((4.2) and (4.3)) is replaced by a product over the K outcomes. Further, the parameter ϑ in (4.4) denotes quantities relevant to constructing the symbol (e.g. the number of bins and their locations), and so is fixed in this setting and is therefore omitted in the notation.

Following similar arguments to Heitjan (1989), Beranger et al. (2018), the symbolic likelihoods $L_{\mathrm{SM}}(\boldsymbol{s};\boldsymbol{\beta})$ and $L_{\mathrm{SO}}(\boldsymbol{s};\boldsymbol{\beta})$ can each be shown to approach their classical equivalent, $L_{\mathrm{M}}(\boldsymbol{x}, y;\boldsymbol{\beta})$ and $L_{\mathrm{O}}(\boldsymbol{x}, y;\boldsymbol{\beta})$, as the number of bins in each histogram approaches infinity and the volume of each bin approaches zero. In this scenario, in the limit each bin will either be empty ($s_{\boldsymbol{b}_k} = 0$) or will contain exactly one point ($s_{\boldsymbol{b}_k} = 1$) observed at each value of $x_n^{(k)}$, which then recovers the classical likelihood term. In this manner, the histogram-based likelihoods can be viewed as approximations to the standard likelihood functions for each model.

To establish conditions for the existence of the respective MLEs, $\hat{\boldsymbol{\beta}}^{\mathrm{SM}} = \arg\max_{\boldsymbol{\beta}} \log L_{\mathrm{SM}}(\boldsymbol{s};\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}}^{\mathrm{SO}} = \arg\max_{\boldsymbol{\beta}} \log L_{\mathrm{SO}}(\boldsymbol{s};\boldsymbol{\beta})$, we need to consider modified definitions of complete and quasi-complete separation of the data, in analogy with Definitions 4.2.1 and 4.2.2, to account for the fact that the location of each covariate vector x_n is only known up to the histogram bin in which it resides.

Definition 4.3.1. (Histogram-based multinomial model separation) There is complete separation of the set of histograms $\boldsymbol{s}$ if for all $k \in \Omega$, a $\boldsymbol{b} = (b_1, \ldots, b_K)$, $b_k \in \mathbb{R}^D$, exists such

that

$$(b_k - b_j)^\top x > 0 \text{ for all } x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_k} \text{ such that } s_{\boldsymbol{b}_k} > 0, j \neq k$$
$$(b_k - b_j)^\top x < 0 \text{ for all } x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_{k'}} \text{ such that } s_{\boldsymbol{b}_{k'}} > 0, j \neq k \text{ and } k' \neq k.$$

There is quasi-complete separation of the set of histograms $\boldsymbol{s}$ if for all $k \in \Omega$, a $\boldsymbol{b} = (b_1, \ldots, b_K)$, $b_k \in \mathbb{R}^D$, exists such that

$$(b_k - b_j)^\top x \geq 0 \text{ for all } x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_k} \text{ such that } s_{\boldsymbol{b}_k} > 0, j \neq k$$
$$(b_k - b_j)^\top x \leq 0 \text{ for all } x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_{k'}} \text{ such that } s_{\boldsymbol{b}_{k'}} > 0, j \neq k \text{ and } k' \neq k.$$

with equality for at least one point in the non-empty histogram bins for the k-th class.

Definition 4.3.2. (Histogram-based OvR model separation) There is complete separation of the histogram for the k^{th} class, s_k, if a vector b_k exists such that

$$b_k^\top x > 0 \text{ for all } x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_k} \text{ such that } s_{\boldsymbol{b}_k} > 0$$
$$b_k^\top x < 0 \text{ for all } x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_{k'}} \text{ such that } s_{\boldsymbol{b}_{k'}} > 0, k \neq k'.$$

There is quasi-complete separation of the histogram for the k^{th} class if there exists a vector b_k such that

$$b_k^\top x \geq 0 \text{ for all } x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_k} \text{ such that } s_{\boldsymbol{b}_k} > 0$$
$$b_k^\top x \leq 0 \text{ for all } x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_{k'}} \text{ such that } s_{\boldsymbol{b}_{k'}} > 0, k \neq k'.$$

with equality for at least one point in the non-empty histogram bins for the k-th class, and equality for at least one point in any non-empty histogram bin not in the k-th class.

Proposition 4.3.2. *If the set of histograms $\boldsymbol{s} = (s_1, \ldots, s_K)$ does not exhibit complete or quasi-complete separation as described under Definition 4.3.1, then $L_{SM}(\boldsymbol{s}; \boldsymbol{\beta})$ has a unique global maximum. If the set of histograms does not exhibit complete or quasi-complete separation for any class $k \in \Omega$ as described under Definition 4.3.2, then $L_{SO}(\boldsymbol{s}; \boldsymbol{\beta})$ has a unique global maximum. For a proof see Appendix A.1.2.*

From Definitions 4.2.1–4.3.2 it can be seen that if there is separation of the histograms then there also has to be separation of the underlying data. In this sense, the definitions of complete and quasi-complete separation for random histograms (Definitions 4.3.1 and 4.3.2) are stronger conditions than those for standard random vectors (Definitions 4.2.1 and 4.2.2). As a result, from Proposition 4.3.2 this means that if no histogram-based MLE ($\hat{\boldsymbol{\beta}}^{\text{SM}}$ or $\hat{\boldsymbol{\beta}}^{\text{SO}}$) exists, then the equivalent standard multinomial model MLE ($\hat{\boldsymbol{\beta}}^{\text{M}}$ or $\hat{\boldsymbol{\beta}}^{\text{O}}$) also does not exist. That is, the standard MLE can't exist without the histogram-based MLE also existing. Conversely, however, it is possible to have separation of the underlying data but no separation in the derived histograms. As a result it is possible that this histogram-based MLE exists without the standard

multinomial model MLE existing. (In this particular setting we therefore have the interesting case of the existence of an MLE for a given histogram converging to the non-existence of an MLE in the limit as the number of histogram bins become large while the volume of each bin approaches zero.). We also note that it is also possible for the histogram-based MLEs to exist for one histogram derived from an underlying dataset, but not exist for a different histogram (e.g. with different numbers of and/or locations of bins) derived from the same dataset. Consequently, if the underlying classical data is available to the practitioner, the separation conditions (Definitions 4.2.1 and 4.2.2) should be examined prior to aggregation, to ensure the appropriateness of the multinomial model. Furthermore, if there is complete separation in the underlying microdata, then it would be an interesting future research project to investigate the utility of the histogram approach described above. Potentially binning the data is analogous to adding a small term to the diagonal of an otherwise singular matrix, which might provide some interesting benefits.

Assuming that the MLEs exist, the benefits of using histograms as data summaries for logistic regression modelling are obvious in the presence of large amounts of data. The effective number of likelihood terms in $L_{\mathrm{SM}}(\boldsymbol{s};\boldsymbol{\beta})$ and $L_{\mathrm{SO}}(\boldsymbol{s};\boldsymbol{\beta})$ is the number of histogram bins multiplied by the number of classes. For very large datasets this can be much smaller than the N terms in $L_{\mathrm{M}}(\boldsymbol{x},y;\boldsymbol{\beta})$ and $L_{\mathrm{SO}}(\boldsymbol{x},y;\boldsymbol{\beta})$, and so computing MLEs given the histogram summaries can be much more efficient. The trade off is the loss of some accuracy due to the loss of information in the binned data.

Despite its computational advantages, the above construction has some limitations. We now discuss these and propose some statistical and computational improvements.

4.3.2 Using both classical data and histograms

When constructing histograms, for example using the method described in Section 4.3.1, the number of observations $s_{\boldsymbol{b}_k}$ within each bin $\boldsymbol{b}_k$ will typically vary widely over bins, from very low to very high counts. Where bins have high counts, large computational efficiencies are obtained in the evaluation of $L_{SM}(\boldsymbol{s};\boldsymbol{\beta})$ and $L_{SO}(\boldsymbol{s};\boldsymbol{\beta})$ over the standard logistic regression likelihood functions. However, when a bin has low numbers of underlying data points, it may be that the computational cost in evaluating the bin-specific integrals in (4.6) and (4.7) (such as e.g. $\int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}} P_{\mathrm{M}}(Y=k|X=x)\mathrm{d}x$) is just as high or higher than evaluating the standard likelihood contributions (e.g. $P_{\mathrm{M}}(Y=k|X=x_n)\mathrm{d}x$) for each of the underlying datapoints in that bin. Taken together with the loss of information in moving from the underlying data to a count of datapoints in a bin, in this case it is obviously worse statistically (and perhaps also computationally) to work with the histogram bin rather than the original data in that bin. Creating bins with low data counts becomes more likely as the dimension D of the predictors increases.

To avoid this situation, we introduce a lower bound, $\tau_k \in \{1,\ldots,N_k\}$, on the number of underlying datapoints in a bin region $\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}$ that is required before these data can be summarised into a histogram bin for their contribution to the likelihood function. When the number of underlying datapoints is lower than τ_k, the original data x_n are retained, and contribute to the

likelihood in the standard way.

Under the assumption that the underlying data are available (which may not always be the case), we therefore propose to use the modified aggregation function

$$S_k = \tilde{\pi}(\boldsymbol{X}^{(k)}) : \mathbb{R}^{N_k \times D} \to \{\tau_k, \ldots, N_k\}^u \times \mathbb{R}^{v \times D}$$
$$\boldsymbol{x}^{(k)} \mapsto \left(\begin{cases} s_{\boldsymbol{b}_k} = \sum_{n=1}^{N_k} \mathbb{1}\{x_n^{(k)} \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_k}\} & \text{if } s_{\boldsymbol{b}_k} \geq \tau_k \\ \boldsymbol{x}_{\boldsymbol{b}_k}^{(k)} = \{x_n^{(k)} : x_n^{(k)} \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_k}\} & \text{otherwise} \end{cases}, \boldsymbol{b}_k = \mathbf{1}_k, \ldots, \boldsymbol{B}_k \right),$$

where $\tau_k \in \{1, \ldots, N_k\}$, $u \in [0, \ldots, B_k^1 \times \ldots \times B_k^D]$ is the number of histogram bins containing at least τ_k observations, and $v = N_k - \sum s_{\boldsymbol{b}_k}$ is the number of retained classical datapoints in bins containing less than τ_k observations. That is, the resulting S_k is a mixture of those histogram bins that contain at least τ_k observations, combined with any remaining predictor vectors X_n that would otherwise be put into bins with less than τ_k observations.

In the context of logistic regression modelling, this mixture histogram construction produces likelihood functions that are a mixture of the standard and histogram-based likelihood functions given in Sections 4.2.1, 4.2.2 and 4.3.1. For example, we can construct the likelihood function for a mixture of histogram and classical data under the multinomial logistic regression model as

$$L_{\mathrm{MM}}(\boldsymbol{s}; \boldsymbol{\beta}) \propto \prod_{k \in \Omega} \prod_{\boldsymbol{b}_k = \mathbf{1}_k}^{\boldsymbol{B}_k} \left(\int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}} P_{\mathrm{M}}(Y = k | X = x) \mathrm{d}x \right)^{s_{\boldsymbol{b}_k} \mathbb{1}\{s_{\boldsymbol{b}_k} \geq \tau_k\}} \left(\prod_{x \in \boldsymbol{x}_{\boldsymbol{b}_k}^{(k)}} P_{\mathrm{M}}(Y = k | X = x) \right)^{\mathbb{1}\{s_{\boldsymbol{b}_k} < \tau_k\}}, \tag{4.8}$$

where MM denotes the multinomial mixture. A similar mixture-likelihood, $L_{\mathrm{MO}}(\boldsymbol{s}; \boldsymbol{\beta})$, can be constructed for the OvR logistic regression model.

Because $L_{\mathrm{MM}}(\boldsymbol{s}; \boldsymbol{\beta})$ can be considered as a special case of $L_{\mathrm{SM}}(\boldsymbol{s}; \boldsymbol{\beta})$ (and $L_{\mathrm{MO}}(\boldsymbol{s}; \boldsymbol{\beta})$ a special case of $L_{\mathrm{SO}}(\boldsymbol{s}; \boldsymbol{\beta})$) in which the retained classical data vectors x_n can be thought of as residing in zero-volume bins, one for each retained vector, it is immediate that $L_{\mathrm{MM}}(\boldsymbol{s}; \boldsymbol{\beta}) \to L_M(\boldsymbol{x}, y; \boldsymbol{\beta})$ also approaches that classical data likelihood function as the number of bins becomes large and the volume of each bin approaches zero (similarly $L_{\mathrm{MO}}(\boldsymbol{s}; \boldsymbol{\beta}) \to L_O(\boldsymbol{x}, y; \boldsymbol{\beta})$).

Similarly, by considering the obvious definition of complete and quasi-complete separation for the mixture of histogram and retained classical data vectors as a combination of those in Definitions 4.2.1 and 4.3.1 (for the standard multinomial regression model) and Definitions 4.2.2 and 4.3.1 (for the OvR regression model), similar statements to Proposition 4.3.2 about the existence of the MLEs $\hat{\boldsymbol{\beta}}^{\mathrm{MM}} = \arg\max_{\boldsymbol{\beta}} \log L_{\mathrm{MM}}(\boldsymbol{s}; \boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}}^{\mathrm{MO}} = \arg\max_{\boldsymbol{\beta}} \log L_{\mathrm{MO}}(\boldsymbol{s}; \boldsymbol{\beta})$ can be made. For example, if no full-histogram MLE exists ($\hat{\boldsymbol{\beta}}^{\mathrm{SM}}$ or $\hat{\boldsymbol{\beta}}^{\mathrm{SO}}$) then no mixture-likelihood MLE exists ($\hat{\boldsymbol{\beta}}^{\mathrm{MM}}$ or $\hat{\boldsymbol{\beta}}^{\mathrm{MO}}$) and no standard likelihood MLE ($\hat{\boldsymbol{\beta}}^{\mathrm{M}}$ or $\hat{\boldsymbol{\beta}}^{\mathrm{O}}$) exists. That is, the standard MLE can't exist without the mixture-likelihood MLE existing, which can't itself exist without the full-histogram MLE existing. However, the full-histogram model MLE can exist without the mixture-likelihood MLE existing, and the mixture-likelihood MLE can exist without the standard MLE existing.

The choice of τ_k, for all $k \in \Omega$, controls the tradeoff between computational efficiency and

information loss. On one hand, if τ_k is too large then we face the original issue of having a huge number of terms slowing down evaluation of the likelihood function. On the other hand if τ_k is too low then we risk a loss of efficiency (and perhaps higher computation) compared to higher τ_k. As a result, one option is to set τ_k to be the value such that integrating e.g. P_{M} over a bin $\boldsymbol{b}_k$ is less computationally expensive than evaluating it τ_k times. Some strategies along these lines are explored in the simulations in Section 4.4.

4.3.3 Composite likelihoods for logistic regression models

Mixing histogram and micro data can lead to substantial statistical efficiency improvements, but it does not address the issue of grid-based multivariate histograms becoming highly inefficient as data summaries as the number of covariates (D) increases. In particular, the integrals required to compute the likelihood function $L_{SM}(\boldsymbol{s};\boldsymbol{\beta})$ (4.6) have no analytical solution when the outcome has more than two possible classes ($K > 2$) and there are more than two explanatory variables ($D > 2$). Similarly the integrals in the likelihood function $L_{SO}(\boldsymbol{s};\boldsymbol{\beta})$ (4.7) have no analytical solution when more than two explanatory variables are considered ($D > 2$). In all non-trivial settings, then, these integrals must be computed numerically. This can be computationally costly when D is large, which can then defeat the purpose (i.e. improved computational efficiency) of using data aggregates.

To circumvent the issue of computing the probabilities of data falling in high-dimensional bins, Whitaker et al. (2019) proposed implementing a composite likelihood approach. This consisted of approximating the likelihood function of a high-dimensional histogram by the weighted product of likelihood functions for lower-dimensional *marginal* histograms, which yielded asymptotically consistent likelihood-based parameter estimates (Lindsay, 1988, Varin et al., 2011). Assuming all weights are equal for simplicity, a j-wise composite likelihood function can be expressed as $L^{(j)}(\theta) \propto \prod_{i=1}^{m} L_i(\theta)$, where $L_i(\theta)$ is the likelihood function of one of m j-wise marginal events for a given parameter vector θ. In the current context, $L_i(\theta)$ corresponds to a likelihood contribution based on the subset of covariates represented by a j-dimensional marginal histogram, $\theta = \boldsymbol{\beta}$ and $m = \binom{D+1}{j}$.

However, omitting an important variable in probit and logistic regression analyses will depress the estimated vector of the remaining coefficients towards zero (Wooldridge, 2002, Cramer, 2007). It is therefore non-viable to directly apply a composite likelihood approach to logistic regression problems. However, in the OvR setting and under the assumption that all predictors are independent, Cramer (2007) showed that the non-omitted coefficients of a logistic regression can be written as functions of the regression coefficients in the scenario that no regressor is omitted. This result was primarily aimed at highlighting the effect of omitting variables in a regression analysis context, and had no practical use for e.g. compensating for zero-depressed parameter estimates, since the established correspondences required information about the variances of the omitted variables, which were unavailable. However, in the composite likelihood setting such information about each covariate is available, and the result of Cramer (2007) can therefore be implemented within each marginal likelihood contribution to compensate for the

covariates that are omitted in that term. We implement this concept in Proposition 4.3.3 below.

In the remainder of this section we construct composite likelihoods for the OvR and histogram-based OvR logistic regression model (the results of Cramer, 2007, do not hold for multinomial logistic regression). Let $\boldsymbol{i} = (i_1, \ldots, i_I) \subseteq \{1, \ldots, D\}$, where for convenience $i_1 < \ldots < i_I$, and define by $\mathcal{I}_j = \{\boldsymbol{i} : |\boldsymbol{i}| = j\}$ the set of all j-dimensional subsets of $\{1, \ldots, D\}$. We adopt the notation that that a vector with superscript $\boldsymbol{i}$ denotes the subvector containing those elements corresponding to the index set $\boldsymbol{i}$. For matrices with the superscript $\boldsymbol{i}$, the operation is replicated column-wise. E.g. for $\boldsymbol{i} \in \mathcal{I}_j$, $\boldsymbol{X}^{(k)\boldsymbol{i}} = (X_1^{(k)\boldsymbol{i}}, \ldots, X_{N_k}^{(k)\boldsymbol{i}}) \in \mathbb{R}^{j \times N_k}$ where $X_n^{(k)\boldsymbol{i}} \in \mathbb{R}^j$ is a subvector of $X_n^{(k)}$, $n = 1, \ldots, N_k$. Then if $S^{\boldsymbol{i}}_{\boldsymbol{b}^{\boldsymbol{i}}_k}$ is the random number of observed data points in $\boldsymbol{X}^{(k)\boldsymbol{i}}$ that fall in bin $\boldsymbol{b}^{\boldsymbol{i}}_k$, we may construct an I-dimensional random *marginal* histogram $\boldsymbol{S}^{\boldsymbol{i}}_k = (S^{\boldsymbol{i}}_{\mathbf{1}^{\boldsymbol{i}}_k}, \ldots, S^{\boldsymbol{i}}_{\boldsymbol{B}^{\boldsymbol{i}}_k})$ as the associated vector of random counts from the first bin $\mathbf{1}^{\boldsymbol{i}}_k = (1, \ldots, 1)$ to the last bin $\boldsymbol{B}^{\boldsymbol{i}}_k = (B^{i_1}_k, \ldots, B^{i_I}_k)$. The vector $\boldsymbol{S}^{\boldsymbol{i}}_k$ has length $B^{i_1}_k \times \ldots \times B^{i_I}_k$ and satisfies $\sum_{\boldsymbol{b}^{\boldsymbol{i}}_k} S^{\boldsymbol{i}}_{\boldsymbol{b}^{\boldsymbol{i}}_k} = N_k$.

The following proposition establishes how to perform approximate composite likelihood estimation for the OvR and histogram-based OvR regressions models using the results in Cramer (2007). As independence between predictors, as assumed by Cramer (2007), is unrealistic, the Proposition also extends the results of Cramer (2007) to account for the correlation between the included set of predictor variables within each composite likelihood term and the omitted variables. Without loss of generality, consider a random vector $X \in \mathbb{R}^D$ and for $\boldsymbol{i} \in \mathcal{I}_j$, let $X^{\boldsymbol{i}} \in \mathbb{R}^j$ represent the observed variables of X. Further let $\mathcal{I}_1^{-\boldsymbol{i}} = \{1, \ldots, D\} \backslash \{\boldsymbol{i}\}$ such that for all $i' \in \mathcal{I}_1^{-\boldsymbol{i}}$, $X^{i'}$ represents an omitted variable of X. Following Cramer (2007) we define the omitted variables $X^{i'}$ to be a linear function of the observed variables via $X^{i'} = \alpha_{\boldsymbol{i}i'}^\top X^{\boldsymbol{i}} + \epsilon_{\boldsymbol{i}i'}$, where $\alpha_{\boldsymbol{i}i'} = (\alpha_{i_1 i'}, \ldots, \alpha_{i_j i'})^\top \in \mathbb{R}^j$ and $\epsilon_{\boldsymbol{i}i'} \sim N(0, \lambda^2_{\boldsymbol{i}i'})$. Denote $\text{Cov}(\epsilon_{\boldsymbol{i}i'_1}, \epsilon_{\boldsymbol{i}i'_2}) = \lambda_{\boldsymbol{i}i'_1 i'_2}$.

Proposition 4.3.3. *The j-wise approximate composite likelihood functions for the standard and the histogram-based D-dimensional OvR logistic regression models are respectively given by*

$$L_O^{(j)}(\boldsymbol{x}, y; \boldsymbol{\beta}) = \prod_{\boldsymbol{i} \in \mathcal{I}_j} L_O(\boldsymbol{x}^{\boldsymbol{i}}, y; \widetilde{\boldsymbol{\beta}}^{\boldsymbol{i}}) \quad \textit{and} \quad L_{SO}^{(j)}(\boldsymbol{s}; \boldsymbol{\beta}) = \prod_{\boldsymbol{i} \in \mathcal{I}_j} L_{SO}(\boldsymbol{s}^{\boldsymbol{i}}, y; \widetilde{\boldsymbol{\beta}}^{\boldsymbol{i}}),$$

where the lower dimensional regression observed coefficients are given by $\widetilde{\boldsymbol{\beta}}^{\boldsymbol{i}} = \left(\widetilde{\beta}^{\boldsymbol{i}}_1, \ldots, \widetilde{\beta}^{\boldsymbol{i}}_K\right) \in \mathbb{R}^{(j+1) \times K}$ where

$$\widetilde{\beta}^{\boldsymbol{i}}_k = \frac{\beta^{\boldsymbol{i}}_k + \left[0, \left(\sum_{i' \in \mathcal{I}_1^{-\boldsymbol{i}}} \beta^{i'}_k \alpha_{\boldsymbol{i}i'}\right)^\top\right]^\top}{\sqrt{1 + \frac{\pi^2}{3} \sum_{i'_1 \in \mathcal{I}_1^{-\boldsymbol{i}}} \left[(\beta^{i'_1}_k \lambda_{\boldsymbol{i}i'_1})^2 + 2 \sum_{i'_2 \in \mathcal{I}_1^{-\boldsymbol{i}}, i'_2 \neq i'_1} \beta^{i'_1}_k \beta^{i'_2}_k \lambda_{\boldsymbol{i}i'_1 i'_2}\right]}} \in \mathbb{R}^{(j+1)}. \tag{4.9}$$

If there is neither complete nor quasi-complete separation in the full D-dimensional dataset $\boldsymbol{x}$ (Definition 4.2.1) for any binary logistic model, then $L_O^{(j)}(\boldsymbol{x}, y; \boldsymbol{\beta})$ will have a unique global maxima. Similarly, if there is neither complete nor quasi-complete separation in any of the sets of marginal histograms $\boldsymbol{s}^{\boldsymbol{i}}$, $\boldsymbol{i} \in I_j$, then $L_{SO}^{(j)}(\boldsymbol{s}; \boldsymbol{\beta})$ will have a unique global maxima. See Appendix A.1.3 for a derivation and proof.

Note that $L_{\mathrm{O}}^{(j)}(\boldsymbol{x}, y; \boldsymbol{\beta})$ and $L_{\mathrm{SO}}^{(j)}(\boldsymbol{s}; \boldsymbol{\beta})$ are *approximate* composite likelihood functions rather than true composite likelihood functions. The resulting maximum composite likelihood estimators ($\hat{\boldsymbol{\beta}}_{\mathrm{O}}^{(j)}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{SO}}^{(j)}$) are not unbiased or consistent. They are, however, reasonable parameter estimates if one is motivated to estimate logistic regression model parameters within the composite likelihood framework (as is the case here), that are more accurate than those estimated via a naive composite likelihood implementation (which we define here as simply $\tilde{L}_{\mathrm{SO}}^{(j)}(\boldsymbol{s}; \boldsymbol{\beta}) = \prod_{\boldsymbol{i} \in \mathcal{I}_j} L_{\mathrm{SO}}(\boldsymbol{s}^{\boldsymbol{i}}, y; \boldsymbol{\beta}^{\boldsymbol{i}})$). We numerically demonstrate the performance of these estimators against the naive implementation in the simulation study in Section 4.4.2. However, when our primary aim is model predictive accuracy, we will demonstrate that the performance of the fitted model using the approximate composite likelihood MLE ($\hat{\boldsymbol{\beta}}_{\mathrm{O}}^{(j)}$ or $\hat{\boldsymbol{\beta}}_{\mathrm{SO}}^{(j)}$) is highly competitive with using the full-data standard MLE, $\hat{\boldsymbol{\beta}}_{\mathrm{O}}$, while, in the case of $\hat{\boldsymbol{\beta}}_{\mathrm{SO}}^{(j)}$, being much more computationally efficient to obtain.

Following similar arguments to before, it is clear that $L_{\mathrm{SO}}^{(j)}(\boldsymbol{x}, y; \boldsymbol{\beta}) \to L_{\mathrm{O}}^{(j)}(\boldsymbol{s}; \boldsymbol{\beta})$ as the number of histogram bins becomes large while the volume of each bin approaches zero. Whitaker et al. (2019) proved the asymptotic normality and consistency of the histogram-based (true) composite likelihood estimator, and highlighted that parameter variance consistency requires the number of bins in each histogram to become large and the number of histograms representing the data be close to N, the number of data points. While these results are not directly applicable under the approximate composite likelihood of Proposition 4.3.3, we intuitively expect that that the estimated variances of $\hat{\boldsymbol{\beta}}_{\mathrm{O}}^{(j)}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{SO}}^{(j)}$ will be inflated compared to that of $\hat{\boldsymbol{\beta}}_{\mathrm{O}}$ unless the same conditions hold.

In specific cases we can obtain a closed-form approximate composite likelihood function for a D-dimensional random histogram, $L_{\mathrm{SO}}(\boldsymbol{s}; \boldsymbol{\beta})$. In the particular case of a binary outcome ($K = 2$) and using $j = 1$ to construct the composite likelihood from all univariate marginal events (the set $\mathcal{I}_1 = \{1, \ldots, D\}$), we have

$$L_{\mathrm{SO}}^{(1)}(\boldsymbol{s}; \boldsymbol{\beta}) = \prod_{i \in \mathcal{I}_1} \prod_{k=1}^{2} \prod_{b_{i_k}=1}^{B_k^i} \left[\left(\frac{1}{\widetilde{\beta}_{k1}^i} \right)^2 \log \left(\frac{1 + e^{\widetilde{\beta}_{k0}^i + \widetilde{\beta}_{k1}^i y_{b_i}^i}}{1 + e^{\widetilde{\beta}_{k0}^i + \widetilde{\beta}_{k1}^i y_{b_i - 1}^i}} \right) \log \left(\frac{1 + e^{-\widetilde{\beta}_{k0}^i - \widetilde{\beta}_{k1}^i y_{b_i - 1}^i}}{1 + e^{-\widetilde{\beta}_{k0}^i - \widetilde{\beta}_{k1}^i y_{b_i}^i}} \right) \right]^{s_{b_k}^i}, \tag{4.10}$$

where $\widetilde{\beta}_k^i = (\widetilde{\beta}_{k0}^i, \widetilde{\beta}_{k1}^i) \in \mathbb{R}^2$ and the bin indexed by $\boldsymbol{b}_k^i = b_{i_k}$ is constructed over the space $\Upsilon_{\boldsymbol{b}_k^i}^i = (y_{b_i - 1}^i, y_{b_i}^i]$. In all other cases, the integrals in $L_{\mathrm{SO}}(\boldsymbol{s}; \boldsymbol{\beta})$ (4.7) require numerical estimation.

Evaluating (4.9) within the approximate composite likelihood requires knowledge of $\alpha_{\boldsymbol{i}i_1'}$, $\lambda_{\boldsymbol{i}i_1'}$ and $\lambda_{\boldsymbol{i}i_1'i_2'}$ for all $i_1', i_2' \in \mathcal{I}_1^{-\boldsymbol{i}}, i_1' \neq i_2'$. The $\alpha_{\boldsymbol{i}i_1'}$ terms are the coefficients explaining the variations of an unobserved variable $X^{i_1'}$ as a linear function of the observed variables $X^{\boldsymbol{i}}$ (i.e. $X^{i'} = \alpha_{\boldsymbol{i}i'}^\top X^{\boldsymbol{i}} + \epsilon_{\boldsymbol{i}i'}$). In the case where $\boldsymbol{i} = i \in \mathcal{I}_1$ (i.e. a simple linear regression with $j = 1$ as in (4.10)), an estimate of $\alpha_{ii_1'}$ is $\mathrm{Cov}(X^i, X^{i_1'})/\mathrm{Var}(X^i)$. In this context, rewriting $\epsilon_{ii_1'} = X^{i_1'} - \alpha_{ii_1'} X^i$, we also have that

$$\hat{\lambda}_{ii_1'}^2 = \mathrm{Var}(X^{i_1'}) - \frac{\mathrm{Cov}(X^i, X^{i_1'})^2}{\mathrm{Var}(X^i)} \quad \text{and} \quad \hat{\lambda}_{ii_1'i_2'} = \mathrm{Cov}(X^{i_1'}, X^{i_2'}) - \frac{\mathrm{Cov}(X^i, X^{i_1'})\,\mathrm{Cov}(X^i, X^{i_2'})}{\mathrm{Var}(X^i)}.$$

Knowledge of variances and covariances between the covariates X is similarly required when two or more predictors are considered in each approximate composite likelihood contribution i.e. when $\boldsymbol{i} \in \mathcal{I}_j$, for $j \geq 2$. Ideally these variances and covariances should be computed and stored prior to the data aggregation process i.e. on the full dataset, but if this information is unavailable, variance and covariance estimates can be derived directly from the histograms either with (Beranger et al., 2018) or without (Billard and Diday, 2003, Billard, 2011) parametric assumptions. Clearly, the assumption that a missing variable can be written as a linear combination of observed variables may not hold. Where viable, transformations (e.g. Box-Cox) or more flexible regression models can be applied to provide a more realistic model, either regressing on the transformed covariates or modifying the form of $\tilde{\beta}_k^{\boldsymbol{i}}$ (4.9) as required.

Finally, suppose that we again consider the case $j = 1$ so that the approximate composite likelihood is constructed with each term comprising each covariate separately (4.10). This is the most computationally efficient histogram-based likelihood as it is based solely on univariate histograms (and known covariances with the other covariates). This construction implies that only univariate marginal histograms are required. Beranger et al. (2018) introduced two likelihood constructions for histogram-valued variables. The first, which we have used until now, assumes that histogram bins are fixed and the corresponding counts are random, which works straightforwardly in D-dimensions. The second construction is a quantile-based approach for univariate variables only, where the bin locations are assumed random and the counts fixed. Defining bin locations using quantiles can better describe the behaviour of the underlying data, and also has the advantage of retaining some of the micro-data (at the observed quantiles) which resembles the mixture of histogram and standard likelihood approach of Section 4.3.2.

In this setting, for each $k \in \Omega$ and each marginal component $i \in \mathcal{I}_1$, define a vector of order statistics $t = (t_1, \ldots, t_B)^\top$ where $1 \leq t_1 \leq \ldots \leq t_B \leq N_k$, such that a quantile-based histogram-valued random variable is obtained through the aggregation function

$$\begin{aligned} S_k^i = \dot{\pi}(\boldsymbol{X}^{(k)i}) : \mathbb{R} \to \mathcal{S} = \{(a_1, \ldots, a_B) \in \mathbb{R}^B : a_1 < \ldots < a_B\} \times \mathbb{N} \\ \boldsymbol{x}^{(k)i} \mapsto s_k^i = \left(\boldsymbol{x}_{(t_1)}^{(k)i}, \ldots, \boldsymbol{x}_{(t_B)}^{(k)i}, N_k\right), \end{aligned} \tag{4.11}$$

where $\boldsymbol{x}_{(t_b)}^{(k)i}$ denotes the t_b-th order statistic of $\boldsymbol{x}^{(k)i} \in \mathbb{R}$. (Note that to ease notation we have omitted superscripts and subscripts related to i and k in the order statistics t.) The b-th histogram bin is then defined over the range $(s_{kb-1}^i, s_{kb}^i]$ with fixed counts of underlying datapoints $t_b - t_{b-1}$, for $b = 1, \ldots, B+1$, where $s_0 = -\infty$, $s_{B+1} = +\infty$, $t_0 = 0$ and $t_{B+1} = N_k + 1$. If each covariate is aggregated via (4.11), then the resulting approximate composite likelihood function is

$$L_{\text{OO}}^{(1)}(\boldsymbol{s}; \boldsymbol{\beta}) = L_{\text{O}}^{(1)}(\{\boldsymbol{x}^{(k)i}\}; \boldsymbol{\beta}) L_{\text{SO}}^{(1)}(\boldsymbol{s}; \boldsymbol{\beta}),$$

where $\boldsymbol{s} = (s_1, \ldots, s_K)$ and $s_k = (s_k^1, \ldots, s_k^D)$ with s_k^i defined in (4.11), and where $L_{\text{SO}}^{(1)}$ is the likelihood shown in (4.10).

4.4 Simulation studies

We now examine the parameter estimation and classification capabilities of the methods developed in Section 4.3 based on simulated data. We consider both the statistical and computational performance of the histogram-based analyses compared to the standard full-data approach.

In the following we set the number of possible outcomes K to define $\Omega = \{1, \ldots, K\}$, the domain of the response variable Y. We obtain the $(D \times 2N)$ matrix of covariates $\boldsymbol{X}$ by generating $2N$ observations from a specified D-dimensional distribution. Given a fixed matrix of regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^{(D+1)\times K}$, for each $n = 1, \ldots, 2N$ we compute the probability of every outcome in Ω using (4.1). These probabilities are then used to generate $\boldsymbol{Y} \in \Omega^{2N}$ from a multinomial distribution. The dataset $(\boldsymbol{Y}, \boldsymbol{X})$ is then split into equally sized training and test datasets, and the estimates $\hat{\boldsymbol{\beta}}^{\mathrm{M}}, \hat{\boldsymbol{\beta}}^{\mathrm{O}}, \hat{\boldsymbol{\beta}}^{\mathrm{MM}}, \hat{\boldsymbol{\beta}}^{\mathrm{O}(j)}$ and $\hat{\boldsymbol{\beta}}^{\mathrm{SO}(j)}$ are obtained by maximising their respective likelihood functions on the training dataset. Using the test dataset we compute the prediction accuracy (PA) of a model (and estimation procedure) as

$$PA = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{Y_n^{\mathrm{Pred}} = Y_n\}$$

where $Y_n^{\mathrm{Pred}} = \mathrm{argmax}_{k\in\Omega} P(Y = k | X = X_n), n = 1, \ldots, N$, denotes the predicted outcome under the model (multinomial or OvR model) with estimated coefficients $\hat{\boldsymbol{\beta}}^{\mathrm{M}}, \hat{\boldsymbol{\beta}}^{\mathrm{O}}, \hat{\boldsymbol{\beta}}^{\mathrm{MM}}, \hat{\boldsymbol{\beta}}^{\mathrm{O}(j)}$ or $\hat{\boldsymbol{\beta}}^{\mathrm{SO}(j)}$. After repeating the above analysis 1 000 times we report the mean squared errors (MSE) and mean prediction accuracies of the above estimators.

In Section 4.3.2 we proposed to improve the statistical and computational efficiency of the MLE by retaining the data underlying a histogram bin, rather than using the bin itself, if the number of underlying observations in the bin $s_{\boldsymbol{b}_k}$ is less than τ_k. Quadrature is utilised in order to perform the integrations in (4.6). In order to perform this approximation, the number of function evaluations used to approximate the integral needs to be specified. A simple method of determining the minimum number of function evaluations required to obtain good results is to iteratively increase the value until the change in results is negligible, thus indicating the approximation of the integral is sufficient. Similar approaches can be utilised to determine the minimum number of marginal bins B needed to obtain comparable results to the classical model for the various histogram-based models described in Section 4.3, whereby increasing values of B are iteratively investigated until the change in results is negligible. This approach is used to indicate convergence to the classical results in the real data analyses in Section 4.5. We note that it would be an interesting future research project to explore potentially more rigorous methods of determining the optimal values of these parameters. Through experimentation we determined that using 2^j function evaluations for each integral globally produced small enough approximation errors when integrating over j-dimensional bins to obtain comparable results to the classical model. As this implies the minimum number of evaluations necessary for a reasonable approximation of the integrals across all bins, we set $\tau_k = 2^j$.

4.4.1 Varying the number of bins, B

We specify (training and test) datasets each comprising $N = 20\,000$ observations for which the response variable Y can take values in $\Omega = \{1, 2, 3\}$ ($K = 3$) conditional on $D = 5$ explanatory variables. The true vector of regression coefficients $\boldsymbol{\beta}_{true}$ has entries randomly drawn from a $U[-5, 5]$ distribution. The explanatory variables are drawn from D-dimensional normal and skew-normal distributions, with correlation matrices containing zero correlations (the identity matrix) or correlations drawn from $U[-0.75, 0.75]$. The elements of the skew-normal slant vector are drawn from $U[-7, 7]$. While the correlation parameter of the skew-normal distribution is not equivalent to the correlation matrix of the associated random variable, skew-normal data simulated using the identity matrix as the correlation parameter typically have low correlations. When aggregating the design matrix $\boldsymbol{X}$ into a histogram through (4.5), an equal number B of bins is set for each margin and each outcome k, i.e. $\boldsymbol{B}_k = (B, B, B, B, B, B)$ for all $k \in \Omega$.

We use both the full multinomial regression (M) model fit using (4.2) and the OvR (O) model fit using the full likelihood (4.3) as reference fits. We also fit the multinomial mixture (MM) model of histogram and underlying classical data (4.8), and the univariate approximate composite likelihood $L_{\mathrm{O}}^{(1)}$ (see Proposition 4.3.3). For the histogram-based, univariate approximate composite likelihood (4.10) we make the assumption that the covariates are either independent ($\alpha_{\boldsymbol{i}i'} = 0$ and $\lambda_{\boldsymbol{i}i'_1i'_2} = 0$ in (4.9)) or that the variance-covariances of the covariates are to be estimated.

Figure 4.1 illustrates the mean prediction accuracies (over 1 000 replicates) as a function of the number of bins B, obtained for each of the above models and estimation procedures. For the full dataset, using an approximate composite likelihood approach to fit the OvR model (dashed grey line) yields comparable prediction accuracies to the full likelihood approach (solid grey line), in particular when the covariates are independent of each other (left panels).

When the empirical variance-covariance matrix is used, the histogram-based OvR model fitted using the $L_{\mathrm{SO}}^{(1)}$ approximate composite likelihood (grey dotted line) is able to obtain comparable prediction accuracies to the $L_{\mathrm{O}}^{(1)}$ full-data approximate composite likelihood OvR model (solid grey line), for a reasonable (≈ 10 marginal bins) level of data aggregation, and for any covariate distribution (rows). This clearly demonstrates that $L_{\mathrm{SO}}^{(j)}(\boldsymbol{x}, y; \boldsymbol{\beta}) \to L_{\mathrm{O}}^{(j)}(\boldsymbol{s}; \boldsymbol{\beta})$ as discussed in Section 4.3.3. It also performs well compared to the analysis on the full data (solid black line) when the covariates are independent (left panels; note the small y-axis scale).

When the empirical variance-covariance matrix is unavailable and the correlations between explanatory variables are assumed to be independent ($\alpha_{\boldsymbol{i}i'} = 0$ and $\lambda_{\boldsymbol{i}i'_1i'_2} = 0$), the histogram-based approximate composite likelihood model (dot-dashed grey line) still produces reasonable prediction accuracies. However, as should be expected, there is a clear loss in performance compared to when the covariances are known, for densities with high covariate correlations (right panels) and asymmetric distributions (bottom panels).

The prediction accuracies obtained from the multinomial model (dashed black line) have converged relatively quickly to its classical equivalent (solid black line) requiring only around $B = 5$ bins per margins. The multinomial model gives higher overall prediction accuracies

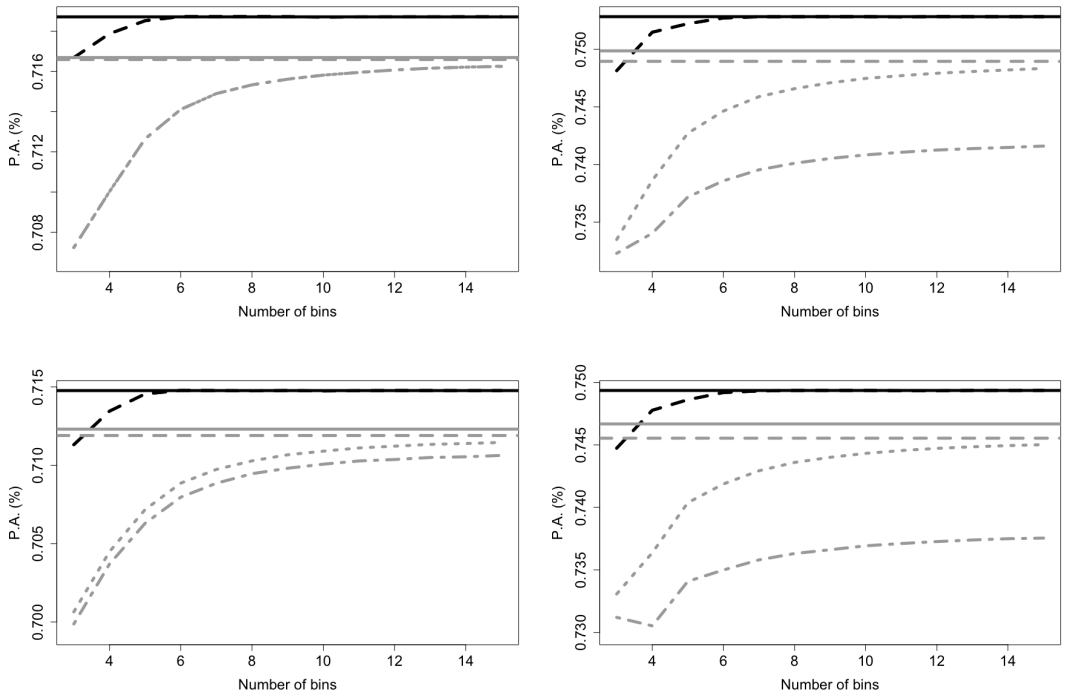

Figure 4.1: Average prediction accuracy (P.A.) computed over 1 000 replications for the multinomial model using full likelihood (solid black), mixture multinomial model (dashed black), OvR model using full likelihood (solid grey) and approximate composite likelihood (dashed grey), histogram-based OvR model using approximate composite likelihood assuming independence of the covariates (dot dashed grey) and using additional covariate assumption (dotted grey). Top panels consider covariates simulated from the multivariate normal distribution and bottom panels using the skew normal distribution. Left panels assume the covariates have zero correlation parameter, and right panels use non-zero correlations.

than the OvR model in each case, however recall that the data were simulated according to the multinomial model, giving it a natural advantage. While the multinomial model is generally preferred over the OvR model here, in practice the One-vs-Rest approach can outperform the multinomial model for some datasets, and can give almost as good results in many other cases (e.g. Eichelberger and Sheng, 2013).

Figure 4.2 illustrates the mean computational efficiency of fitting each model. It highlights the computational superiority of the histogram-based OvR model when univariate components are included in the likelihood ($L^{(1)}_{\text{SO}}(\boldsymbol{s};\boldsymbol{\beta})$; grey lines) against the full data multinomial model (black lines). The computation time increases as the number of bins increases when $L^{(1)}_{\text{SO}}(\boldsymbol{s};\boldsymbol{\beta})$ is used, however comparable predictions to the full multinomial model are achieved for $B \approx 10$ (c.f. Figure 4.1) at a significantly cheaper computational cost. For the mixture multinomial model (black dashed lines), the computation time increases strongly with increasing B. This phenomena is due to the intractability of the integrands in (4.6) requiring numerical integration, and our choice of τ_k. However, for this setup $B \approx 5$ is sufficient to provide comparable predictions to the multinomial model (Figure 4.1), and with lower computational overheads than the

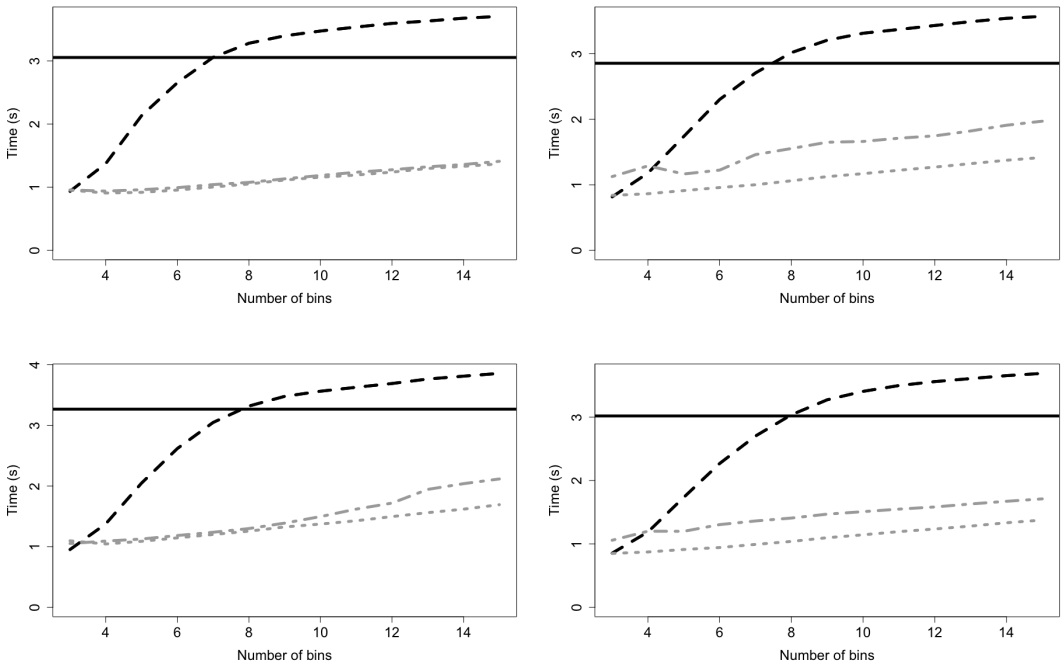

Figure 4.2: Average computation time (in CPU seconds) over 1 000 replications for the multinomial model using full likelihood (solid black), mixture multinomial model (dashed black), histogram-based OvR model using approximate composite likelihood assuming independence of the covariates (dot dashed grey) and using additional covariate assumption (dotted grey). Top panels consider covariates simulated from the multivariate normal distribution and bottom panels using the skew normal distribution. Left panels use covariates from a zero correlation parameter, and right panels use non-zero correlations.

complete data case.

Even though the $L^{(1)}_{\text{SO}}(\boldsymbol{s};\boldsymbol{\beta})$ and $L_{MM}(\boldsymbol{s};\boldsymbol{\beta})$ likelihood approaches for histogram-valued data increase in computational intensity with increasing B, relative to the classical M model, we need to keep in mind that $N = 20\,000$ observations are considered. Increasing N will directly increase the computational time of the classical approach (solid black line), while computational overheads will remain relatively unchanged for the histogram-based methods (where computation is proportional to the number of bins, not datapoints within bins). Consequently, there are clear computational benefits to employing a histogram likelihood approach when analysing extremely large datasets. We explore this in Section 4.4.2.

4.4.2 Varying the number of underlying observations, N and comparison with subsampling

Aggregating data into summaries and performing an analysis on these new "datapoints" seems a good strategy when the sample size is large. It is natural to compare this approach to other popular techniques for downsizing data volume, such as subsampling algorithms. We use the two-step subsampling scheme given by Wang et al. (2018) for logistic regression modelling,

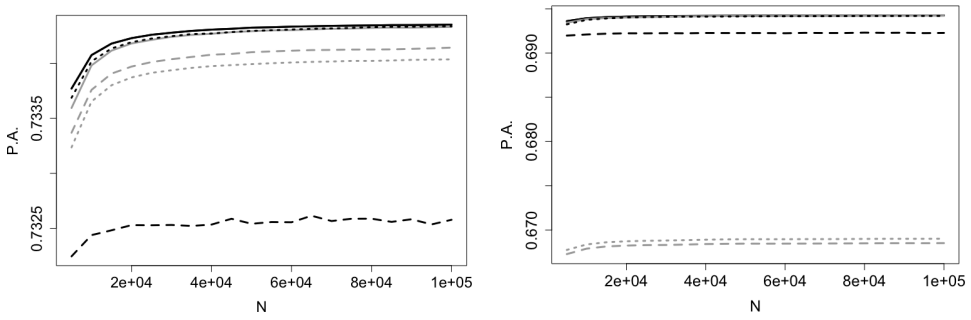

Figure 4.3: Mean prediction accuracies (P.A.) using the multinomial model on the full data (solid black line), subsampled data (dashed black line) and the histogram-based OvR model using $L_{\mathrm{SO}}^{(1)}$ with independence assumption (dashed grey line), $L_{\mathrm{SO}}^{(1)}$ with correlations (solid grey line), $L_{\mathrm{SO}}^{(2)}$ (dotted black line) and the naive composite likelihood model (dotted grey line) as a function of the number of datapoints N. The covariates are generated from 8-dimensional skew-normal distributions, considering zero (left) and non-zero (right) correlation parameters. The responses have two possible outcomes ($K = 2$). Results are based on 1 000 replicate analyses.

which first uniformly draws a subsample of size $r_0 = 1\,000$ from the dataset to produce an MLE estimate $\hat{\boldsymbol{\beta}}_0$, and uses this to produce an optimal weight for each datapoint. The second step then draws with replacement a subsample of size $r = 1\,000$ using the optimal weights, and then determines the final estimate of $\boldsymbol{\beta}$ using the total subsample of size $r_0 + r$.

In the following binary response ($K = 2$) experiment each element in the true vector of regression coefficients is generated from $U[-1, 1]$, and the number of observations, N, varies between 5 000 and 100 000. The explanatory variables are drawn from 8-dimensional skew-normal distributions ($D = 8$), with either zero correlations (identity matrix) or correlations drawn uniformly on $[0, 0.75]$. The slant vector of the skew-normal distribution is drawn from $U[-10, 10]$. (Recall that the identity correlation matrix for the skew-normal distribution does not lead to independent covariates, but rather low correlations between the covariates.) Note that in the case of binary responses the multinomial and OvR models are identical. After aggregating the covariates into histograms with $B = 15$ bins for each margin, the histogram-based OvR model is fitted using $L_{\mathrm{SO}}^{(1)}$ and $L_{\mathrm{SO}}^{(2)}$ (univariate and bivariate marginal histograms), including covariate correlations. The $L_{\mathrm{SO}}^{(1)}$ model is also fitted assuming independence between covariates.

Figure 4.3 shows that the mean prediction accuracies obtained by each method are increasing functions of the sample size N. Overall the symbolic based methods yield higher prediction accuracies than the subsampling approach when the covariate correlations are incorporated, and the more informative bivariate histogram setup $L_{\mathrm{SO}}^{(2)}$ will outperform the univariate histogram-based $L_{\mathrm{SO}}^{(1)}$. When the covariates exhibit low correlations, the $L_{\mathrm{SO}}^{(1)}$ model provides significant improvements over the naive composite likelihood analysis, regardless of whether correlations are incorporated. When there are correlations between the covariates, the $L_{\mathrm{SO}}^{(1)}$ model only provides significant improvements over the naive composite likelihood analysis if the covariate

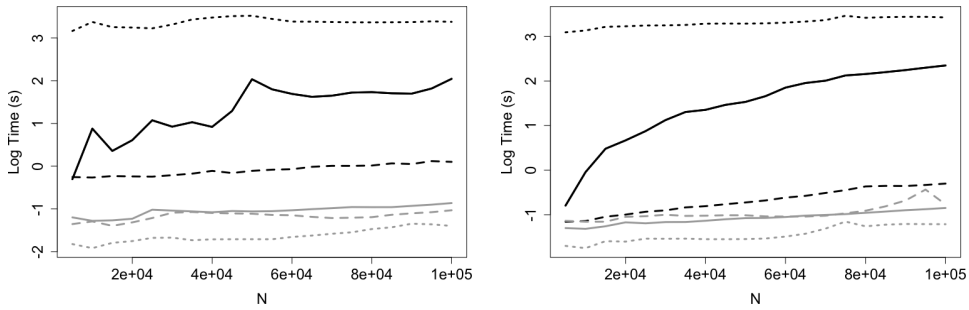

Figure 4.4: Mean total computation times (in CPU seconds) for the multinomial model on the full data (solid black line), subsampled data (dashed black line) and the histogram-based OvR model using $L_{\text{SO}}^{(1)}$ with independence assumption (dashed grey line), $L_{\text{SO}}^{(1)}$ with correlations (solid grey line), $L_{\text{SO}}^{(2)}$ (dotted black line) and the naive composite likelihood model (dotted grey line) as a function of the number of datapoints N. The covariates are generated from 8-dimensional skew-normal distributions, considering zero (left) and non-zero (right) correlation parameters. Results are based on 1 000 replicate analyses.

correlations are included. The magnitude of the variations in the prediction accuracy confirms that the extra efforts to use $L_{\text{SO}}^{(2)}$ are not justified in this case, and that univariate marginal histograms provide enough information and produce comparable results to a classical full data analysis.

Figure 4.4 supports these conclusions by providing the overall computation times (including aggregation and optimisation) for each model in Figure 4.3. We observe that the mean computational time required for the univariate symbolic model is significantly lower than for the multivariate model with full data, with an increasing disparity as the sample size N increases, as the number of terms in the histogram-based OvR model depends on the histogram construction and not N (and so is constant in these plots). In addition to its prediction superiority, $L_{\text{SO}}^{(1)}$ also computationally outperforms the subsampling approach of Wang et al. (2018). Note that Figure 4.4 indicates that $L_{\text{SO}}^{(2)}$ (dotted black line) is computationally more demanding than using the full data (solid black line), making it superfluous in this setting. However note that as computation for $L_{\text{SO}}^{(2)}$ is constant in N, there is some value N_0 such that if $N > N_0$ then the computational overheads for $L_{\text{SO}}^{(2)}$ will be more efficient than for the full data analysis.

Figure 4.5 explores parameter estimator performance via the mean mean squared error (MMSE) of a model's MLE, $\hat{\theta}^{\text{Model}}$, defined as $\text{MMSE}(\hat{\theta}^{\text{Model}}) = S^{-1}\sum_{s=1}^{S} \|\hat{\theta}_s^{\text{Model}} - \theta_s^{\text{true}}\|^2$, where $\|\cdot\|$ denotes the Euclidean norm, θ_s^{true} is the true parameter vector, and $S = 1\,000$ the number of replicate analyses. Figure 4.5 demonstrates that subsampling methods perform better than the histogram-based methods if a low MMSE is desired. Using $L_{\text{SO}}^{(1)}$ instead of the naive composite likelihood analysis leads to a lower MMSE, with the results further improving if the covariate correlations are included.

Figure 4.6 explores parameter estimate accuracy further, displaying the mean estimates for a selection of the regression parameter over the 1 000 replicate analyses. The estimates obtained

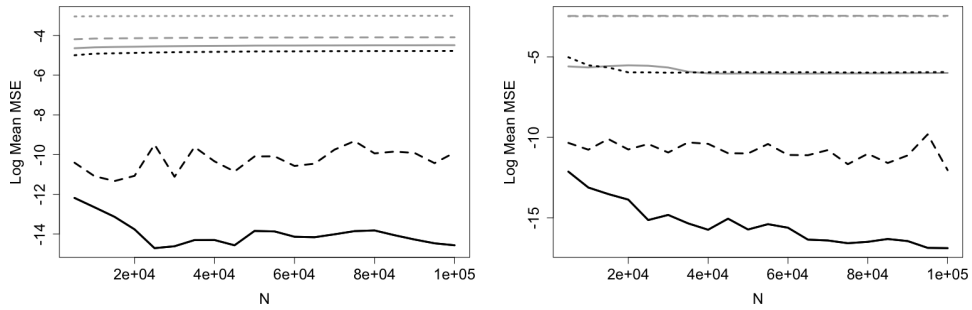

Figure 4.5: MMSE for the multinomial model on the full data (solid black line), subsampled data (dashed black line) and the histogram-based OvR model using $L_{\text{SO}}^{(1)}$ with independence assumption (dashed grey line), $L_{\text{SO}}^{(1)}$ with correlations (solid grey line), $L_{\text{SO}}^{(2)}$ (dotted black line) and the naive composite likelihood model (dotted grey line) as a function of the number of datapoints N. The covariates are generated from 8-dimensional skew-normal distributions, considering zero (left) and non-zero (right) correlation parameters. Results are based on 1 000 replicate analyses.

from $L_{\text{SO}}^{(1)}$ are much closer to those of the full model analysis than that of the naive composite likelihood analysis, with accuracy improving if covariate correlations are incorporated into the model. While the subsampling method provides more accurate parameter estimates, the better performance of the histogram-based models for predictions can potentially be explained by the fact that the histogram-based models incorporate the entire dataset, whereas a subsampling scheme can still potentially omit important observations. For predictions and the binary model, an observation x_n is assigned to class 1 if $\beta^\top x_n < 0$, and class 2 otherwise. Consequently, the same predictions are obtained from any parameter vector $m\beta$, $m > 0$ for any dataset. In this case, the histogram-based models are more accurately estimating the model parameters to proportionality compared to the subsampling scheme, despite having a larger MSE.

In summary, this experiment suggests that if predictions are desired for logistic regression models, histogram-based solutions can be more accurate and computationally more efficient than subsampling-based methods, such as that in Wang et al. (2018). The use of bivariate histograms to represent the covariate information improves the prediction of the response outcomes, but at an often impractical computational cost compared to univariate histograms. This simulation study suggests that marginally aggregating the covariates into univariate histograms, in combination with knowledge of covariate correlations, provides the best trade off between accuracy and speed.

4.5 Real data analyses

We illustrate the applicability of our proposed methodology to two real data problems. We first consider a logistic regression problem where the goal is to distinguish between a process where new supersymmetric particles are produced and a background process. Secondly we tackle a multinomial regression problem which consists of predicting crop types based on satellite-based

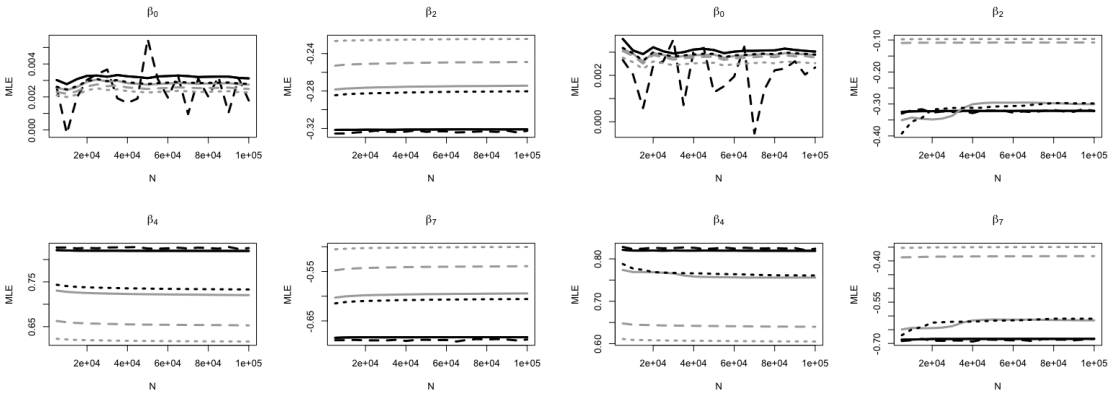

Figure 4.6: Mean MLEs using the multinomlial model on the full data (solid black line), subsampled data (dashed black line) and the histogram-based OvR model using $L_{\text{SO}}^{(1)}$ with independence assumption (dashed grey line), $L_{\text{SO}}^{(1)}$ with correlations (solid grey line), $L_{\text{SO}}^{(2)}$ (dotted black line) and the naive composite likelihood model (dotted grey line) as a function of the number of replicates N. The covariates are generated from 8-dimensional skew-normal distributions, considering zero (left two columns) and non-zero (right two columns) correlation parameters. Results are based on 1 000 replicate analyses.

pixel observations.

4.5.1 Supersymmetric benchmark dataset

The Supersymmetry dataset (SUSY) is available from the Machine Learning Repository (Dua and Graff, 2017) and comprises 5 million Monte Carlo observations generated by Baldi et al. (2014). The binary response variable ($K = 2$) discriminates a signal process which produces supersymmetric particles from a background process which does not. There are 18 features ($D = 18$); the first 8 are low-level features representing the kinematic properties measured by the particle detectors, while the remaining 10 are high-level features derived as function of the previous 8 by physicists to help discriminate between the two outcomes. This dataset was analysed by Wang et al. (2018) to test their optimal subsampling scheme for logistic regression. Following Wang et al. (2018), we consider a training dataset of 4 500 000 randomly chosen observations, and a test dataset with the remaining 500 000 observations.

Following the conclusions from Section 4.4, we fit the histogram-based OvR model using univariate marginal histogram aggregates ($L_{\text{SO}}^{(1)}$). While until now the focus has been on histograms with random counts (fixed bins), here we fit $L_{\text{OO}}^{(1)}$ to explore the performance of using random bin (fixed counts) histograms. For an integer B, we construct histograms for each covariate by partitioning the data into B bins with roughly equal counts. For example, for $B = 4$ we would use the $0.0, 0.25, 0.5, 0.75$ and 1.0 empirical quantiles to construct the histogram for each covariate.

The likelihood functions are optimised using ridge regularisation with 10-fold cross-validation and, for $L_{\text{OO}}^{(1)}$, for a range of values of B. Prediction accuracies obtained on the test dataset and the optimisation times (in seconds) on the training dataset are reported in Table 4.1. For

Likelihood	Bins							
	6	8	10	12	15	20	25	
$L_{\mathrm{OO}}^{(1)}$	74.9	75.9	76.6	77.7	78.1	77.9	78.1	
	(11.7)	(14.5)	(12.2)	(15.0)	(18.9)	(21.3)	(27.6)	
$L_{\mathrm{SO}}^{(1)}$	74.4	73.5	75.8	77.8	77.4	78.0	78.0	
	(13.3)	(12.6)	(11.5)	(13.9)	(16.8)	(18.0)	(21.4)	
Subsampling Wang et al. (2018)								78.2 (86.1)

Table 4.1: Percentage prediction accuracy with computing time (in seconds) for the Supersymmetry dataset, using histograms with B bins per margins, and the subsampling approach of Wang et al. (2018).

the histogram-based models, there is an increase in prediction accuracy as the number of bins increases, which is as expected since these are more informative summaries. The improvement in prediction accuracy slows down at around the $B = 12$ bin mark, whereas the computation time naturally increases with the number of bins. The performance when $B = 12$ reaches only slightly inferior levels than those in Figure 10(b) in Wang et al. (2018) when $r_0 = 200$ and $r = 1\,000$ with various subsampling probabilities. When replicating this method for comparison, we obtain a prediction accuracy of 78.2% with a computation time of 86.1 seconds i.e. about 3–4 times more computation than for the histogram-based models with $B = 25$. That is, the histogram-based methods offer as good prediction accuracy with much smaller computational overheads compared to state-of-the-art subsampling approaches.

4.5.2 Crop type dataset

We examine a crop type dataset (QUT, 2016) which consists of 247 210 observations, each representing a $25 \times 25\mathrm{m}^2$ pixel located over farmland across the state of Queensland, on the east coast of Australia (Figure 4.7). For each pixel the ground-truth crop type is available (observed at one of three possible times) as well as numerous vegetation indices, based on reflectance data taken from a LANDSAT 7 satellite. The aim of this analysis is to predict the crop type based on the vegetation indices. After selecting the most meaningful covariates by iteratively removing variables with correlations greater than 0.85, we retained $D = 7$ variables corresponding to various colour reflectances measured by the satellite and functions of these indicators.

As poor prediction accuracy of classes with low numbers of observations is a well known issue, we only retain crop types that are observed more than 10 000 times, reducing the dataset to 234 485 observations. The set of possible outcomes of our multinomial response variable $Y =$ "Crop type" is then $\Omega = \{\text{Bare soil}, \text{Cotton}, \text{Maize}, \text{Pasture natural}, \text{Peanut}, \text{Sorghum}, \text{Wheat}\}$ and thus $K = 7$. The resulting dataset is identical to the one used in a previous analysis in QUT (2016) which used the standard multinomial model $L_{\mathrm{M}}(\boldsymbol{x}, y; \boldsymbol{\beta})$.

As the approximate composite likelihood relies on the assumption of a linear relationship between the predictor variables, we use the R package `bestNormalize` to select the best transformation to achieve approximate predictor Gaussianity, according to the Pearson P-test statistic. The dataset is randomly partitioned into a training dataset of size 200 000 used for parameter

		Bins						
Crop type	N_k	6	8	10	12	15	20	$L_{\mathrm{M}}(\boldsymbol{x}, y; \boldsymbol{\beta})$
Cotton	72 450	90.5	90.6	92.8	93.6	94.0	94.1	92.2
Sorghum	66 751	74.6	74.8	75.7	76.4	76.2	76.3	80.3
Pasture Natural	27 479	75.7	75.4	76.0	76.8	77.0	77.1	77.6
Bare Soil	26 173	88.0	89.6	89.2	90.0	89.5	90.1	91.0
Peanut	17 868	81.2	81.3	81.5	81.5	81.9	81.6	82.9
Maize	12 986	9.7	9.9	10.2	10.4	10.3	10.4	14.2
Wheat	10 778	3.4	4.0	4.8	5.0	5.2	5.7	10.3
Overall	234 485	74.6	75.5	76.4	77.1	77.2	77.2	78.1
		(164)	(162)	(221)	(229)	(276)	(508)	(6071)

Table 4.2: Crop specific and overall prediction accuracies (%) using univariate marginal histograms with B bins. The likelihood optimisation times (in seconds) are reported in the last row. The full model is the standard multinomial likelihood $L_{\mathrm{M}}(\boldsymbol{x}, y; \boldsymbol{\beta})$ (4.2) with LASSO regularisation, as implemented by QUT (2016).

estimation and a test dataset with the remaining 34 485 observations to evaluate the prediction accuracy. We perform constrained likelihood optimisation with a LASSO regularisation, and use 10-fold cross validation to determine the best regularisation parameter.

Table 4.2 presents the prediction accuracies for the $L_{SO}^{(1)}$ model for each crop type and the overall prediction accuracy when the covariate information is collapsed into univariate marginal histograms with $B = 6, 8, 10, 12, 15$ and 20 bins. The last column of Table 4.2 provides a comparison with the full data multinomial likelihood ($L_{\mathrm{M}}(\boldsymbol{x}, y; \boldsymbol{\beta})$) using the same LASSO regularisation, as implemented in the original analysis by QUT (2016). The overall and crop-specific prediction accuracies have achieved good predictive performance compared to the full data multinomial model analysis using only $B \approx 10$–12 bins. Two particular crops produce notable results. The prediction accuracies for Wheat are around 5% for the histogram-based analysis compared to the ∼10% accuracy of the full-data analysis. While both of these are low due to this crop being the least well represented of all crops in the study (less than 5% of all observations), and perhaps lowly informative vegetation indices for this crop, the 50% predictive underperformance for the histogram-based analysis suggests that categories with less data in a model are more sensitive to the degree of binning than those categories with larger representation in the dataset (although this is less apparent for Maize).

In the case of Cotton, which is the largest representative category (at ∼31%), the histogram-based prediction accuracies are even higher (at ∼94%) than for the full data analysis (92.2%). While this is not immediately understandable intuitively, in that by constructing histograms information in the dataset is certainly being lost and so performance should perhaps always be worse, the difference is only 2%, and moreover the likelihoods are not directly comparable in the sense that the limit of the approximate composite likelihood $L_{\mathrm{SO}}^{(j)}(\boldsymbol{s}; \boldsymbol{\beta})$ as $B \to \infty$ is not the full data multinomial model $L_{\mathrm{M}}(\boldsymbol{x}, y; \boldsymbol{\beta})$ as used in QUT (2016). So here the discrepancy is that the two likelihoods simply have different performances for these data. This argument notwithstanding, proponents of symbolic data analysis sometimes ascribe to the idea that in-

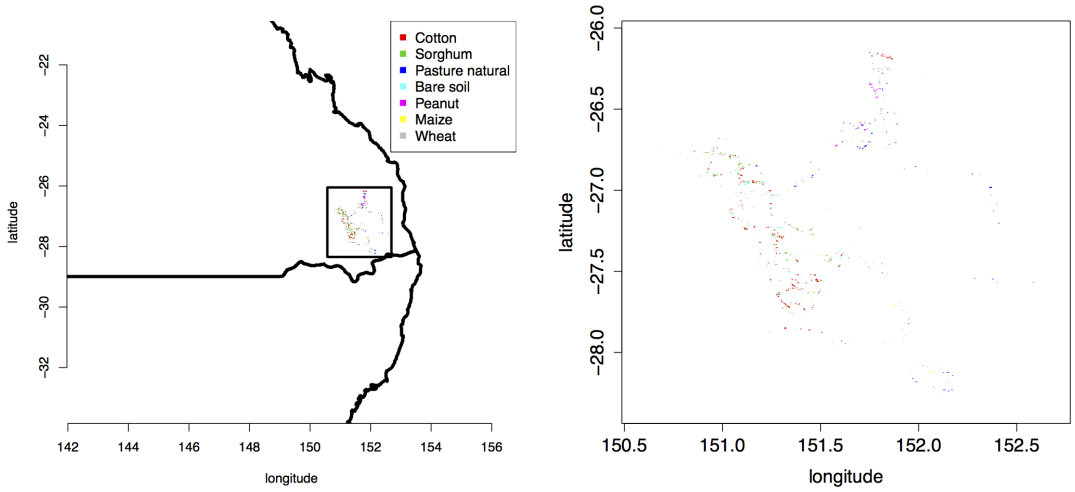

Figure 4.7: The crop type dataset with different colours for each crop. Left: Location of the study area in the state of Queensland on the east coast of Australia. Right: farm location and crop type detail.

ference using the 'shape' of the data may sometimes be more useful than an analysis of the full underlying dataset (Edwin Diday, personal communication).

Overall, while the histogram-based analysis gives comparable prediction accuracies to the full data analysis, the real gains are in the computational overheads required for each model. The full multinomial analysis takes considerably longer (more than $25\times$ the $B = 12$ analysis) to implement than the histogram based analysis. Finally, note that while the computational savings here are substantial, this dataset only contains $N = 234\,485$ observations. For larger datasets the computational overheads will skyrocket for the standard multinomial model analysis (where computation is proportional to N), and yet will remain roughly constant for the histogram-based approach (where computation is proportional to B).

4.6 Discussion

In this article, we have developed a novel approach for classifying binary and multinomial random variables that alleviates the computational bottleneck that arises with very large datasets. The strategy relies on concepts from the field of symbolic data analysis (Beranger et al., 2018), aggregating the covariate data into histogram-valued random variables which have lower computational overheads to analyse and store, albeit with some loss of information. When computation for any histogram bin is larger than that for the standard likelihood contribution of the datapoints within that bin, the standard likelihood contribution for these datapoints can be used instead. However, because high-dimensional histograms are not efficient distributional summaries, we additionally introduced an approximate composite likelihood methodology, which quantitatively builds on the qualitative results of Cramer (2007). The individual components of the approximate composite likelihood are constructed from marginal histograms derived from the full D-dimensional histogram. This concept of approximate composite likelihoods for logistic

regression does not solely apply to aggregated data and can be used in more general settings.

We have demonstrated through simulation studies and real data analyses that these histogram-based strategies can produce fitted models that have comparable prediction accuracies to the standard full data analysis, but at a much lower computational cost, even compared to state-of-the-art computational techniques for logistic regression such as subsampling (Wang et al., 2018). On the down side, the resulting parameter estimates are biased, though not as much as for naive composite likelihood-based approaches.

One aspect of implementing histogram-based inference that we have not explored is principled ways of constructing the histograms for subsequent analysis. If the number of bins is too low then important information in the data will be lost and model predictions may be poor (see e.g. Figure 4.1). In contrast, as the number of bins becomes large then inferential accuracy can approach the level of the full data analysis (within the context of the inferential model being used). However, the price of more accurate inference is an increase in the computational costs. A simple approach was used in the real data analyses in Section 4.5, in which increasing values of B were investigated until the change in results was negligible, thus demonstrating convergence to the comparable classical model. A downside of this approach is that it requires the optimisation of multiple likelihood functions until convergence is reached, whereas an ideal method would be to determine the optimal value of B prior to any aggregation. We note however that for huge datasets like those analysed in Section 4.5, there are still significant computational gains associated with the use of the models presented in this paper, despite this drawback. An 'optimal' approach could therefore consider balancing computational complexity and inferential accuracy, or alternatively by minimising a loss function constructed over some useful criterion. This is clearly an important component of the current methodology, and is the focus of current research.

Acknowledgements

This research is supported by the Australian Research Council through the Australian Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS; CE140100049), and the Discovery Project scheme (FT170100079).

Chapter 5

Non-Parametric Estimating Equations for symbolic data

5.1 Introduction

Symbolic Data Analysis (SDA) is a branch of statistics that has been introduced to deal with some of the issues associated with the growing size and complexity of modern datasets (see Billard and Diday (2003, 2006), Bertrand and Goupil (2000)), with applications in various fields such as environmental and climate sciences, economics and medicine. Large datasets are summarised into non-standard observations such as distributions, intervals and histograms, which can contain qualitative and quantitative data of a manageable size. When classical data is aggregated into a non-standard form, knowledge of the underlying data is generally lost. Furthermore, classical statistical techniques are often not appropriate due to the internal variation associated with symbolic objects, meaning new methods of analysis and inference need to be developed.

An important distinction between classical and non-standard data is that non-standard data possess internal variation, whereas classical data does not. An analysis of symbolic data must therefore account for the variation of the underlying microdata within a symbolic object, as well as the variation between symbolic observations (Billard, 2011), which isn't possible with existing classical methods of analysis due to the lack of internal variation within classically observed pointwise data. With the exception of a small number of works (e.g. see Le Rademacher and Billard (2011) and Beranger et al. (2018)), most of the existing literature focuses on exploratory analyses of symbolic observations without a connection to the underlying microdata from which the symbols arose. Bertrand and Goupil (2000), Billard and Diday (2003), Billard (2011) proposed estimates for the sample symbolic mean, variance and correlations for interval and histogram valued datasets, which we will show in this paper is equivalent to the classical results if the microdata is uniformly distributed within each symbol. Oliveira et al. (2018) describe the various definitions of symbolic variances and correlations that have been proposed in recent years, and derive the conditions on the underlying microdata that are required for the symbolic estimates to be equivalent to the classical case. These estimates are limited to assump-

tions that in practise are often violated, such as the aforementioned uniformity assumption or the assumption of normality within each symbol.

Non-parametric estimating equations were defined by Godambe (1990) as a method of obtaining estimates for non-parametric statistics (i.e. means, variances, skewness, quantiles etc.) of a classical dataset without the need for a parametric assumption. Parametric methods of estimation, such as Maximum Likelihood Estimation, can be considered a special case in which a parametric density is assumed for the observed dataset. Various methods have been developed for obtaining variances and confidence intervals for these estimates, such as the bootstrap (Efron, 1979) and more recently, Empirical Likelihood (EL) (Owen, 1988, 1990). Empirical Likelihood has emerged as an effective means of obtaining confidence intervals for quantities of interest, without making distributional assumptions on the data. Some common applications include the EL analysis of econometric data (Bravo, 2004), censored survival data (Zhou and Li, 2008, Zhou, 2015) and time series data (Nordman and Lahiri, 2014, Piyadi et al., 2017). Qin (2017) provide some examples of simple applications of EL analysis. While these methods are well developed for classical-valued datasets, they need to be extended to the non-standard case so that variances and confidence intervals can be obtained for symbolic-valued datasets. Elashoff and Ryan (2004) propose an EM algorithm for estimating equations for missing data, whereby the expected values of the estimating equations for the missing data are obtained using the observed data, and substituted into the usual estimating equations. Wang and Pepe (2000) proposed an expected estimating equations approach to accomodate error in the measurement of covariates, whereby the error is modelled via an assumed parametric density. This methodology was extended to accomodate missing data and missclassification by Wang et al. (2008). In some cases these methods are insufficient for data arriving in a non-standard form however, as they either require some classical data to be observed or the assumption of some parametric structure which may not be verifiable or reasonable. As a result, new methods need to be developed to accomodate the case whereby data arrives in a non-standard form (e.g. intervals, histograms, distributions, etc.) and for which we don't want to (or can't) assume a parametric form for the complete underlying density or the distribution within each symbol.

There has been much work done on the estimation of non-parametric densities from non-standard data. The simplest example is the often utilised histogram density estimator (Silverman, 1986), which is easily derived from any observed histogram. Minnotte (1996) proposed the 'bias-optimized frequency polygon', which is an extension of the frequency polygon (Oliver, 2014) with preserved bin probabilities for each histogram bin. Scott and Sheather (1985) and Hall (1996) provide theoretical results for the errors involved in performing a Kernel Density Estimation (KDE) analysis on the midpoints of binned data, weighted by their respective counts. Blower and Kelsall (2002) improve on this by constructing a smooth kernel density estimator through the integration of the classical KDE over the domain of the underlying microdata. Minnotte (1998) and Koo and Kooperberg (2000) estimate the underlying density from binned data using splines fitted to the observed proportions, although it is shown in Minnotte (1998) with a real histogram dataset that this can lead to negative density estimates at some points if low count bins are sandwiched between two high count bins. Yongho et al. (2015) develop

a method of non-parametric density estimation for intervals by using weighted local Gaussian Kernels. Conditional means and variances of each interval are estimated using weighted sums of the same quantities from neighbouring intervals, where the weight is determined by the distance between each interval, and the distances are calculated according to either midpoints or edges of each interval. Each of these methodologies are effective in estimating the underlying density of the microdata, however require extra decisions from the practioner, such as kernel choice, bandwidth selection, number of splines in the case of Minnotte (1998) and Koo and Kooperberg (2000), and distance metric in the case of Yongho et al. (2015). Furthermore, the main objective of these works revolve around the estimation and visualisation of an underlying classical data density, and not of the values estimating equations would take from that underlying density.

Le Rademacher and Billard (2011) utilise the empirical densities derived by Bertrand and Goupil (2000) and Billard and Diday (2003) to derive a likelihood-based approach for interval-valued symbolic data, from which a histogram symbolic likelihood function is a simple extension. These likelihood functions are effective in deriving the internal statistics of a symbol given an assumed parametric distribution, however there is no mechanism for understanding the structure of the underlying data from which the symbolic data arose. Heitjan (1989) explored the use of Sheppard's correction in the estimation of a sample variance from rounded data, which states that if a random variable X is i.i.d. normally distributed, and instead of observing X, we observe rounded values Y, then $Var(X) = Var(Y) - \frac{\delta^2}{12}$, where $\frac{\delta}{2}$ is the degree of rounding. This can be generalised to any parametric framework, but provides no mechanism of estimating variance if the underlying parametric family is unknown. Furthermore, Sheppard's correction is restricted in that it is unreliable for datasets that are non-symmetrical, multimodal, and of a low sample size. Beranger et al. (2018) proposed a general symbolic likelihood function whereby Maximum Likelihood Estimators (MLEs) for the underlying classical data can be obtained from the symbolic data. This construction automatically assumes a non-uniform truncated parametric density within each symbol, therefore removing the unreliable uniformity assumption previously utilised. While these models are effective in estimating an underlying microdata parametric density, they provide no mechanism for modelling the underlying data and obtaining estimates for parameters if we don't assume a parametric form.

In this paper we propose a methodology for obtaining non-parametric estimates for statistics from symbolic datasets that are comparable to that of a classical analysis of the original microdata for a certain level of information retention/data aggregation. These estimates do not require the assumption of a uniform within-symbol distribution, and in fact can be obtained for any symbolic dataset for which we can estimate the internal structure of the microdata. Furthermore, our model does not require a parametric assumption on the underlying data or a uniformity assumption, which is of particular significance in the estimation of confidence intervals for quantiles from symbolic data. We show that for statistics that follow the concept of sufficiency derived by Elashoff and Ryan (2004), estimates can be written as a function of estimates of the same statistics for each individual symbol. The symbolic mean, variance and covariance proposed by Billard and Diday (2003), Bertrand and Goupil (2000), Billard (2011), Oliveira et al. (2018) can be considered a special case of these results whereby the microdata

is assumed to follow independent uniform distributions within each symbol. Furthermore, we extend the EL methodology proposed by Owen (1988) to the symbolic setting, enabling us to obtain variances and confidence intervals for the symbolic estimates. We then develop specific methodologies to estimate the within-symbol structure of the microdata for each symbol in interval and histogram-valued datasets by utilising information in neighbouring symbols, thus allowing more accurate estimation of within-symbol statistics such as the within-symbol mean, variance and skewness. The idea behind these constructions is that if the classical underlying data is generated from the same underlying process, then we can utilise the information in all the symbols to estimate the individual parameters for each symbol, instead of just treating each within-symbol distribution as independent.

The structure of this paper is as follows. In Section 5.2 we provide the necessary background information on Estimating Equations, Empirical Likelihood and SDA. In Section 5.3 we derive the general form for symbolic estimating equations such that the estimates obtained are comparable to those obtained from the underlying classical dataset for a certain level of data aggregation, and show that EL can be used to obtain variances and confidence intervals for these parameters. We then derive the general symbolic estimating equations for specific statistics, such as quantiles, means, variances, skewnesses, etc. In Section 5.4 we then derive specific methodologies to estimate the within-symbol structure of the microdata for each symbolic observation using the entire symbolic dataset for interval and histogram valued datasets, allowing better estimates for statistics of the underlying data such as mean, variance, skewness and in particular quantiles. In Section 5.5 we then illustrate the increased efficiency of these constructions compared to previous results through the use of various simulation studies, and in Section 5.6 we then apply these methodologies to real datasets.

5.2 Background Information

In this section we first provide a brief overview of the theory of estimating equations and then define a framework that allows for statistical analysis of data summaries. This sets the foundations to develop estimating equations for aggregated data in the following section.

Consider a random vector $X = (X_{[1]}, \ldots, X_{[D]}) \in \mathcal{D}_X \subset \mathbb{R}^D$ with unknown distribution function F_X and let $\boldsymbol{X} = (X_1, \ldots, X_N)$ be the collection of N i.i.d. replicates of X with realisation given by $\boldsymbol{x} = (x_1, \ldots, x_N)$. Without any parametric assumption about F_X we are interested in making statistical inference on the parameter vector $\theta \in \mathcal{D}_\theta \subset \mathbb{R}^M$, using $R \geq M$ functionally independent estimating equations (EE) defined through $g(X, \theta) = (g_1(X, \theta), \ldots, g_R(X, \theta))^\top$, with the condition

$$\mathbb{E}_{F_X}[g_r(X, \theta)] = 0, \text{ for all } r = 1, \ldots, R \tag{5.1}$$

uniquely at θ_0 the true parameter value, ensuring unbiasedness.
Because we do not wish to make assumptions about distribution function F_X and the dataset $\boldsymbol{x}$ is available, we consider an empirical alternative. The empirical likelihood (EL) associated with

the observed sample $\boldsymbol{x}$ is given by

$$L(F_X; \boldsymbol{x}) = \prod_{n=1}^{N} \mathrm{d}F_X(x_n) = \prod_{n=1}^{N} P(X = x_n), \tag{5.2}$$

meaning that only the distributions that put an atom of probability on each x_n are considered. In other terms F can be seen as a discrete distribution on $\{x_1, \ldots, x_N\}$ with probability vector $p = (p_1, \ldots, p_N)$ defined such that (C1): $p_n = P(X = x_n) > 0, n = 1, \ldots, N$ and (C2): $\sum_{n=1}^{N} p_n = 1$. If no other conditions on $\boldsymbol{x}$ are imposed other than (C1) and (C2), then the EL likelihood (5.2) is maximised by the empirical distribution, i.e. $\hat{p}_n = 1/N$, for all n and from (5.1) an estimate of θ is given by

$$\frac{1}{N} \sum_{n=1}^{N} g_r\left(x_n, \hat{\theta}\right) = 0, \text{ for all } r = 1, \ldots, R. \tag{5.3}$$

In the empirical setting described above, the EE condition (5.1) can be used to define the additional condition (C3): $\sum_{n=1}^{N} p_n g_r(x_n, \theta) = 0$ for all $r = 1, \ldots, R$. Then, letting $\lambda = (\lambda_1, \ldots, \lambda_R)$ be the vector of Lagrange multipliers associated with (C3), constraint maximisation of (5.2) gives

$$\hat{p}_n = \frac{1}{N\{1 + \lambda^\top g(x_n, \theta)\}}.$$

Plugging the above expression in (C3) allows to determine λ as a function of θ. Owen (1990) showed that for the true parameter θ_0, $-2\sum_{n=1}^{N} \log(N\hat{p}_n) \sim \chi^2_R$, allowing to construct hypothesis testing and confidence intervals for θ. As an example, suppose $\theta = (\mu, \sigma^2) = (\mathbb{E}(X), \mathbb{V}(X))$, a natural choice of estimating function with $R = 2$ is then $g(X, \theta) = \left(X - \mu, (X - \mu)^2 - \sigma^2\right)$, which, from (5.3), gives the estimates $\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$ and $\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N} (x_n - \hat{\mu})^2$.

The notions presented until now evidently rely on the data being available. We now introduce a framework where only summaries of the information contained in the data are available. For $c = 1, \ldots, C$, let $\boldsymbol{X}^{(c)} = \{X_1^{(c)}, \ldots, X_{n_c}^{(c)}\}$ denote the c-th subset of $\boldsymbol{X}$ of size n_b such that $\bigcup_{c=1}^{C} \boldsymbol{X}^{(c)} = \boldsymbol{X}$ and $\sum_{c=1}^{C} n_c = N$ which can be interpreted as the set of members belonging to a *class* c (e.g. Billard and Diday, 2003). The so-called *symbolic object* $S_c \in \mathcal{D}_{S_c}$ is a summary of the information contained in $\boldsymbol{X}^{(c)}$ obtained through an aggregation function $\pi(\cdot)$ which may contain some deterministic elements ϑ. For observations given by $\boldsymbol{x}^{(c)}$, the information is summarised in $s_c = \{n_c, \Upsilon_c, \alpha_c\}$, where $\Upsilon_c = \mathcal{D}(\boldsymbol{X}^{(c)})$ and α_c contains the summary statistics specific to the aggregation function $\pi(\cdot)$.

If a parametric assumption is made on the underlying data is made with likelihood function $L(\boldsymbol{x}^{(c)}; \theta) = \prod_{i=1}^{n_c} f_X(x_i^{(c)}|\theta)$, then the likelihood of the symbol s_c is given as

$$L(s_c; \theta, \vartheta) = \int_{(\mathcal{D}_{\mathbf{X}})^{n_c}} L\left(\boldsymbol{x}^{(c)}; \theta\right) f_{S_c|\boldsymbol{X}^{(c)}}\left(s_c|\boldsymbol{x}^{(c)}, \vartheta\right) \mathrm{d}\boldsymbol{x}^{(c)}. \tag{5.4}$$

(see Beranger et al., 2018). For example, aggregating n_c univariate ($D = 1$) observations $x^{(c)}$ into

an interval defined by some l-th and u-th order statistics, i.e. for $\vartheta = (l, u)$ taking $\alpha_c = (\alpha_{c,l}, \alpha_{c,u})$ where $\alpha_{c,l} = x_{(l)}^{(c)}$ and $\alpha_{c,u} = x_{(u)}^{(c)}$, gives

$$L(s_c; \theta, \vartheta) \propto F_X\left(\alpha_{c,l}|\theta\right)^{l-1} \left(F_X\left(\alpha_{c,u}|\theta\right) - F_X\left(\alpha_{c,l}|\theta\right)\right)^{u-l-1} \left(1 - F_X\left(\alpha_{c,u}|\theta\right)\right)^{n_c-u} f_X\left(\alpha_{c,l}|\theta\right) f_X\left(\alpha_{c,u}|\theta\right).$$

The advantage of this construction is the ability to go beyond the omnipresent assumption that data are uniformly distributed within a symbol and to draw conclusions at the data level. Maximum likelihood estimation is a particular case of estimating equations where for the parameter vector $\theta = (\theta_1, \ldots, \theta_R)$ we have $g_r(x_n, \theta) = \partial/\partial\theta_r \log f_X(x_n|\theta)$. The aim of this paper is to define a framework for estimating equations constructed from data aggregates where the within symbol distribution does not require a uniformity assumption.

5.3 Estimating Equations using data summaries

If the data were available then the estimating equations would be those defined in (5.1) but if only summaries of the data are accessible, then a new set of equations is considered. These new equations are a transformation of the original ones which depends on the aggregation procedure.

Let $\phi_c \coloneqq f_{X|S_c=s_c}$ denote the density of the underlying random variable X given an observed summary s_c which is linked to $\phi \coloneqq f_{X|\boldsymbol{S}=\boldsymbol{s}}$, its equivalent when a set of summaries $\boldsymbol{s}$ is observed, through

$$\phi_c(x) = \frac{\mathbb{1}(x \in s_c)\phi(x)}{\mathbb{P}(s_c)}, \tag{5.5}$$

where the indicator restricts to the c-th symbol and the denominator corresponds its probability of occurrence $\mathbb{P}(s_c) = \int \phi(y)\mathbb{1}\{y \in s_c\}\mathrm{d}y$. If the symbols are assumed independent then we can take $\mathbb{P}s_c = \frac{n_c}{N}$ such that $\sum_{c=1}^{C} \mathbb{P}s_c = 1$ which yields

$$\phi(x) = \sum_{c=1}^{C} \frac{n_c}{N} \phi_c(x) \tag{5.6}$$

and for ease of notation let also $\phi(\boldsymbol{x}) = \prod_{n=1}^{N} \phi(x_n)$. The intuition behind these density will be studied in the following section with some detailed examples. Using the above we define the estimating equations for symbolic inputs as

$$\mathbb{E}_{F_S}[g'_r(S, \theta, \vartheta)] = 0, \text{ for all } r = 1, \ldots, R \tag{5.7}$$

where g'_r result from the aggregation on the original g_r functions and F_S is the symbolic distribution function. Before explicitly defining the functions g'_r we derive the F_S in the empirical context as follows.

Let's assume a discrete distribution on the symbols $s_1, \ldots, s_C$ with probabilities $p_1, \ldots, p_C$, such that (C4): $p_c = \mathbb{P}(S_c = s_c) > 0, c = 1, \ldots, C$ and (C5): $\sum_{c=1}^{C} p_c = 1$. Each observation $x_n^{(c)}$, $n = 1, \ldots, n_c$, $c = 1, \ldots, C$ aggregated into a symbol s_c has probability $q_n^{(c)} = \frac{p_c}{n_c}$, meaning

that $\sum_{c=1}^{C}\sum_{n=1}^{n_c} q_n^{(c)} = 1$. The equivalent of (5.2) for data aggregates $\boldsymbol{s}$ is thus

$$L(F_S; \boldsymbol{s}) = \prod_{c=1}^{C} \mathbb{P}(S = s_c) = \prod_{c=1}^{C}\prod_{n=1}^{n_c} q_n^{(c)}, \tag{5.8}$$

which is maximised, under (C4) and (C5), for $\hat{q}_n^{(c)} = \frac{1}{N}$ and implies $\hat{p}_c = \frac{n_c}{N}$ for $c = 1, \ldots, C$. As a consequence an estimate of θ using data aggregates is

$$\sum_{c=1}^{C} \frac{n_c}{N} g'_r(s_c, \hat{\theta}, \vartheta) = 0, \text{ for all } r = 1, \ldots, R. \tag{5.9}$$

In the empirical setting described above, the EE condition (5.7) can be used to define the additional condition (C6): $\sum_{c=1}^{C} p_c g'_r(s_c, \theta) = 0$ for all $r = 1, \ldots, R$. Then letting $\lambda = (\lambda_1, \ldots, \lambda_R)$ be the vector of Lagrange multipliers associated with (C6), constraint maximisation of (5.8) gives

$$\hat{q}_n^{(c)} = \frac{1}{N\{1 + \lambda^\top g'(s_c; \theta)\}}, \; c = 1, \ldots, C,$$

refer to Appendix B.1.1 for a proof. Setting $T(\theta; \boldsymbol{s}) = \frac{L(\hat{\boldsymbol{q}}; \boldsymbol{s})}{L(\boldsymbol{q}; \boldsymbol{s})}$ with $\boldsymbol{q}$ the vector of probabilities, then variances and confidence intervals for $\hat{\theta}$ are establish using $2\log(T(\theta; \boldsymbol{s})) = -2\sum_{c=1}^{C} n_c \log(N\hat{q}_n^{(c)}) \to \chi^2_R$.

The symbolic estimating equations (5.7) are obtained by integrating the estimating equation (5.1) over all the data points $\boldsymbol{x}$ from which the symbols $\boldsymbol{s}$ are produced with their corresponding weights, i.e.

$$\mathbb{E}_{F_S}[g'_r(S, \theta, \vartheta)] = \int_{\mathcal{D}_X^N} \mathbb{E}_{F_X}[g_r(X, \theta)] f_{\mathbf{S}|\boldsymbol{X}}(\boldsymbol{s}|\boldsymbol{x}, \vartheta)\phi(\boldsymbol{x})d\boldsymbol{x}.$$

Noting that $f_{\mathbf{S}|\boldsymbol{X}}(\boldsymbol{s}|\boldsymbol{x}, \vartheta) = \mathbb{1}\left\{\pi\left(\boldsymbol{x}^{(c)}\right) = s_c; c = 1, \ldots, C\right\}$ and using (5.3) yields

$$\begin{aligned}
\mathbb{E}_{F_S}[g'_r(S, \theta, \vartheta)] &= \int_{\mathcal{D}_X^N} \mathbb{1}\left\{\pi\left(\boldsymbol{x}^{(c)}\right) = s_c; c = 1, \ldots, C\right\} \left\{\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} \mathbb{1}\{x_n \in \boldsymbol{x}^{(c)}\} g_r(x_n; \theta)\right\} \phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} \mathbb{1}\{x_n \in \boldsymbol{x}^{(c)}\} \int_{\mathcal{D}_X^N} \mathbb{1}\left\{\pi\left(\boldsymbol{x}^{(c)}\right) = s_c\right\} g_r(x_n; \theta)\phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \sum_{c=1}^{C} \frac{n_c}{N} \int_{\Upsilon_c} \phi_c(x) g_r(x; \theta)\mathrm{d}x,
\end{aligned}$$

from which we conclude that $g'_r(s_c, \theta, \vartheta) = \int_{\Upsilon_c} \phi_c(x) g_r(x, \theta)\mathrm{d}x$ for all $r = 1, \ldots, R$ and $c = 1, \ldots, C$. The symbolic estimating function g' for s_c corresponds to the average of the regular estimating function g weighted by the density ϕ_c.

If a parametric assumption is made then ϕ_c corresponds to the model density restricted to Υ_c and it is easy to show that $g'(s_c, \theta) = \frac{\partial}{\partial\theta}\log\{\mathbb{P}(X \in \Upsilon_c|\theta)\}$ and that for $\theta = (\theta_1, \ldots, \theta_R)$ the solution of (5.9) is equivalent to maximum likelihood estimator of (5.4). Considering the example of intervals constructed from l-th and u-th order statistics given at the end of Section 5.2, it

is straightforward that these can be re-written as the product of five "sub-symbols" and fit in the context of (5.9).

The notion of sufficiency for estimating equations is described in Elashoff and Ryan (2004) as follows. Consider a vector-valued function $\beta^c(\boldsymbol{x}) = \left(\beta_1^c\left(\boldsymbol{x}^{(c)}\right), \ldots, \beta_R^c\left(\boldsymbol{x}^{(c)}\right)\right)$ and two distinct datasets $\boldsymbol{x}$ and $\boldsymbol{x}'$ with c-th subsets indicated by $\boldsymbol{x}^{(c)}$ and $\boldsymbol{x}'^{(c)}$. If $\beta^c(\boldsymbol{x}) = \beta^c(\boldsymbol{x}')$ for all $c = 1, \ldots, C$, then using the empirical likelihood we obtain

$$\frac{1}{n_c}\sum_{n=1}^{n_c} g(x_n^{(c)}; \theta) = \frac{1}{n_c}\sum_{n=1}^{n_c} g(x_n'^{(c)}; \theta) = A(\theta) + \sum_{r=1}^{R} k_r(\theta)\beta_r^c(\boldsymbol{x})$$

for some functions $A(\theta)$, $k_r(\theta)$, $r = 1, \ldots, R$ and an estimate of θ can be obtained by solving

$$\mathbb{E}_{F_X}\left[g\left(\boldsymbol{X}, \hat{\theta}\right)\right] = \sum_{c=1}^{C}\frac{n_c}{N}\left(A(\theta) + \sum_{r=1}^{R} k_r(\theta)\beta_r^c(\boldsymbol{x})\right) = 0.$$

We now show give an analog definition of sufficiency for symbolic estimation equations. Since we have shown that $g'(s_c, \theta, \vartheta) = \mathbb{E}_{\phi_c}[g(X, \theta)]$ for all $r = 1, \ldots, R$, we can easily show that

$$g'(s_c, \theta, \vartheta) = A(\theta) + \sum_{r=1}^{R} k_r(\theta)\beta_r'^c(\boldsymbol{s}),$$

where $\beta_r'^c(\boldsymbol{s}) = \mathbb{E}_{\phi_c}[\beta_r^c(\boldsymbol{X})]$, using the relationship $\mathbb{E}_{\phi_c}[g(X, \theta)] = \mathbb{E}_{\phi_c}[\mathbb{E}_{F_X}[g(\boldsymbol{X}, \theta)]]$. As a result, if two distinct sets of symbols yield to the same value for all $\beta_r'^c$ then the same estimates of θ are obtained from solving the symbolic estimating equations (5.9). Furthermore if, for all $c = 1, \ldots, C$, there exist a function $q^c(\boldsymbol{s}) = (q_1^c(s_c), \ldots, q_R^c(s_c))$ such that $q^c(\boldsymbol{s}) = \beta^c(\boldsymbol{x})$ then it implies $\beta'^c(\boldsymbol{s}) = \beta^c(\boldsymbol{x})$. As a consequence the both estimates from full and aggregated data are equal since $\mathbb{E}_{F_X}\left[g\left(\boldsymbol{X}, \hat{\theta}\right)\right] = \mathbb{E}_{F_S}\left[g'\left(\boldsymbol{S}, \hat{\theta}, \vartheta\right)\right] = 0$.

Returning to the example where $\theta = (\mu, \sigma^2)$, let's assume that $\beta^c(\boldsymbol{x}) = (\mu_c, \sigma_c^2)$ is the vector of sample mean and sample variance from the subset $\boldsymbol{x}^{(c)}$. We can write

$$\mathbb{E}_{F_X}\left[g\left(\boldsymbol{X}, \hat{\theta}\right)\right] = \left(\frac{1}{N}\sum_{c=1}^{C} n_c(\mu_c - \hat{\mu}), \frac{1}{N}\sum_{c=1}^{C} n_c((\mu_c - \hat{\mu})^2 + \sigma_c^2 - \hat{\sigma}^2)\right) = 0$$

Thus θ can be estimated from a set of symbolic data through the evaluation of $\hat{\mu}_c = \mathbb{E}_{\phi_c}(X)$ and $\hat{\sigma}_b^2 = \mathbb{V}_{\phi_c}(X)$, the expected value and variance of the observations that were aggregated into the c-th symbol, under some hypothetised distribution. If μ_c and σ_c^2 were recorded in the aggregation process then using complete or aggregated data would obviously lead to the same estimates.

In the following sections we will make use of μ_c, σ_c^2 but also the skewness $\gamma_c = \mathbb{E}_{F_{X^c}}[((X^{(c)} - \mu_b)/\sigma_b)^3]$ and correlation between the d-th and e-th components $\rho_{cde} = \mathbb{E}_{F_{X^c}}[(X_{[d]}^{(c)} - \mu_{c[d]})(X_{[e]}^{(c)} - \mu_{c[e]})]\sigma_{c[d]}^{-2}\sigma_{c[e]}^{-2}$, where F_{X^c} defines the empirical distribution restricted to the c-th symbol. This

quantities will be estimated by taking expected values with respect to ϕ_c for which we propose a novel estimation procedure.

5.4 Estimating the within-symbol density

In the previous section the construction of estimating equations has been shown to depend on ϕ_c the density of the underlying (unobserved) random variable given some observed summary of a group c. In this section we focus on estimating these densities for classes c where $c = 1, \ldots, C$ with summaries taking the form of intervals or histograms. In particular information about the splitting process into subsets is incorporated. The key idea is that if the attribution to a class c is done completely at random then a point in class c could have been attributed to a different class c', leading to different aggregates $\boldsymbol{s}$ and $\boldsymbol{s}'$ and consequently the density ϕ_c should not only rely on the symbol s_c.

The most popular approach to estimate ϕ, the density of the underlying random variable given a set of summaries, does not take into account any information about the class allocation process and assumes the summaries to be independent as in (5.6). Bertrand and Goupil (2000), Billard and Diday (2003), Le Rademacher and Billard (2011) and Oliveira et al. (2018) utilise this approach to estimate symbolic means, variances and covariances from interval or histogram-valued observations, assuming ϕ_c to be a uniform distribution. We will refer to this methodology as the symbolic independent uniforms model (SIU). For each class $c = 1, \ldots, C$ and dimension $d = 1, \ldots, D$ we have a $\alpha_c = (\alpha_{c,l}, \alpha_{c,u})$ defining the upper and lower bounds of an interval or the range of a histogram bin which gives the estimates $\hat{\mu}_c = \frac{\alpha_{c,l}+\alpha_{c,u}}{2}$, $\hat{\sigma}_c^2 = \frac{(\alpha_{c,u}-\alpha_{c,l})^2}{12}$, $\hat{\gamma}_{cd} = 0$ and $\hat{\rho}_{cde} = 0$. Oliveira et al. (2018) and Le Rademacher and Billard (2011) extend this methodology to other distributions such as independent triangular distributions or Dirac distributions centred on the midpoints or endpoints of each summary.

We consider that the attribution of a realisation to a class is random, i.e. that there is a discrete random variable Λ_x taking values in $\{1, \ldots, C\}$ indicating the list of symbols in which an observation x could have been aggregated in. We assume Λ_x has distribution function H_x. We introduce a share of the information by re-defining ϕ, the density of a point given the observed set of symbols $\boldsymbol{s}$, as the integral of the density $f_{X|\boldsymbol{S}}$ weighted by the density of H, i.e.

$$\phi(x) = \int_{\mathcal{D}(\Lambda_x)} f_{X|\boldsymbol{S}=\boldsymbol{s}'}(x) \mathrm{d}H_x(\lambda) \mathrm{d}\lambda \tag{5.10}$$

where $\boldsymbol{s}'$ denotes the set of symbols given that the observation x is assumed to be grouped in the λ-th class. It can be noticed that if an observation x can only be associated to a unique symbol s_c then H_x is a Dirac delta function at the correct allocation of an observation to a class c then (5.10) reduces to $\phi(x) = f_{X|\boldsymbol{S}=\boldsymbol{s}'}(x)$. The density of an observation given a symbol s_c is given by

$$\phi_c(x) = \frac{\mathbb{1}\{x \in s_c\}\phi(x)}{P_H^{s_c}}, \tag{5.11}$$

where the normalising term $P_H^{s_c} = \int_{\Upsilon_c} \phi(x) \mathrm{d}H_x(c) \mathrm{d}x$ is an integral with respect to all points that

could have been aggregated into s_c. Because this construction of ϕ and ϕ_c relies on borrowing information from adjacent symbols we refer to this model as the symbolic dependent distributions (SDD).

Note that in the literature there is a prevalence of models assuming uniformity over a summary which, in the current context, translate to $f_{X|S=s_c} = |\Upsilon_c|^{-1}$. It is worth highlighting that it may not always be a good strategy to borrow information from neighbouring summaries. For example in the mushroom dataset analysed by de A Lima Neto et al. (2011), each group corresponds to different species with possibly very different characteristics.

The following subsections will focus on the specific cases where data are aggregated into intervals and histograms with the uniformity assumption within the summary and where it is justified to borrow information with adjacent symbols. We will show that for independent intervals the distribution H_x is a discrete distribution whereas histograms can be interpreted as a set of deterministic intervals and in that case H_x is continuous.

5.4.1 Interval-valued data

In this subsection subsets $\boldsymbol{x}^{(c)}$ of $\boldsymbol{x}$ are aggregated into D-dimensional intervals resulting in the summaries s_c with $\alpha_c = (\alpha_{c,l}, \alpha_{c,u})$ where $\alpha_{c,l}$ and $\alpha_{c,u}$ are D-dimensional vectors denoting the lower and upper bounds, and the region $\Upsilon_c = [\alpha_{c,l}, \alpha_{c,u}]$. For an observation x, if it belong to a region where some intervals from the set $\boldsymbol{s}$ overlap then the random variable Λ_x has a discrete outcome λ in the form of a list of two or more values in $\{1, \dots, C\}$ with non-zero probability. For all $x \in \mathbb{R}^D$ the distribution Λ is discrete and we propose its weights to be defined as

$$H_x(\lambda, c) = \frac{n_c}{\sum_{a \in \lambda} n_a}, \text{ for all } c \in \lambda,$$

meaning that the probability that an observation would have been aggregated into a symbol s_c is proportional to the number of points in this symbols.

In order to incorporate the information about H_x in the estimation of $\phi(x)$ and $\phi_c(x)$ the range of $\boldsymbol{x}$ is split into a grid of subintervals denoted by $\upsilon_{\boldsymbol{b}}$ with $\boldsymbol{b} = (b_1, \dots, b_D)$ and $b_d = 1, \dots, (2C+1); d = 1, \dots D$. The subinterval $\upsilon_{\boldsymbol{b}}$ defines the region $(z^1_{b_1-1}, z^1_{b_1}) \times \dots \times (z^D_{b_D-1}, z^D_{b_D})$ where $z^d_0, \dots, z^d_{2C+1}$ is the ordered sequence of the d-th component of $(\alpha_{c,l}, \alpha_{c,u}), c = 1, \dots, C$. The density ϕ given in (5.10) simply reduces to

$$\phi(x) = \sum_{c=1}^{C} \frac{n_c}{N|\Upsilon_c|} \mathbb{1}\{x \in \Upsilon_c\},$$

while the normalising term in (5.11) is $P_H^{s_c} = m_c(1)$ where the function $m_c(f(x)) = \int_{\Upsilon_c} f(x)\phi(x)\mathrm{d}H_x(c)\mathrm{d}x$ is given by

$$m_c(f(y)) = \frac{1}{N} \sum_{c'=1}^{C} \frac{n_{c'}}{|\Upsilon_{c'}|} \left(\sum_{\boldsymbol{b}} \mathbb{1}\{\upsilon_{\boldsymbol{b}} \subset \Upsilon_{c'}\} H_{\boldsymbol{b}}(\lambda, c) \int_{\upsilon_{\boldsymbol{b}}} f(y)\mathrm{d}y \right),$$

using the notation $H_{\boldsymbol{b}}(\lambda, c) = H_y(\lambda, c)$ for $y \in \upsilon_{\boldsymbol{b}}$ and $\lambda = \{c = 1, \dots, C \text{ s.t. } \upsilon_{\boldsymbol{b}} \subset \Upsilon_c\}$.

Proposition 5.4.1. *The estimates of the mean, variance, skewness and correlation within an interval c are obtained using the density ϕ_c and are given as follows*

$$\hat{\mu}_{cd} = \frac{m_c\left(x_{[d]}\right)}{m_c(1)}, \quad \hat{\sigma}^2_{cd} = \frac{m_c\left(x^2_{[d]}\right)}{m_c(1)} - \hat{\mu}^2_{cd}, \quad \hat{\gamma}_{cd} = \frac{m_c\left(x^3_{[d]}\right)}{m_c(1)\hat{\sigma}^3_{cd}} - \frac{\hat{\mu}^3_{cd}}{\hat{\sigma}^3_{cd}} - 3\frac{\hat{\mu}_{cd}}{\hat{\sigma}_{cd}}, \quad \hat{\rho}_{cde} = \frac{m_c\left(x_{[d]}x_{[e]}\right)}{m_c(1)\hat{\sigma}_{cd}\hat{\sigma}_{ce}} - \frac{\hat{\mu}_{cd}\hat{\mu}_{ce}}{\hat{\sigma}_{cd}\hat{\sigma}_{ce}},$$

for $d = 1, \ldots, D$.

Appendix B.1.2 provides the technical details for the evaluation of the integral in m_c.

5.4.2 Histogram-valued data

Suppose now that $\boldsymbol{x}$ is aggregated into a $D-$dimensional histogram $\boldsymbol{s} = (s_{\mathbf{1}}, \ldots, s_{\boldsymbol{C}})$ where $s_c = (n_{\boldsymbol{c}}, \Upsilon_{\boldsymbol{c}}, \alpha_{\boldsymbol{c}})$ is a summary of the information in the bin $\boldsymbol{c} = (c^1, \ldots, c^D)$ with $c^d = 1, \ldots, C^d$ and $d = 1, \ldots, D$. For a deterministic bin locations $\Upsilon_{\boldsymbol{c}} = \left((y^1_{c_1-1}, y^1_{c_1}) \times \cdots \times (y^D_{c_D-1}, y^D_{c_D})\right)$ we define $n_{\boldsymbol{c}} = \alpha_{\boldsymbol{c}} = \sum_{n=1}^{N} \mathbb{1}\{x_n \in \Upsilon_{\boldsymbol{c}}\}$. Assume the marginal bin width to be equal and given by $\delta_d = y^d_{c_d} - y^d_{c_d-1}$ for all c_d and d such that the area of each bin is $|\Upsilon_{\boldsymbol{c}}| = \prod_{d=1}^{D} \delta_d$. This histogram construction can be thought as $C^1 \times \cdots \times C^D$ non-overlapping histograms and as a consequence $P_H^{s_c} = \int_{\Upsilon_c} \phi(y)\mathrm{d}H_y(c)\mathrm{d}y = \int_{\Upsilon_c} \phi(y)\mathrm{d}y$ since y can only belong to s_c.

The choice of the bin locations $\Upsilon_{\boldsymbol{c}}$ is arbitrary and thus, keeping their width constant, other histograms could have arisen by shifting the bin locations by up to half of their width to the left or to the right. We define new bin locations $\Upsilon' = \Upsilon + \boldsymbol{u}$ where $\boldsymbol{u} = (u_1, \ldots, u_D)$ and u_d is a realisation from a uniform distribution on $\left(-\frac{\delta_d}{2}, \frac{\delta_d}{2}\right)$. The random variable Λ_x representing the bin Υ in which x could have been aggregated to is now a continuous as the set of outcomes contains all the shifted bins Υ' that will include x. Because of Υ is fixed and the shift is uniformly distributed we can write

$$H_x(\Upsilon'_c) = \frac{1}{\prod_{d=1}^{D} \delta_d} \mathbb{1}\{x \in \Upsilon'_c\}$$

Assuming that histogram bins can be shifted relates to the ideas of Heitjan and Rubin (1991) who described the conditions in which a parametric analysis of coarsened data remains the same regardless of whether the likelihood of the coarsening process itself was incorporated. As a result, no additional attention needs to be paid to the coarsening process for binned data, which in practise is equivalent to assuming a uniform distribution for all possible coarsening configurations, given the fixed parameters such as number of bins, their locations and width. We now use this concept in a non-parametric setting to determine within-symbol quantities such as means, variances, etc.

First, from there the density ϕ given in (5.10) simply reduces to

$$\phi(x) = \frac{1}{\prod_{d=1}^{D} \delta_d^3} \sum_{\boldsymbol{c'}=\mathbf{1}}^{\boldsymbol{C}} \sum_{\boldsymbol{c''}=\boldsymbol{c'}-\mathbf{1}}^{\boldsymbol{c'}+\mathbf{1}} \frac{n_{\boldsymbol{c''}}}{N} J_{\boldsymbol{c'},\boldsymbol{c''}}(x), \tag{5.12}$$

while the normalising term in (5.11) is $P_H^{s_{\boldsymbol{c}}} = m_c(1)$ where the function $m_c(f(x))$ is given by

$$m_c(f(x)) = \frac{1}{\prod_{d=1}^{D} \delta_d^3} \sum_{\boldsymbol{c'}=\boldsymbol{c}-\mathbf{1}}^{\boldsymbol{c}+\mathbf{1}} \sum_{\boldsymbol{c''}=\boldsymbol{c'}-\mathbf{1}}^{\boldsymbol{c'}+\mathbf{1}} \frac{n_{\boldsymbol{c''}}}{N} \int_{\Upsilon_{\boldsymbol{c}}} f(x) J_{\boldsymbol{c'},\boldsymbol{c''}}(x) \mathrm{d}x.$$

Refer to Appendix B.1.3 for technical details to obtain (5.12) and a definition of $J_{\boldsymbol{c'},\boldsymbol{c''}}(x)$.

Proposition 5.4.2. *The estimates of the mean, variance, skewness and correlation within a histogram bin* $\boldsymbol{c}$ *are given as follows*

$$\hat{\mu}_{\boldsymbol{c}d} = \frac{m_c\left(x_{[d]}\right)}{m_c(1)}, \quad \hat{\sigma}^2_{\boldsymbol{c}d} = \frac{m_c\left(x^2_{[d]}\right)}{m_c(1)} - \hat{\mu}^2_{\boldsymbol{c}d}, \quad \hat{\gamma}_{\boldsymbol{c}d} = \frac{m\left(x^3_{[d]}\right)}{m_c(1)\hat{\sigma}^3_{\boldsymbol{c}d}} - \frac{\hat{\mu}^3_{\boldsymbol{c}d}}{\hat{\sigma}^3_{\boldsymbol{c}d}} - 3\frac{\hat{\mu}_{\boldsymbol{c}d}}{\hat{\sigma}_{\boldsymbol{c}d}}, \quad \hat{\rho}_{\boldsymbol{c}de} = \frac{m\left(x_{[d]}x_{[e]}\right)}{m_c(1)\hat{\sigma}_{\boldsymbol{c}d}\hat{\sigma}_{\boldsymbol{c}e}} - \frac{\hat{\mu}_{\boldsymbol{c}d}\hat{\mu}_{\boldsymbol{c}e}}{\hat{\sigma}_{\boldsymbol{c}d}\hat{\sigma}_{\boldsymbol{c}e}},$$

for $d = 1, \ldots, D$.

The necessary integrals to evaluate m_c require tedious but straightforward calculations that can be found in the Supplementary Material.

5.5 Simulations

The aim of the experiments presented in this section is to demonstrate that symbolic estimating equations have the ability to give good estimates of some statistics of interest. In particular we focus on the scenario where simulated data are aggregated into symbolic objects taking the form of intervals or histograms. Furthermore the interest is in measuring the impact of the level of aggregation on the precision of the estimates since symbolic and classical methods are expected to provide identical results in the limiting case where a symbolic object collapses to a single datum.

5.5.1 Univariate examples

We generate $N = 2\,000$ observations from a standard normal distribution, a standard skew-normal with shape $\alpha = 20$ (implying that $\gamma = 0.99$) and a mixture of three skew-normal with location $\mu = -3, 0, 3$, variance $\sigma^2 = 1, 1, 2$, shape $\alpha = -10, -20, -50$ respectively, and probabilities $1/14, 4/14$ and $9/14$. In order to make comparisons between summary functions, the simulated data is aggregated under two scenarii: first into a histogram with C equally spaced bins and second into C minimum/maximum intervals. The performance of the naive but widely used SIU approach is compared to the proposed SDD approach with respect to their ability to estimates various statistics such as the mean, variance, skewness and some quantile levels. A comparison to the classical counterparts is also made to measure the effect of information loss and to assess the effectiveness of both aggregation functions. Experiments are repeated 1 000 times.

Billard and Diday (2003) proposed a histogram estimator of a set $\boldsymbol{s}$ of intervals by partitioning the domain of the underlying process into R subintervals I_r, $r = 1, \ldots R$ representing the

histogram bins and defining the respective probabilities by

$$p_r = \sum_{c=1}^{C} \frac{n_c}{N} \frac{|I_r \cap \Upsilon_c|}{|\Upsilon_c|}, \quad r = 1, \ldots, R, \tag{5.13}$$

where $|A|$ denotes the area of the region A. Both SIU and SDD approaches will also be applied to these histogram estimate with $R = 30$ equally spaced bins.

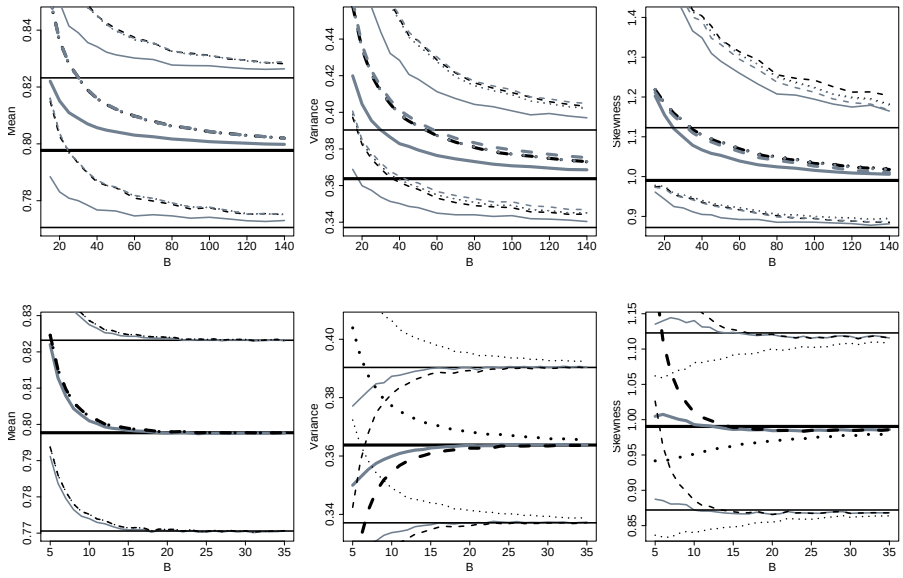

Figure 5.1: Estimates and 95% confidence intervals of the mean (left), variance (centre) and skewness (right) as function of C the number of symbols when the aggregates take the form of intervals (top) or histograms (bottom). Classical estimates are given by black horizontal lines, SIU estimates by dotted black lines, SDD estimates by solid grey lines and rounded analysis with Sheppard's correction by dashed black lines. Estimates obtained from the estimated histogram (5.13) are given in the top row respectively by dashed and dotted grey lines for the SIU and SDD approaches. Original data were simulated from the skew-normal distribution.

Figure 5.1 illustrates the ability of both interval- (top) and histogram- (bottom) based methodologies to recover the classical estimates (solid black line) as the number of aggregates (i.e. the level of information retained) increases. Each line corresponds to the mean and 95% confidence interval for a fixed number of aggregates, calculated from 1 000 replicates. The following conclusions drawn from Figure 5.1 focus on the scenario of skew-normal underlying data, however similar conclusions can be made for the normal and mixture of skew-normal distributions as per Figure B.1 and B.2 in Appendix B.1.4. Comparing the different estimation procedures it is obvious that the SDD approach consistently provides the most accurate results for all each statistic studied here and in particular for low numbers of symbols (up to 50 for intervals and up to 10 for histograms). Unsurprisingly histograms provide better estimates than intervals when comparing estimation procedure since, by construction, they retain a greater amount of information. For example, comparing the solid grey lines between the top and bottom panels highlights that the SDD approach yields estimates and confidence intervals much closer to the

classical ones using histograms rather than intervals. For ≈ 20 histogram-valued symbols the estimates and confidence intervals can barely be dissociated from their classical counterparts whereas ≈ 20 interval-valued symbols give a much rougher approximation. The dashed black lines represent estimates where a Sheppard's correction has been applied, to the estimated histograms obtained from (5.13) (top row) or to the midpoints of the bins (bottom row). These estimates are often not as accurate as the SDD estimates and in most cases are only similar to those obtained through the SIU approach. The exception is noted for the variance of the normal distribution when a low number of symbols are used (see Figure B.1). Overall the approaches investigated here have different levels of over-estimation of the true statistics which is expected since some information has been lost in the aggregation procedure. However we note that for histogram-valued data the variance is under-estimated (bottom middle panels), a phenomena that can be observed for the estimation of all statistics when the data are drawn from a mixture of skew-normal distributions (see Figure B.2). These results are influenced by the estimation of the densities ϕ and $\phi_c, c = 1, \ldots, C$ and their estimation. We have shown that considering dependence between classes/symbols in the symbolic estimation setting significantly improves the current methods available in the literature (SIU approach), in particular for small symbolic sample sizes (C) and skewed, multi-modal distributions.

5.5.2 Bivariate intervals simulations

The aim of the following experiments is to assess the ability of the proposed SSD methodology to estimate statistics from bivariate aggregated data. We generate $N = 2\,500$ observations from two bivariate normal distributions with marginal means $\mu_1 = -1$, $\mu_2 = 1$, marginal variances $\sigma_1^2 = \sigma_2^2 = 1$ and respective correlations $\rho = -0.2$ and 0.9, and two bivariate skew-normal distributions with the same mean and covariance matrices as above and respective skewness $\alpha = (\alpha_1, \alpha_2) = (1, 0.5)$ and $(6, 3)$. The simulated data is aggregated into C minimum/maximum rectangles. Similarly to the previous section, the experiment is repeated 1 000 times to allow for a comparison of the SIU, SDD and classical approaches which also includes a naive approach that consists in a classical analysis performed on the midpoints of each observed rectangle. The statistics of interest are the marginal means and variances as well as the covariance which can be calculated from Proposition 5.4.1. Figure 5.2 illustrates the ability to estimate some statistics of interest, $\mu_2, \sigma_1^2, \sigma_{12}$ and $\mu_2, \gamma_1, \sigma_{12}$ respectively using data drawn from a normal (top row) and skew-normal (bottom row) distribution and aggregated into rectangles. Figure B.3 provides similar results for the remaining two distributions. The estimates obtained using the underlying data are used as reference and are represented by a solid black line. It appears that the SDD estimates (solid grey line) provide the best estimates of the marginal means, variances and skewnesses and of the covariance for larger number of symbols (> 150). For small numbers of symbols, the naive estimator calculated from the midpoints seems to yield estimates closer to the classical results than the SDD approach for the marginal variance σ_1^2 of the normal distribution and marginal skewness γ_1 of the skew-normal distribution. Convergence towards the empirical covariance appears very slow for the naive estimator whereas the SDD estimates

converge at a much faster rate. To conclude we observe an increase in the number of symbols from the univariate experiments in order to obtain comparable results than an analysis on the full data and overall the SDD methodology provides a clear improvement in the evaluation of some statistics when only aggregates are available.

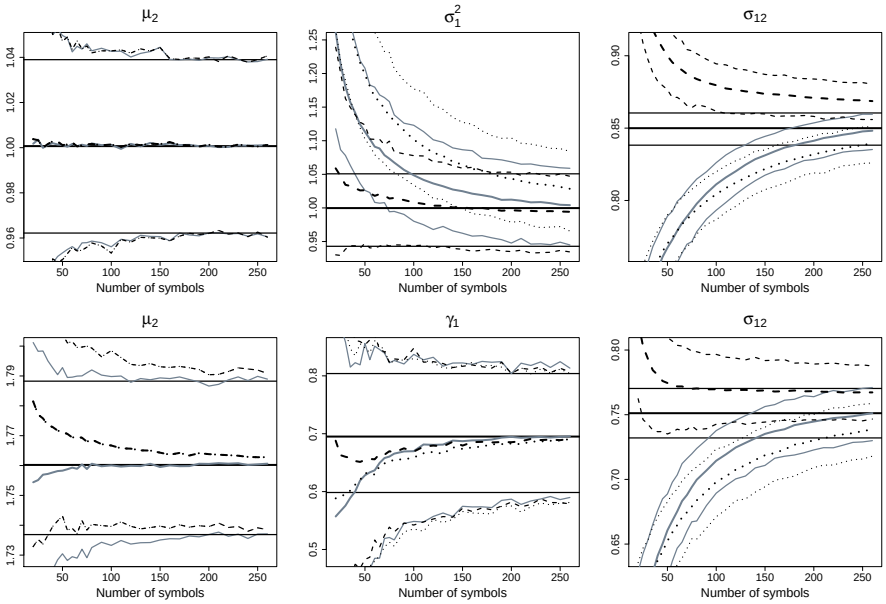

Figure 5.2: Estimates and 95% confidence intervals for some of the statistics of interest as a function of C the number of rectangles. Classical estimates are given by black horizontal lines, SIU estimates by dotted black lines, SDD estimates by solid grey lines and estimates using solely the midpoints of each rectangles by a black dashed line. Original data were simulated from a normal distribution with correlation $\rho = 0.9$ (top row) and skew-normal distribution with $\alpha = (6, 3)$.

5.6 Real Data Analyses

Through the analysis of several datasets with non-standard (i.e. not pointwise) observations we now demonstrate the utility of the symbolic estimating equation framework developed in this paper. These analyses are performed in the context where the underlying data are not available and the only information given is the result of some aggregation function. The first two datasets are respectively sets of observed intervals and histograms from which simple statistics such as means and variances are estimated. The third example focuses on estimating quantile levels from a histogram.

5.6.1 Soccer Dataset

In this analysis, we investigate the weight, height and age of 531 soccer players from 20 of the French Football Professional Championship. The dataset is freely available from the R package `iRegression` and consists in the minimum and maximum records for each team. This interval-valued dataset has been studied by de A Lima Neto et al. (2011) where the authors aimed to

predict the weight of soccer players as a function of height and age. In this particular example, the assumption that a point (player) can contribute to multiple intervals (team) seems reasonable given that players are allowed to move between teams during or in between Championship seasons. Additional information about the number of players per team was also retrieved from the SODAS software. These range from 22 to 30 players, weakening the assumption of equal number of players per team used by de A Lima Neto et al. (2011). Table 5.1 gives the estimates and 95% confidence intervals of the marginal means and variances as well as correlations using the SIU and SDD approaches on the 20 observed 3-dimensional hyperrectangles. Note that the confidence intervals are computed using the asymptotic distribution linked to the symbolic empirical likelihood as seen in Section 5.3. We can observe that some of the 95% confidence intervals from the SIU and SDD methods barely overlap or not at all, in particular for the variances and correlations. For example the SIU approach gives a 95% confidence interval for ρ_{13} that contains 0 whereas the SDD approach yields and interval only defined on the negative part. Based on the conclusions of Section 5.5 we believe the SDD estimates to be the closest to the truth. Such discrepancies in the estimation of the correlation coefficients can have a major impact on the estimates of linear regression coefficients and thus lead to incorrect prediction errors.

Statistic	SIU			SDD		
μ_1	73.340 (	73.965,	74.506)	74.313 (	74.056,	74.504)
μ_2	180.685 (	179.759,	180.275)	180.005 (	179.756,	180.210)
μ_3	25.269 (	26.199,	26.432)	26.201 (	26.094,	26.298)
σ_1^2	47.844 (	45.812,	50.355)	44.487 (	42.704,	46.238)
σ_2^2	48.590 (	46.490,	50.692)	45.221 (	43.452,	46.974)
σ_3^2	20.666 (	20.067,	21.244)	20.214 (	19.697,	20.815)
ρ_{12}	0.049 (	0.043,	0.065)	0.057 (	0.044,	0.069)
ρ_{13}	-0.003 (	-0.009,	0.006)	-0.006 (	-0.011,	0.000)
ρ_{23}	0.017 (	0.011,	0.022)	0.013 (	0.007,	0.016)

Table 5.1: Estimates and 95% confidence intervals for the means, variances and correlations of the weight, height and age of soccer players using the SIU and SDD methodologies.

5.6.2 Weight Dataset

Billard and Diday (2003, Table 7) uses a weight (kg) dataset in histogram form to illustrate their proposed methodology to compute the mean and variance from data aggregates. This histogram is obtained by combining histograms from seven age groups using (5.13) with $R = 10$. Through the SIU approach we estimate a symbolic mean of 143.9, symbolic variance of 485.4 and symbolic skewness of -0.272 whereas for the SDD approach these quantities refer directly to the underlying data and are respectively estimated as $143.9, 453.9$ and -0.295. As expected, both approaches yield comparable estimates for the mean, however there is a significant difference in the variance and skewness estimates. While the authors in Billard and Diday (2003) do not claim that their method estimates these quantities at the underlying data level, and instead define their

estimates as symbolic equivalents of the classical versions, we have shown in Section 5.3 that their formulas are equivalent to the sample mean and variance of the microdata if it uniformly distributed within each bin. Given the continuous nature of the weight variable, we expect this assumption to be violated and thus our estimates to be closer to the truth. Furthermore, using the far more informative set of underlying histograms per age group(Billard and Diday, 2003, Table 6), the authors show that estimates of the mean and variance are 143.9 and 447.5 which implies that the variance estimate obtained from the SDD approach performed on the aggregated histogram is significantly more accurate than the SIU. In the previous section we have shown that the symbolic estimating equations depend on the symbolic empirical distribution which is given here through the bin proportions, however the sample sizes required for the computation of confidence intervals aren't available.

5.6.3 Protein solubility dataset

Assuming independent uniform distributions Dedduwakumara and Prendergast (2018) illustrate the ability to estimate quantiles from histogram-valued data on the protein solubilities (in %) dataset analysed in Niwa et al. (2009). It is worth noting that the original microdata is freely available at `http://www.taguchi.bio.titech.ac.jp/eng/paper-e/paper-e.html` and that three extra low count bins weren't included in the original analysis of (Niwa et al., 2009), and subsequently in Dedduwakumara and Prendergast (2018), yielding us to expect slightly different results from theirs when considering the SIU approach. Figure 5.3 presents the estimates (solid lines) and 95% confidence intervals (dashed lines) for the $10, 50, 75$ and 90% quantiles using the microdata (black) and histogram data (SIU in blue and SDD in red) for various number of bins C. Refer to the Supplementary Material for technical details about estimating equations for quantiles. Figure 5.3 shows convergence, as the number of bins (C) increases, of the SDD quantile estimates to the classical ones, at a faster rate than the SIU estimates. From $C = 15$, the estimates from the SDD method appear to be similar than when using the full information. For example, an estimate of the median using all the data is $42.48\,(40.41, 44.72)$ and for histograms with $C = 15$ the SIU approach gives $42.64\,(40.32, 44.9)$, the SDD approach $42.43\,(40.29, 44.60)$, while Dedduwakumara and Prendergast (2018) report an estimate of the median of $41.83\,(39.56, 44.10)$. Similarly an estimate of the 75% quantile using all the data is $78.65\,(77.14, 80.39)$ and for histograms with $C = 15$ the SIU approach gives $78.51\,(76.78, 80.28.9)$, the SDD approach $78.58\,(76.91, 80.28)$, while Dedduwakumara and Prendergast (2018) report a lower estimate of $77.83\,(75.982, 79.674)$ implying some difficulties in the tail. Overall the SDD approach provides the best surrogate to the classical estimate in the case where only a histogram summary of the data is available.

5.7 Discussion

In this research paper we have defined a framework for non-parametric estimating equations where data summaries are observed rather than the data itself. This extends the work of Be-

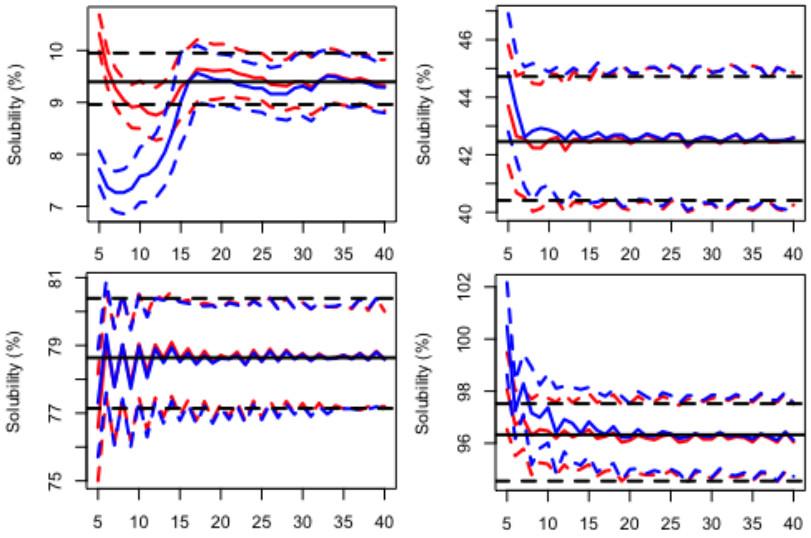

Figure 5.3: Estimates (solid lines) and 95% confidence intervals (dashed lines) for the $10, 50, 75$ and 90% quantiles of the protein solubility dataset using the microdata (black) and histogram data (SIU in blue and SDD in red) for various number of bins C.

ranger et al. (2018) which allows for parametric inference on such type of non-standard data and interpretation at the original microdata level. Estimating equations are commonly used to estimate means, variances, and other statistics. In the context of aggregated data, which we call "symbol", we show that our general methodology recovers the results of Bertrand and Goupil (2000) and Billard and Diday (2003) as special case when summaries are assumed independent with data uniformly distributed over the summary (interval or histogram). The latter uniformity assumption is over-prevalent in the symbolic data analysis literature despite being consistently violated due to the continuous nature of the underlying density of the microdata. We provided confidence interval for our estimates and have showed through synthetic and real examples that both estimates and confidence intervals converge as the summary collapses into a pointwise observation, to those in the classical setting. Furthermore we have highlighted that our proposed methodology is more flexible and consistently yields accurate estimates since it allows for information to be transmitted between summaries.

For interval-valued symbols, the major assumption made in the proposed (SDD) approach approach is that every classical latent observation could have possibly been aggregated into any of the intervals. For example, the soccer dataset considered in Section 5.6.1, it seems reasonable that a player from a particular team could have played for any other team and therefore could have contributed to this team's summary. An example of where this assumption might fail is in the mushroom dataset analysed in de A Lima Neto et al. (2011). There each group of observations correspond to a specific species and a mushroom from one species could not possibly have contributed to another summary.

As much as maximum likelihood estimation can be recovered from estimating equations, the symbolic likelihood function derived by Beranger et al. (2018) can be recovered as a special case of the symbolic estimating equations introduced in this article. If a parametric assumption is

undesirable then the SDD approach allows to estimate within-symbol statistics which are then used to obtain population level statistics. In order to do so we propose an estimation procedure for the within-symbol density without any distributional assumption about the underlying data. The main focus has been on non-parametric data summaries such as intervals and histograms.

Additionally to proposing better non-parametric estimates than the references in the literature, the setting described in this paper opens the door to the use of more flexible summaries. Indeed distributional-valued symbols can easily be used as estimates of the underlying data distribution for a particular class.

Chapter 6

Parameter estimation in Generalized Linear Models for rounded discrete data, with an application to the Athena SWAN award dataset.

6.1 Introduction

This paper is motivated by the evaluation of a diversity initiative in the UK, in which the effect of various variables on the proportion of females in various academic positions in different fields at university is investigated. There are two main questions which are being explored. The first question concerns the significance of various variables in the prediction of the number and proportions of females for various departments around universities in England, Scotland, Ireland and Wales, and the second question aims to determine the effect of these same variables, along with the proportions/counts of females, on the outcome of the Athena SWAN rating of these departments, which is an ordinal categorical rating assigned based on various factors revolving around the progressiveness and diversity of that department. Details of the Athena SWAN rating can be found at the Athena SWAN charter webpage https://www.ecu.ac.uk/equality-charters/athena-swan/. Due to data confidentiality reasons, rather than actual proportions, the number of males, number of females and total department sizes for each department at various employment levels are reported rounded to the nearest 5. We wish to apply a generalized linear model (glm) framework to this non-standard rounded dataset, with the aim of determining the useful predictors in the above questions.

Generalized Linear Models were first proposed by Nelder and Wedderburn (1972), and are now a well developed methodology for modelling the effect of covariate data on a broad range of types of response variables, such as continuous, ordinal, categorical and count data. As the name suggests, glm's are a generalisation of ordinary linear regression whereby the distribution of the differences between the predicted and observed response data is not restricted to the gaussian

family. McCullagh and Nelder (1989), Dunteman and Moon-Ho (2006) and Faraway (2010) all provide a good overview of the theory and application of glm's. Estimates for the regression coefficients in these models are generally obtained using maximum likelihood estimation, however unlike ordinary linear regression there usually isn't a closed form solution to the optimisation of the likelihood for glm's. As a result, maximum likelihood estimates (MLEs) are obtained via algorithms such as iteratively reweighted least squares (Nelder and Wedderburn, 1972) and the Newton-Raphson method (Jennrich and Sampson, 1976). These approaches are well developed for the case where the response and predictor variables are fully observed, i.e. they arrive in the form of standard pointwise data. However, it is becoming increasingly common for the practitioner to observe data (either predictor or response) in non-standard forms, due to reasons such as computational savings or privacy. Examples of these non-standard forms include the coarsening or rounding of some variables (Schneeweiss et al., 2010, Heitjan and Rubin, 1991) due to reasons such as privacy (Willenborg and de Waal, 1996, 2001) or a natural grouping of the observations (Haitovsky, 1983), or when the underlying classical data (microdata) is aggregated into distribution-valud objects such as intervals or histograms (Billard and Diday, 2000, Billard, 2011, Billard and Diday, 2003).

Armstrong (1985) consider the case where a single covariate is measured with error, and the distribution of the coarsened value, given the true latent observation can be assumed. The likelihood contribution of each coarsened observation is then calculated as the integral of the classical glm density over the domain of the latent value. Johnson (2006) assume Gaussian distributions for coarsened covariates and utilise this assumption to model their likelihood within the glm framework. Lee et al. (2018) consider the case where only a subset of observations are subject to coarsening, and assume a Bahadur type multivariate distribution for the modelling of whether a given observation is coarsened. These models require the specification of the distribution of the latent data given the coarsened observation, and as a result run into problems if these distributional assumptions are violated, or if not enough information is available from which parametric assumption can be made. Little (1993) develop a likelihood-based approach for the modelling of masked data, given a parametric assumption and fully observed additional variables. Lipsitz et al. (2004) propose a method for implementing a generalized linear model when one of the covariates is coarsened (binned) for only a subset of observations, whereby the contribution of each observation towards the likelihood is the integral of the glm density over all possible values that variable could have taken weighted by a density that is dependent on the uncoarsened fully observed variables. Estimates for glm coefficients are then obtained via the EM algorithm. Johnson and Wiest (2014) propose a simulated Bayesian approach for coarsened covariates in glm's in which a prior distribution is placed on the model parameters, and a distribution is assumed for the coarsened covariates, given the uncoarsened data.

These methods rely on the occurence of uncoarsened fully observed additional regression covariates, or that not every observation is subject to coarsening for the variables of interest, so that a distribution of the underlying latent microdata for the coarsened/rounded data can be fit using available classically observed data. However, these methods are not designed to accomodate situations for which it isn't possible to assume a parametric form for the coars-

ened/rounded data, or covariate or predictor data is coarsened for every observation and there are no appropriate additional classically observed variables from which the practitioner can fit a distribution for the underlying classical data to.

It is not possible to fit a model (parametric or otherwise) to the underlying values of the rounded observations of the Athena SWAN dataset, given every observation is subject to the rounding mechanism and the additional predictor variables are all categorical observations with small numbers of distinct categories, meaning any distribution fit using these covariates will likely not be very informative. In this paper we address these problems with the applied analysis described above by developing a methdology of estimating the parameters of glm's for rounded discrete data that incorporate additional variables not included in the original analysis. Estimates are obtained for various glm models by solving a set of estimating functions in which the contribution of each observation is calculated as the average of the estimating equation over all sets of underlying observations that could have occured, given the rounded observations. By using the rounded values for the numbers of males and total department size, as well as some available proportions for large departments, the domain of the underlying microdata for the rounded numbers of females and total department sizes can be greatly restricted.

As a simple example, if we observe values of 5, 5 and 15 respectively for the rounded numbers of females, males and people in a given department, and only the number of females is included in the original analysis, then without any additional thought the log-likelihood contribution for that observation reduces to a uniform summation over the set $\{3, 4, 5, 6, 7\}$ for that variable, with each possible underlying value is assigned a weight of $\frac{1}{5}$. However, if we utilise the fact that the numbers of females and males must sum up to the total department size, then the possible complete sets for that observation become $\{(6, 7, 13), (7, 6, 13), (7, 7, 14)\}$, and as a result a uniform summation over these sets within the likelihood for that observation leads to a only the values 6 and 7 being included for the numbers of females, with given weights $\frac{1}{3}$ and $\frac{2}{3}$ respectively. Given the misspecified nature of these log-likelihoods, the Godambe information matrix (Godambe, 1960) is used instead of the usual Fisher information matrix to determine the variances of the parameter estimates obtained from this analysis, allowing the evaluation of confidence intervals for each parameter, and subsequent inference to be performed.

This paper proceeds as follows. In Section 6.2 we describe the motivating dataset and define the relevant variables, taking note of which are subject to a rounding mechanism. In Section 6.3 we provide a brief overview of glm theory, including methods of estimation and inference. In Section 6.4 we provide the general framework for obtaining glm parameter estimates from rounded discrete data, with a focus on the utilisation of additional variables not originally included in the classical glm model that we wish to approximate. In Section 6.5 we then provide the specific models for the applied data analyses motivating this paper, along with simulations demonstrating their utility and results for the real data analyses. In Section 6.6 we conclude with a brief discussion of the impacts of this work, as well as potential extensions and directions for future work.

6.2 The Athena SWAN dataset

We will now describe the motivating dataset for this paper. We focus on data from the academic fields of Psychology and Physics, for the years 2012 and 2016, however similar analyses described in this paper can be performed for the greater dataset containing all academic fields. Each observation in each dataset represents a university cost centre that has engaged with the Athena SWAN process. We are interested in determining what variables are predictive of both the ordinal-valued Athena Award Status (Gold, Silver, Bronze or None) and also the number/proportion of females in a given academic cost centre. Given some of the observed count variables are subject to rounding, a classical glm analysis is insufficient. We now describe the dataset, along with the rounding procedures for some of the variables.

6.2.1 Data Description

For an observation x subject to rounding, denote x as the true underlying value, and x^* as the rounded observation, i.e. $x \to x^*$. For each observation $n = 1, \ldots, N$, every count variable is rounded to the nearest $R = 5$. The variables included in the dataset for each field and year are as follows.

— Athena SWAN award status : $a_n \in \{\text{None, Bronze, Silver, Gold}\}$

— Country: $l_n \in \{\text{England (1), Scotland (2), Wales (3), Northen Ireland (4)}\}$

— Research Intensity REF rating: $r_n \in \{1, 2, 3\}$, with higher values representing a more intense research rating

— Length of engagement with Athena SWAN process (months): $t_n \in \mathbb{N}$

— Total number of people: $n_n \in \mathbb{N} \to n_n^*$

— Proportion of females, unknown if $n_n < 22.5$: $p_n \in \{[0, 1], \emptyset\}$

— $w_{ni} \in \mathbb{N} \to w_{ni}^*$, $i = 1, \ldots, 4$: Number of females for the i^{th} employment level, $i \in \{\text{fixed term, professor, research, teaching}\}$

— Number of males for the i^{th} employment level: $v_{ni} \in \mathbb{N} \to v_{ni}^*$, $i = 1, \ldots, 4$:

— Proportion of females for the i^{th} employment level, unknown if $w_{ni} + v_{ni} < 22.5$.: $p_{ni} \in \{[0, 1], \emptyset\}$

Figure 6.1 shows histograms of the rounded total number of females and males for each department, constructed respectively as the rounded sums of the number of females and males in each employment level. Clearly the psychology datasets exhibit much higher proportions of females, while the physics datasets have a much higher prevalence of males. Furthermore, let $l_{ni} = 1$ if and only if $l_n = i$ and $l_{ni} = 0$ otherwise, $i = 1, \ldots, 3$, meaning Northern Ireland is used as the reference category. This allows regression coefficients to be obtained that describe the effect of belonging to each country. For each dataset, there is a maximum of 1 'Gold' Athena SWAN award status, making the fit obtained from an ordinal regression model unreliable. To rectify this, we group the 'Gold' and 'Silver' awards together, denoted simply as 'Silver+' and leaving us with $K = 3$ response categries. Only cost centres that have engaged with the Athena SWAN

process were included in each analysis, leaving $N = 104$ and $N = 58$ observations respectively for the academic fields of psychology and physics for each year. Each proportion value is only reported if the relevant group has a total size greater than $\tau = 22.5$, i.e. $p_{ni} = \emptyset$ if $w_{ni} + v_{ni} < \tau$ and $p_{ni} = \frac{w_{ni}}{w_{ni}+v_{ni}}$ otherwise, $i = 1, \ldots, 4$, and $p_n = \emptyset$ if $w_n + v_n < \tau$ and $p_n = \frac{w_n}{n_n}$ otherwise, where w_n represents the total number of females for the n^{th} department.

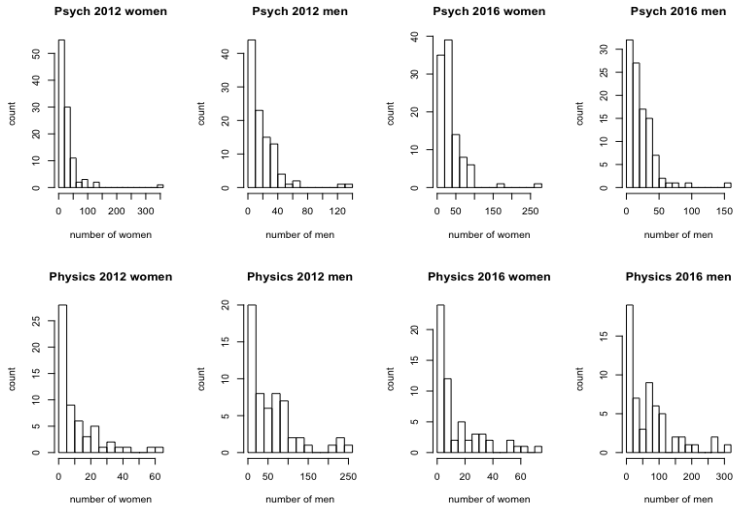

Figure 6.1: Histograms of the total (rounded) number of females and males for each dataset.

6.3 Generalized linear models for rounded discrete data

We now provide a brief overview of glm theory for classically observed variables originally proposed by Nelder and Wedderburn (1972), followed by a description of some specific examples. An overview of the available methodologies for the dataset analysed in this paper is then given. The theory described and examples listed are motivated by the real data analyses undertaken in Section 6.5, in which we wish to approximate the classical glm models by equivalent expressions for rounded discrete data.

6.3.1 Generalized Linear Models

Denote $Y_n \in \mathcal{D}_{Y_n}$ as a random response variable and $X_n = (X_{n1}, \ldots, X_{nD})^\top$, $n = 1, \ldots, N$ as N $D-$dimensional *i.i.d.* vectors of covariates from which we want to predict Y, with $X_{nd} \in \mathcal{D}_{X_{nd}}$, $n = 1, \ldots, N$. We assume the response Y_n is distributed according to some distribution in the exponential family, with the mean μ of the distribution depending on X_n via the link function g such that $\mathbb{E}(Y_n) = \mu_n = g^{-1}(\beta^\top X_n)$, for some function b, where $\beta = (\beta_1, \ldots, \beta_D)^\top \in \mathbb{R}^D$ is a vector of regression parameters. This can be expressed as

$$f(y_n; x_n, \beta) = \exp\left\{\frac{y_n g(\mu_n) - b(\mu_n)}{a(\tau)} + c(y_n, \tau)\right\}$$

for some known functions $a(\cdot)$, $b(\cdot)$ such that $E(Y_n) = \frac{\partial}{\partial x} b(\beta^\top X_n)$, where τ is a dispersion parameter. The variance of the response is then related to the mean through the variance function $V(\cdot)$ such that $\text{Var}(Y_n) = V(\mu_n)a(\tau)$. Note that an intercept term β_0 can easily be added, by setting $X_{n0} = 1$, $n = 1, \ldots, N$. Denote the N i.i.d. realisations of (X, Y) as $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ respectively. The likelihood of the responses $\boldsymbol{y}$ given the observed covariates $\boldsymbol{x}$ and parameter vector β is given as

$$L(\boldsymbol{x}, \boldsymbol{y}; \beta) = \prod_{n=1}^{N} f(y_n; x_n, \beta). \tag{6.1}$$

The log-likelihood is therefore given as

$$\log L(\boldsymbol{x}, \boldsymbol{y}; \beta) = \sum_{n=1}^{N} \log f(y_n; x_n, \beta) = \sum_{n=1}^{N} \frac{y_n g(\mu_n) - b(\mu_n)}{a(\tau)} + c(y_n, \tau). \tag{6.2}$$

Classical maximum likelihood estimates (MLE's) for β (denoted as $\hat{\beta}_c$) based on the observed sample $(\boldsymbol{x}, \boldsymbol{y})$, $n = 1, \ldots, N$ can be obtained via the maximisation of (6.2) or alternatively, by finding the solution with respect to β to the following set of estimating functions.

$$U(\boldsymbol{x}, \boldsymbol{y}, \beta) = \frac{1}{N} \sum_{n=1}^{N} \nabla_\beta \log f(y_n; x_n, \beta) = \sum_{n=1}^{N} x_n \frac{y_n - \mu_n}{V(\mu_n) g'(\mu_n)} = 0. \tag{6.3}$$

Fahrmeir and Kaufmann (1985) showed that under mild regularity conditions, $\hat{\beta}_c$ is is asymptotically normally distributed, i.e.

$$\hat{\beta}_c \sim N(\beta, I(\beta_c)^{-1}), \tag{6.4}$$

with covariance matrix given by the inverse of the Fisher information matrix $I(\beta)$, with the (i, j) element denoted as

$$I(\beta)_{ij} = \mathbb{E}\left\{ \left(\nabla_\beta \log f(y; x, \beta)\right) \left(\nabla_\beta \log f(y; x, \beta)\right)^\top \right\}. \tag{6.5}$$

The fisher information matrix can be estimated via its empirical estimator $\hat{I}(\theta)$, i.e.

$$\hat{I}(\beta) = \frac{1}{N} \sum_{n=1}^{N} \left(\nabla_\beta \log f(y_n; x_n, \beta)\right) \left(\nabla_\beta \log f(y_n; x_n, \beta)\right)^\top.$$

Typically no analytical solution for the maximisation of (6.2) exists, meaning numerical methods are required. Various iterative procedures are readily available that allow the evaluation of the MLE for glm's, such as the Newton-Raphson method (Jennrich and Sampson, 1976), or iteratively re-weighted least squares based methods (Nelder and Wedderburn, 1972).

6.3.2 Existing methodologies for glm's for rounded discrete data

Lipsitz et al. (2004) proposed the following methodology for fitting glm models to data where some covariates are rounded to a known degree. Assume the covariate $x_{nD} \in \{1, \ldots, K\}$ is discrete and possibly coarsened, $n = 1, \ldots, N$, meaning that instead of observing the exact value of x_{nD}, we observe a (possibly) coarse value z_n such that we only know from z_n that x_{nD} can take some subset s_n of values in $\{1, \ldots, K\}$. As an example, suppose x_{nD} represents a count, and that each count has been rounded to the nearest 5. Then observing $z_n = 5M$ for some M tells us that $x_{nD} \in s_n = \{5M-2, 5M-1, 5M, 5M+1, 5M+2\}$. The likelihood of the regression coefficients is then given as

$$L(\beta, \theta) = \prod_{n=1}^{N} \sum_{z \in s_n} f(y_n; x_n^{-D}, z_n, \beta) h(z_n; x_n^{-D}, \theta),$$

where

$$f(y_n; x_n^{-D}, z_n, \beta) = f(y_n; (x_{n1}, \ldots, x_{nD-1}, z_n), \beta)$$

is the classic glm exponential density with z_n substituted in for x_{nD}, $h(z_n; x_n^{-D}, \theta)$ is an assumed distribution of the coarsened covariate given the observed covariates and a parameter θ, and x_n^{-D} represents the covariate vector x_n minus the D^{th} element. Continuing the above example where x_{nD} represents a count and has been coarsened, we could propose a poisson distribution for $x_{nD} = z_n$ given the observed covariates x_n^{-D}, such that

$$h(z; x_n^{-D}, \theta) = \frac{\lambda_n^z \exp -\lambda_n}{z!},$$

where $\log(\lambda_n) = \theta_0 + \theta_1 x_{n1} + \ldots + \theta_{D-1} x_{nD-1}$.

The above methodology relies on the condition that there are additional fully observed covariates included in the model from which the practitioner can use to fit a distribution for the underlying classical data. This approach is unsuitable for the dataset described in Section 6.5, as the fully observed variables (e.g. country, REF rating, etc.) only take a small number of distinct categorical values, meaning it isn't clear how they could be used to determine a distribution for the underlying microdata, given the fully observed covariates and rounded observations. Furthermore, the above methodology requires the specification of a parametric distribution for the coarsened values, given the observed covariates, leading to the requirement that the likelihood is optimised over additional nuisance parameters θ and the potential for a poor model fit if the distributional assumption is violated.

A similar methodology is employed by Johnson and Wiest (2014), who instead employ a simulation-based Bayesian approach and assume that not every observation is subject to the coarsening mechanism, meaning a distribution can be fit to the observations exhibiting coarsening using the fully observed observations. They define the probability model for a coarsened covariate z_n in terms of a posterior distribution $\pi(\beta, \boldsymbol{z}; \boldsymbol{y})$ for the glm parameter β and the

missing values of the covariate $z_n = x_{nD}$, $n = 1, \ldots, N$ as follows.

$$\pi(\beta, \boldsymbol{z}; \boldsymbol{y}) \propto m(\beta) \prod_{n=1}^{N} f(y_n; x_n^{-D}, z_n, \beta) h(z_n; x_n^{-D}) \mathbb{1}\{z_n \in s_n\},$$

where $m(\beta)$ is the assumed prior distribution for the regression parameter β, and $\mathbb{1}\{z_n \in s_n\}$ is the indicator function taking the value 1 if the simulated value z_n belongs to the set of possible underlying values for x_{nD}, and 0 otherwise. In this setting, $h(z_n; x_n^{-D})\mathbb{1}\{z_n \in s_n\}$ can be thought of as a prior distribution for the coarsened value. Inferences for β are then obtained from the set of realisations from simulations in a typical Bayesian framework. This approach is unsuitable for the applied analyses described in Section 6.5, as every observation is subject to rounding for the covariates of interest. Furthermore, the practitioner may wish to avoid a Bayesian or simulation based approach, meaning this methodology might not be appropriate. In the next Section we propose a methodology in which additional rounded variables not used in the glm analysis can be utilised in the glm framework for data of this nature.

6.4 Estimating glm parameters from discrete rounded data

We now propose a general framework from which estimates for the parameters of a glm model can be obtained from data where either the response or some covariate data is discrete and subject to rounding for every observation, and the additional covariates included in the model are not informative enough to use in the fittting of a distribution to the underlying microdata for the rounded variables. Our aim is to approximate the log-likelihood of the classical glm model by averaging the contribution of each observation over the log-likelihoods of all underlying values the rounded value could have arisen from, such that parameter estimates obtained from the 'approximate' log-likelihood are comparable to that of the complete analysis. We conclude this section with a description of the Godambe Information matrix (Godambe, 1960) that is used to obtain variances and subsequent inferences for the MLEs obtained from this model, given the misspecified aspect of the log-likelihood.

6.4.1 Overview

Suppose some of the covariates and/or the response variable are integer valued variables, and also only observed in rounded form. Denote $R(A)$ as the function with domain $\mathbb{Z}^+$, where A is an integer random variable and $R(A) = k$ if A is rounded to the nearest k^{th} integer. Note that for an unrounded integer variable X_d, $R(X_d) = 1$. Furthermore, for ease of notation, for a fully observed non-integer variable X_d, let $R(X_d) = 1$. Denote y_n^* and x_n^* as the observed data, with $y_n^* = y_n$ iff $R(Y) = 1$ and $x_{nd}^* = x_{nd}$ iff $R(X_d) = 1$, $n = 1, \ldots, N$, $d = 1, \ldots, D$. Denote

$$\phi(a) = \mathbb{1}(P(A = a|\cdot) > 0)$$

as the identity function taking the value 1 if the value a is possible for the random variable A, given additional variables and/or information, and 0 otherwise, and let

$$\gamma(a^*) = \{a^* - \lceil \frac{R(A)}{2} - 1 \rceil, \ldots, a^* + \lfloor \frac{R(A)}{2} \rfloor\}$$

represent the set of all possible underlying values the rounded observation a^* could have arisen from, given the degree of rounding and no additional constraints, where $\lceil x \rceil$ and $\lfloor x \rfloor$ represent the integer ceiling and floor functions respectively. Note that for unrounded observations, $\gamma(a^*) = \{a^*\}$. Let $\boldsymbol{x}^* = (x_1^*, \ldots, x_N^*)$ and $\boldsymbol{y}^* = (y_1^*, \ldots, y_N^*)$ represent the complete rounded dataset. Uniformly integrating the classical glm log-likelihood (6.2) over all possible underlying datasets yields the following log-likelihood:

$$\log L(\boldsymbol{x}*, \boldsymbol{y}*; \beta) = \sum_{n=1}^{N} \sum_{y_n \in \gamma(y_n^*)} \sum_{x_{n1} \in \gamma(x_{n1}^*)} \cdots \sum_{x_{nD} \in \gamma(x_{nD}^*)} \frac{\Phi(x_n, y_n)}{A_n} \log f(x_n, y_n; \beta), \tag{6.6}$$

where $\Phi(x_n, y_n) = \phi(y_n) \times \prod_{d=1}^{D} \phi(x_{nd})$ and

$$A_n = \sum_{y_n \in \gamma(y_n^*)} \sum_{x_{n1} \in \gamma(x_{n1}^*)} \cdots \sum_{x_{nD} \in \gamma(x_{nD}^*)} \Phi(x_n, y_n)$$

is the total number of distinct multivariate values the n^{th} underlying observation could have taken. Note that $\Phi(x_n, y_n)$ is the indicator function that takes the value 1 if the potential underlying observation (x_n, y_n) being plugged into (6.6) is possible (i.e. a non-zero probability of occuring given the observed rounded data), and 0 otherwise, and that each summation only has one term (the observation) if the corresponding covariate is unrounded. This leads to the assumption that the possible underlying multidimensional values are uniformly distributed, given the rounded observation, however the marginal distributions of the univariate values are not necessarily uniform. Note therefore that $\Phi(x_n, y_n)$ can possibly be better defined through the use of additional arguments, as seen in Section 6.4.2. Estimates for β can then be obtained via the solution to the following set of estimating functions.

$$U'(\boldsymbol{x}^*, \boldsymbol{y}^*, \beta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{y_n \in \gamma(y_n^*)} \sum_{x_{n1} \in \gamma(x_{n1}^*)} \cdots \sum_{x_{nD} \in \gamma(x_{nD}^*)} \frac{\Phi(x_n, y_n)}{A_n} \nabla_\beta \log f(y_n; x_n, \beta) = 0, \tag{6.7}$$

where $\nabla_\beta \log f(y_n; x_n, \beta) = x_n \frac{y_n - \mu_n}{V(\mu_n) g'(\mu_n)}$.

Each term within the summation over n in (6.7) can be described as the average of the contributions to the classical glm estimating functions of all possible multidimensional observations which could have led to the observed value. Note that when there are constraints on the underlying data, such as when covariates are functions of each other (i.e. sums), some covariates end up having differently weighted underlying marginal values, due to the exclusion of some 'impossible' underlying observations.

Example. Suppose X_d is a count variable rounded to the nearest 5, i.e. $R(X_d) = 5$. If $x_{nd}^* = 5$,

then $x_{nd} \in \gamma(5) = \{3,4,5,6,7\}$ and $\phi(k|\cdot) = 1,\ k = 3,4,5,6,7$. However, if $x^*_{nd} = 0$, then $x_{nd} \in \gamma(0) = \{0,1,2\}$ and $\phi(k|\cdot) = 0$ for $k = -2,-1$. Clearly the potential underlying datapoints below the rounded observation (i.e. $x^*_{nd} - 2$ and $x^*_{nd} - 1$) are possible for $x^*_{nd} \geq 5$, but not for $x^*_{nd} = 0$. As a result, $\phi(x_{nd}|\cdot) = \mathbb{1}(x_{nd} > 0)$.

6.4.2 Utilising additional information

Suppose $z_n \in \mathbb{R}^+$, is an integer realisation of a random variable Z comprising of additional information for the n^{th} observation, such that z_n can be included in the function ϕ for some rounded variables to better specify the probability of the underlying classical observation, given the rounded realisations. Suppose also that z_n can be potentially rounded to the nearest k^{th} integer, i.e. $R(Z) = k$. This additional information can be incorporated into the estimating equations (6.7) as follows.

For ease of notation, first denote $\sum_{\Gamma(x^*_n, y^*_n, \ldots)}$ as the summation over all included and additional information variables for the given model, such that

$$\sum_{\Gamma(x^*_n, y^*_n, z^*_n)} = \sum_{z_n \in \gamma(z^*_n)} \sum_{y_n \in \gamma(y^*_n)} \sum_{x_{n1} \in \gamma(x^*_{n1})} \cdots \sum_{x_{nD} \in \gamma(x^*_{nD})}.$$

The 'averaged' glm estimating function for a rounded sample with additional information incorporated in $\mathbf{z}^*$ is then given as

$$U'(\boldsymbol{x}^*, \boldsymbol{y}^*, \beta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{\Gamma(x^*_n, y^*_n | z^*_n)} \frac{\Phi(x_n, y_n, z_n)}{A_n} \nabla_\beta \log f(y_n; x_n, \beta) = 0. \tag{6.8}$$

where $\Phi(x_n, y_n, z_n) = \phi(z_n|x_n, y_n)\phi(y_n|x_n, z_n) \times \prod_{d=1}^{D} \phi(x_{nd}|x_n, y_n, z_n)$ and

$$A_n = \sum_{\Gamma(x^*_n, y^*_n, z^*_n)} \Phi(x_n, y_n, z_n).$$

Once again this leads to the assumption that the possible underlying multidimensional values for each rounded observation are uniformly distributed, albeit without necessarily leading to uniform distributions for each marginal underlying value.

Example. Suppose the sum of some subset $\boldsymbol{d} = (d_1, \ldots, d_P)$, $d_p \in 1, \ldots, D$, $p = 1, \ldots, P$, $P \leq D$ of the included covariates might be known to be z_n, such that $\sum_{d \in \boldsymbol{d}} x_{nd} = z_n$. In the context of the analysis conducted in this paper, this is comparable to the restrictions that the sums of the males and females for each university cost centre must be equal to the total number of people, i.e. $\sum_{i=1}^{4} w_{ni} + v_{ni} = n_n$. As a result,

$$\begin{aligned} \Phi(x_n, y_n, z_n) &= \phi(z_n|x_n, y_n)\phi(y_n|x_n, z_n) \times \prod_{d=1}^{D} \phi(x_{nd}|x_n, y_n, z_n) \\ &= \mathbb{1}\Big(\sum_{d \in \boldsymbol{d}} x_{nd} = z_n\Big) \times \phi(y_n) \times \prod_{d=1}^{D} \phi(x_{nd}). \end{aligned}$$

Only values for x_n for which the sum of the covariates indexed by $\boldsymbol{d}$ is equal to the additional observation z_n are then included in the estimating equation (6.8).

6.4.3 Variances

The asymptotic distribution of the classical MLE $\hat{\theta}_c$ shown in (6.4), and subsequent covariance matrix, denoted as the Fisher information matrix and shown in (6.5), are specific cases of a general theory outlined by Godambe (1960), which states that the MLE is asymptotically distributed with covariance matrix given as the inverse of the Godambe Information matrix

$$G(\beta) = H(\beta)J(\beta)^{-1}H(\beta), \tag{6.9}$$

such that

$$\hat{\beta}_c \sim N(\beta, G(\beta)^{-1}),$$

where ∇_θ and ∇_{θ^2} represent respectively the gradient and matrix of second derivatives for a given parameter θ, and $H(\beta) = -\mathbb{E}(\nabla^2 \log L(\boldsymbol{x}, \boldsymbol{y}; \beta))$ and $J(\beta) = \text{Var}(\nabla \log L(\boldsymbol{x}, \boldsymbol{y}; \beta))$ are respectively the sensitivity and variability matrices. For regular likelihoods, $H(\beta) = J(\beta)$ and as a result, $G(\beta) = H(\beta) = J(\beta)$, however, when the model is misspecifed, or if there are correlations between the observations, then $H(\beta) \neq J(\beta)$ (White, 1982) and the variance must be obtained by the complete Godambe information matrix.

Given we have a misspecified log-likelihood, and the Godambe information matrix shown in (6.9) must be used to obtain variances for the estimates, denoted as $\hat{\theta}_s$. We now present formulae for the estimators of $H(\beta)$ and $J(\beta)$, respectively denoted as $\hat{H}(\hat{\beta}))$ and $\hat{J}(\hat{\beta}))$. The hessian and jacobian estimators are then respectively given as

$$\hat{H}(\hat{\beta}) = \sum_{n=1}^{N} \nabla^2_\beta \log L(x^*_n, y^*_n; \hat{\beta}) = \sum_{n=1}^{N} \sum_{\Gamma(x^*_n, y^*_n, z^*_n)} \frac{\Phi(x_n, y_n, z_n)}{A_n} \nabla^2_\beta \log f(x_n, y_n; \hat{\beta}) \tag{6.10}$$

$$\begin{aligned}\hat{J}(\hat{\beta}) &= \sum_{n=1}^{N} \left(\nabla_\beta \log L(x^*_n, y^*_n; \beta)\right) \left(\nabla_\beta \log L(x^*_n, y^*_n; \hat{\beta})\right)^\top \\ &= \sum_{n=1}^{N} \left(\sum_{\Gamma(x^*_n, y^*_n, z^*_n)} \frac{\Phi(x_n, y_n, z_n)}{A_n} \nabla_\beta \log f(x_n, y_n; \hat{\beta}) \right) \left(\sum_{\Gamma(x^*_n, y^*_n, z^*_n)} \frac{\Phi(x_n, y_n, z_n)}{A_n} \nabla_\beta \log f(x_n, y_n; \hat{\beta}) \right)^\top .\end{aligned} \tag{6.11}$$

Note that $\nabla_\beta \log f(x_n, y_n; \hat{\beta})$ and $\nabla^2_\beta \log f(x_n, y_n; \hat{\beta})$ are simply the classical gradient and hessian formulae. For fully observed data, $\hat{H}(\hat{\beta}) = \hat{J}(\hat{\beta})$, and $\hat{G}(\theta) = \hat{I}(\theta)$

6.5 Data Analyses

We now undertake the real data analysis of the Athena SWAN dataset to determine which variables are predictive of the ordinal-valued Athena Award Status and the number/proportion of females in a given academic cost centre. As previously stated, the rounded nature of some of

the observed count variables result in the unsuitability of classical glm models, leading to the necessity of the glm framework proposed in this paper. We first outline the desired analyses for each model, and then illustrate their utility with some simulation studies. The actual analyses are then performed to conclude this section.

6.5.1 Data Analyses Description

In each of the real data analyses, the non-standard EE methodology (6.8) is compared to a naive classical analysis of the rounded data, with MLE variances for both methods obtained using the Godambe information matrix. For the classical rounded analysis, we are forced to set $p_n = \frac{\sum_{i=1}^{4} w_{ni}^*}{n_n^*}$ and $p_{ni} = \frac{w_{ni}^*}{w_{ni}^* + v_{ni}^*}$, $i = 1, \ldots, 4$, for observations with respective counts less than $\tau = 22.5$. Furthermore, denote w_n and v_n as the total number of females and males respectively for each observation. For the rounded classical analysis, we are therefore also forced to set $w_n = \sum_{i=1}^{4} w_{ni}^*$ and $v_n = \sum_{i=1}^{4} v_{ni}^*$. Let $\epsilon = 10^{-3}$. For (6.8), when p_n is included in the estimating equations, we set $p_n = \frac{w_n}{n_n}$.

For the non-standard EE methodology (6.8), there are various constraints on the underlying data for the count variables, which we can exploit for better parameter estimates. For example, every underlying count observation can't be negative, and the counts of the males and females for each employment level must add to the relevant total count for that level. For observations where some proportions of females for each level are available, further conditions can be imposed. Utilising all the available information yields the following identity functions

$$
\begin{aligned}
\phi(n_n) &= \mathbb{1}(n_n > 0) \\
\phi(w_n, v_n) &= \mathbb{1}(w_n \geq 0 \cap v_n \geq 0) \\
\phi(w_{ni}, v_{ni}) &= \mathbb{1}(w_{ni} \geq 0, v_{ni} \geq 0) \\
\phi(w_n | \{w_{ni}\}) &= \mathbb{1}(\sum_{i=1}^{4} w_{ni} = w_n), \\
\phi(v_n | \{v_{ni}\}) &= \mathbb{1}(\sum_{i=1}^{4} v_{ni} = v_n), \\
\phi(n_n | w_n, v_n) &= \mathbb{1}((w_n + v_n) = n_n), \\
\phi(p_n | n_n, w_n, v_n) &= \mathbb{1}((n_n < 22.5) \cup (|\frac{w_n}{n_n} - p_n)| < \epsilon), \\
\phi(p_{ni} | \{w_{ni}, v_{ni}\}) &= \mathbb{1}((w_{ni} + v_{ni} < 22.5) \cup (|\frac{w_{ni}}{w_{ni} + v_{ni}} - p_{ni})| < \epsilon),
\end{aligned}
$$

such that

$$
\begin{aligned}
\Phi(w_n, v_n, n_n | \{w_{ni}, v_{ni}, p_{ni}\}, p_n) =& \phi(n_n)\phi(w_n, v_n)\phi(w_{ni}, v_{ni})\phi(w_n | \{w_{ni}\})\phi(v_n | \{v_{ni}\}) \\
& \times \phi(n_n | w_n, v_n)\phi(p_n | n_n, w_n, v_n)\phi(p_{ni} | \{w_{ni}, v_{ni}\}) \qquad (6.12)
\end{aligned}
$$

and

$$A_n = \sum_{n_n \in \gamma(n_n^*)} \sum_{w_{n1} \in \gamma(w_{n1}^*)} \cdots \sum_{w_{n4} \in \gamma(w_{n4}^*)} \sum_{v_{n1} \in \gamma(v_{n1}^*)} \cdots \sum_{v_{n4} \in \gamma(v_{n4}^*)} \Phi(w_n, v_n, n_n | \{w_{ni}, v_{ni}, p_{ni}\}, p_n).$$

These expressions are then included in the estimating functions (6.8), with variances obtained using the components of the Godambe information matrix shown in (6.10) and (6.11).

For each analysis, we define the response observation and covariate vector for the n^{th} observation as y_n and x_n respectively. Expressions for the log-likelihood, estimating functions and components of the estimated Godambe information matrix (6.10) and (6.11) are given for each model, so that estimates and variances can be obtained for each parameter vector.

Poisson Regression Analysis

For predicting the number of females w_n for each observation (university cost centre) using the other covariates, a poisson regression is appropriate. Denote the response variable $y_n = w_n$, the predictor variables $x_n = (1, a_n, t_n, r_n, l_{n1}, l_{n2}, l_{n3})^\top$ and the vector of regression coefficients as $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_{41}, \beta_{42}, \beta_{43})^\top$. The classical glm likelihood is given as

$$L(\boldsymbol{x}, \boldsymbol{y}; \beta) = \prod_{n=1}^{N} P(Y_n = y_n | x_n, \beta),$$

where

$$P(Y_n = y | x_n, \beta) = \frac{1}{y!} \exp\left(y(\beta^\top x_n) - \exp(\beta^\top x_n)\right).$$

The classical poisson log-likelihood can therefore be expressed as

$$\log L(\boldsymbol{x}, \boldsymbol{y}; \beta) = \sum_{n=1}^{N} \{y_n(\beta^\top x_n) - \exp(\beta^\top x_n) - \log(y_n!)\},$$

with the gradient for the variance calculation and estimating function (6.8) given as

$$\nabla_\beta \log L(\boldsymbol{x}, \boldsymbol{y}; \beta) = \sum_{n=1}^{N} x_n(y_n - \exp(\beta^\top x_n)).$$

Binomial Regression Analysis

For predicting the proportion of females p_n for each university cost centre using the other covariates, a binomial regression model can be used, with the counts of females w_n and total size n_n included in the response matrix. Denote the predictor observations as $x_n = (1, n_n, a_n, t_n, r_n, l_{n1}, l_{n2}, l_{n3})^\top$ and the vector of regression coefficients as $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_{51}, \beta_{52}, \beta_{53})^\top$. The classical glm likelihood can be expressed as

$$L(\boldsymbol{x}, \boldsymbol{y}; \beta) = \prod_{n=1}^{N} \frac{\exp(w_n \beta^\top x_n)}{(1 + \exp(\beta^\top x_n))^{n_n}}.$$

The classical poisson log-likelihood can therefore be expressed as

$$\log L(\boldsymbol{x}, \boldsymbol{y}; \beta) = \sum_{n=1}^{N} \{w_n \beta^\top x_n - n_n \log(1 + \exp(\beta^\top x_n)\},$$

with the gradient given as

$$\nabla_\beta \log L(\boldsymbol{x}, \boldsymbol{y}; \beta) = \sum_{n=1}^{N} x_n \left(w_n - \frac{n_n \exp(\beta^\top x_n)}{1 + \exp(\beta^\top x_n)} \right).$$

Ordinal Logistic Regression Analysis

An ordinal logistic regression model can be used to predict the Athena SWAN award status a_n for each university cost centre using the other covariates. Denote the response as $y_n = a_n$, the predictor observations as $x_n = (n_n, p_n, t_n, r_n, l_{n1}, l_{n2}, l_{n3})^\top$, the vector of regression coefficients as $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_{51}, \beta_{52}, \beta_{53})^\top$ and an ordered vector of threshold parameters as $\alpha = (\alpha_1, \alpha_2)^\top$, $\alpha_2 > \alpha_1$. The classical ordinal logistic log-likelihood we want to approximate can be expressed as follows.

$$L(\boldsymbol{x}, \boldsymbol{y}; \beta, \alpha) = \prod_{n=1}^{N} P(Y_n = y_n | x_n, \beta, \alpha),$$

where

$$P(Y_n = k | x_n, \beta, \alpha) = P(Y_n \leq k | x_n, \beta, \alpha) - P(Y_n \leq k - 1 | x_n, \beta, \alpha), \tag{6.13}$$

$$P(Y_n \leq k | x_n, \beta, \alpha) = \frac{\exp(\alpha_k + \beta^\top x_n)}{1 + \exp(\alpha_k + \beta^\top x_n)}. \tag{6.14}$$

Note that this model has no intercept term. The classical ordinal logistic regression log-likelihood can therefore be expressed as

$$\log L(\boldsymbol{x}, \boldsymbol{y}; \beta, \alpha) = \sum_{n=1}^{N} \log P(Y_n = y_n | x_n, \beta, \alpha).$$

The gradient of the log likelihood, which is dependent on the constraints of the threshold parameter α, is given by Kim (2004).

$$\nabla_\beta \log L(\boldsymbol{x}, \boldsymbol{y}; \beta, \alpha) = \sum_{n=1}^{N} \sum_{k=1}^{K} x_n \mathbb{1}(y_n = k) \left(1 - P(Y_n \leq k | x_n, \beta, \mu) - P(Y_n \leq k - 1 | x_n, \beta, \alpha)\right)$$

$$\nabla_\alpha \log L(\boldsymbol{x}, \boldsymbol{y}; \beta, \alpha) = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{1}(y_n = k) \Big\{ e_{y_n} \left(1 - P(Y_n \leq k | x_n, \beta, \alpha) - \frac{1}{1 - \exp(\alpha_{k-1} - \alpha_k)}\right) + e_{y_n - 1} \left(1 - P(Y_n \leq k - 1 | x_n, \beta, \alpha) - \frac{1}{1 - \exp(\alpha_k - \alpha_{k-1})}\right) \Big\},$$

where e_k is the k^{th} canonical vector, $k = 1, \ldots, K$.

As noted by McCullagh (1980), when estimating the parameters of the ordinal logistic regres-

sion models, the actual quantities being estimated for the regression and threshold parameters are respectively $\frac{\alpha}{\epsilon}$ and $\frac{\beta}{\epsilon}$, where ϵ is a scale parameter. As a result, there is an identifiability issue with the model, in that there are infinitely many sets of parameters (β, α) which will achieve the same log-likelihood score, given adjustments to the scale parameter ϵ. One solution is to fix α_1, ensuring there is a unique solution. This approach is used in the simulations and real data analyses seen later in this paper.

6.5.2 Simulations

We consider some synthetic examples for the poisson, binomial and logistic ordinal regression analyses to illustrate the advantages of using the glm estimating function methodology outlined in Section 6.4 over a naive classical analysis of the observed rounded data. For each of the following analyses, N observations are simulated, with the covariates in each observation designed to mimic the covariates in the motivating dataset described in Section 6.2. The numbers of females w_n and males v_n for each cost centre are simulated directly, rather than from the summation of simulated values for w_{ni} and v_{ni} respectively. For each simulated dataset, the count observations are rounded to the nearest $R = 25, 20, 15, 10, 5, 1$ integers, with $R = 1$ representing the complete classical analysis of the underlying dataset. For each setup, realisations of t_n, r_n and l_n, $n = 1, \ldots, N$, are simulated as

$$
\begin{aligned}
t_n &\in \{1, \ldots, 20\}, \text{ where } P(t_n = k) \propto \frac{1}{k}, k = 1, \ldots, 20 \\
r_n &\in \{1, 2, 3\}, \text{ where } P(X_{n3} = k) = \frac{1}{3}, k = 1, 2, 3 \\
l_n &\in \{1, 2, 3, 4\}, \text{ where } P(X_{n4} = k) = \frac{1}{4}, k = 1, 2, 3, 4.
\end{aligned}
$$

MLE's for β (and μ for the ordinal model) for each analysis are obtained for each degree of rounding (R) using a naive classical analysis of the rounded dataset, with variances obtained using the estimated Godambe information matrix. MLE's are then obtained using (6.8) for each degree of rounding, with varying degrees of additional information incorporated. The following functions Φ are used to incorporate varying amounts of additional information for the estimating functions in (6.8).

$$\Phi(x_n, y_n) = \phi(n_n)\phi(w_n, v_n) \tag{6.15}$$

$$\Phi(x_n, y_n) = \phi(n_n)\phi(w_n, v_n)\phi(n_n|w_n, v_n) \tag{6.16}$$

$$\Phi(x_n, y_n|p_n, \tau) = \phi(n_n)\phi(w_n, v_n)\phi(n_n|w_n, v_n)\phi(p_n|n_n, w_n, v_n). \tag{6.17}$$

Note that these functions Φ represent the incoporation of an increasing amount of additional information, ranging from only the conditions that each count is non-negative, and the total size is also not zero (6.15), to the inclusion of the condition that $w_n + v_n = n_n$ (6.16), as well as the proportions of females conditions (6.17). Each analysis is then replicated 1000 times, so that the average estimates and average estimated Godambe variances can be presented. The Godambe variances are also compared to the observed variance of the MLEs across the 1000 replications,

to establish whether good estimates for the variances are obtained.

For the binomial and ordinal logistic regression simulation studies, the degree of rounding for all observations is then fixed at $R = 5$. The mean of the process used to simulate n_n is then varied inversely proportionally to the regression coefficient associated with that covariate, such that the effect on the other regressors should be minimal. Varying the total size n_n with a fixed degree of rounding allows us demonstrate the point at which the cost centre sizes are so big that the differences between the approaches are negligible.

Poisson Regression simulations

For the poisson regression synthetic example, $N = 100$ observations are simulated, with realisations for $a_n \in \{0, 1, 2\}$, $n = 1, \ldots, N$, simulated by sampling from $\{0, 1, 2\}$, with $P(a_n = k) = \frac{1}{3}, k = 0, 1, 2..$ A corresponding poisson response w_n is then simulated from each x_n and $\beta_{true} = (0.8, 0.2, 0.2, 0.2, 0.3, 0.5, 0.4)$. For each observation, the number of males v_n is then simulated from an poisson distribution with mean $\frac{3\beta^\top x_n}{2}$, and the total cost centre size n_n then follows as $v_n + y_n = z_n$. Univariate histograms of n_n, w_n and v_n, $n = 1, \ldots, N$ are shown in Figure 6.2. Figure 6.3 presents the mean estimates (top row) and standard deviations (middle row) for a subset of the parameters for varying degrees of rounding and additional information incorporation. Results for all parameters are shown in Figures C.1 and C.2 in the Supplementary material in C.1. The bottom row of figure 6.3 presents the observed standard deviation across the 1000 simulation replicates along with the mean estimated Godambe standard deviation for the complete model and the model represented by (6.15).

We see from Figure 6.3 that as we decrease the degree of rounding (i.e. $R \to 1$), the results of each model converge towards the classical results due to the increasing retention in information about the underlying data. The newly developed EE approach (6.8) gives closer estimates and variances to the complete analysis compared to the naive classical analysis of the rounded data (which performs the worst by far) when no additional information is included (6.15), and increasingly outperforms the rounded analysis when additional information is included (6.16) and (6.17). Additionally, the standard deviation plots on the bottom row in Figure 6.3 demonstrate that the Godambe information matrix is able to effectively estimate the standard error of each parameter from the estimating function (6.8), as the mean estimated standard errors are comparable to the observed sample standard deviations, albeit with a slight underestimation of the variance of each parameter.

Binomial Regression simulations

For the binomial regression synthetic example, $N = 100$ observations are simulated, with realisations of n_n and a_n generated as

$$n_n \sim \exp(50) + 5 \text{ (rounded to the nearest integer)}$$
$$a_n \in \{0, 1, 2\}, \text{ where } P(a_n = k) = \frac{1}{3}, k = 0, 1, 2$$

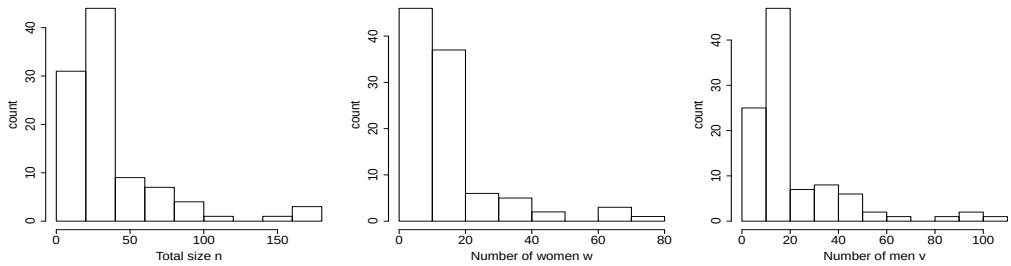

Figure 6.2: Histograms from one replication for the poisson synthetic analysis of the total size (left), number of females (middle) and number of males (right).

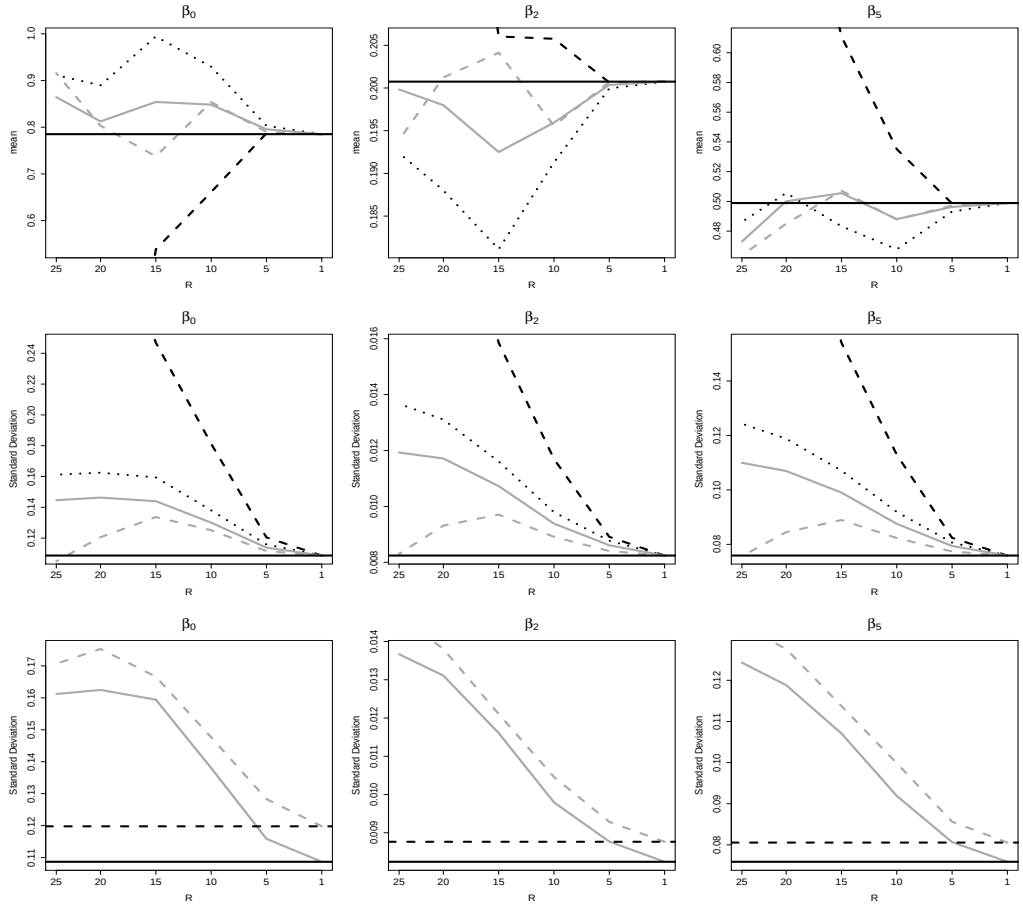

Figure 6.3: Mean estimates (top row), mean estimated standard deviations (middle row) and mean estimated and observed standard deviations (bottom row) for a subset of the parameters (columns) in the poisson regression synthetic analysis. For the top two rows, Classic results are shown in solid black, classical analysis of the rounded data in dashed black, EE results with no additional information in dotted black (6.15), EE results with additional totals information utilised in solid grey (6.16) and EE results with additional proportions information in dashed grey (6.17).. For the bottom row, observed standard deviations are shown with dashed lines, with mean estimated Godambe standard deviations shown in solid lines.

Using a pre-specified parameter vector $\beta_{true} = (-0.5, 0.001, 0.05, 0.05, -0.1, 0.1, 0.2, -0.3)^\top$, the number of females w_n is then simulated for each group, such that $w_n \sim \text{Bin}(n_n, p_n)$, where

$p_n = \beta^\top x_n$, and from these we have the numbers of males for each group $v_n = n_n - w_n$, as well as the true proportions of females $(\frac{w_n}{n_n})$ for groups with total sizes $n_n > \tau$, where $\tau = 22.5$. Univariate histograms of n_n, w_n and v_n, $n = 1, \ldots, N$ are shown in Figure 6.4. Figure 6.5 presents the mean estimates (top row) and standard deviation (middle row) for a subset of the parameters for varying degrees of rounding and additional information incorporation, with the complete results shown in Figures C.3 and C.4 in C.1. The bottom row of figure 6.5 presents the observed standard deviation of the estimates across the 1000 simulation replicates along with the mean estimated Godambe standard deviation for the complete model and (6.15). Figure 6.6 shows the effect of the change in mean of the total size n_n on the differences in the estimated variances between the complete classical model, and the various models investigated for a subset of the parameters, with the complete results shown in Figure C.5 in C.1.

Figure 6.5 demonstrates that as the degree of rounding decreases (i.e. $R \to 1$), the results of each model converge towards the results of the complete classical analysis due to the increasing retention in information about the underlying data. The newly developed EE approach gives closer results to the complete case (in terms of mean estimates and variances) compared to the classical binomial regression performed on the rounded data for the predictor variables (which provides by far the worst results), and when additional information is incorporated into the model, the estimating function results improve in that they are closer to the complete classical results for all parameters. Furthermore, the standard error plots on the bottom row in Figure 6.5 show that good estimates can be obtained for the standard error of each parameter via the estimated Godambe information matrix, as the mean estimated standard errors are comparable to the observed sample standard deviations, albeit slightly underestimated. Figure 6.6 illustrates that that when additional information is utilised, the estimating function methodology (6.8) perform better than the classical model performed on the rounded data for all predictor means, although this improvement gets comparably smaller as the magnitude of the department sizes, numbers of females and numbers of males increases.

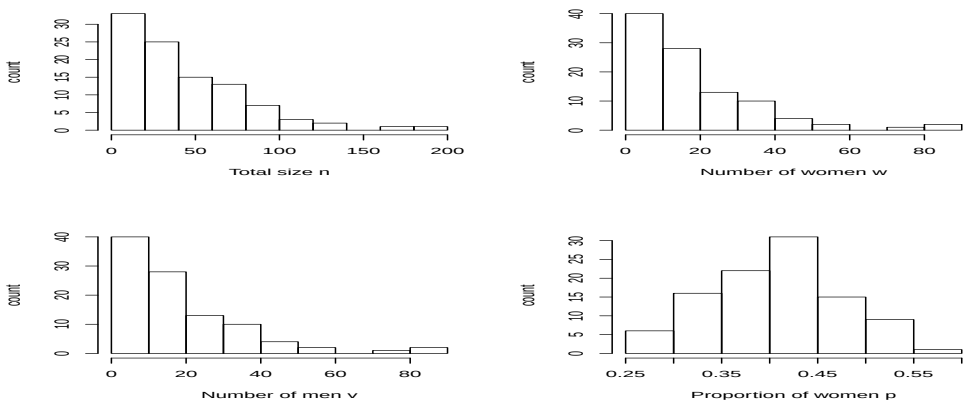

Figure 6.4: Histograms from one replication of the binomial synthetic example of the total size (top left), number of females (top right), number of males (bottom left) and proportion of females (bottom right.

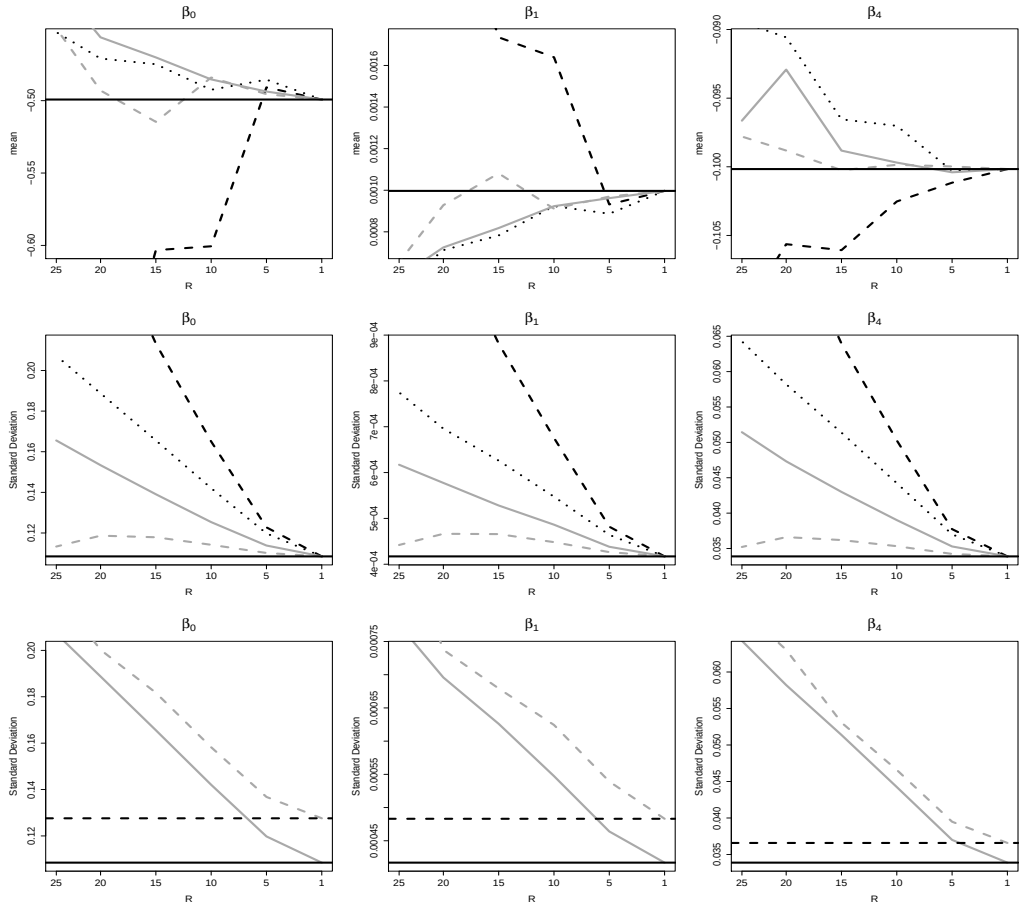

Figure 6.5: Mean estimates (top row), mean estimated standard deviations (middle row) and mean estimated and observed standard deviations (bottom row) for a subset of the parameters (columns) in the binomial regression synthetic analysis. For the top two rows, Classic results are shown in solid black, classical analysis of the rounded data in dashed black, EE results with no additional information in dotted black (6.15), EE results with additional totals information utilised in solid grey (6.16) and EE results with additional proportions information in dashed grey (6.17).. For the bottom row, observed standard deviations are shown with dashed lines, with mean estimated Godambe standard deviations shown in solid lines.

Ordinal Logistic Regression simulations

For the ordinal logistic regression synthetic example, $N = 100$ observations are simulated, with realisations of n_n and p_n simulated as

$$n_n \sim \exp(50) + 5 \text{ (rounded to the nearest integer)}$$
$$p_n \sim U(0.2, 0.75).$$

The number of females and males can then be obtained, such that $w_n = p_n n_n$ (rounded) and $v_n = n_n - w_n$. $K = 3$ classes are used for the ordinal categorical response variable Y, which we simulate as follows. Two equally spaced thresholds $\alpha_{true} = (\alpha_1, \alpha_2)^\top = (0, 3)^\top$ and a parameter vector $\beta_{true} = (0.01, -0.5, -0.01, -0.2, -0.5, -1, 0.5)^\top$ are chosen, and class probabilities are

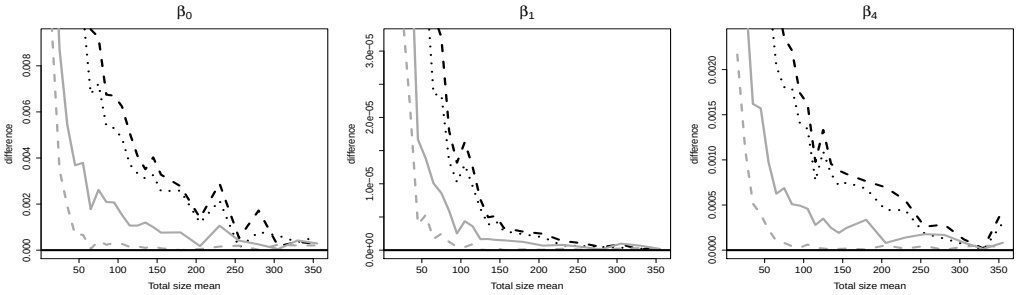

Figure 6.6: Difference between the estimated variances of the complete classical analysis and various models for a subset of the parameters for the binomial regression synthetic example. Classic results are shown in solid black, classical analysis of the rounded data in dashed black, EE results with no additional information in dotted black (6.15), EE results with additional totals information utilised in solid grey (6.16) and EE results with additional proportions information in dashed grey (6.17).

then calculated using (6.13) and (6.14). Each Y_n is then simulated according to the class probabilities. Univariate histograms of n_n, w_n and v_n, $n = 1, \ldots, N$ for one replication are shown in Figure 6.7. Due to the identifiability issues associated with the threshold parameter α, we fix $\alpha_1 = 0$ for each optimisation. Figure 6.8 presents the mean estimates (top row) and standard deviation (middle row) for a subset of the parameters for varying degrees of rounding and additional information incorporation, with the complete results shown in Figures C.6 and C.7 in C.1. The bottom row of figure 6.8 presents the observed standard deviation of the estimates across the 1000 simulation replicates along with the mean estimated Godambe standard deviation for the complete model and (6.15). Figure 6.9 shows the effect of the change in mean of the total size n_n on the differences in estimated variances between the complete classical model, and each model for a subset of the parameters, with the complete results shown in Figure 6.9 in C.1.

Figure 6.8 shows that as the degree of rounding decreases (i.e. $R \to 1$), the results for each model for the regression and threshold parameters (β and α) converge towards the classical results. The naive classical analysis of the rounded data outperforms (6.15) for a large degree of rounding for some parameters, however both analyses obtain comparable results for lower values of R. When additional information is incorporated, i.e. (6.16) and (6.17), closer results to the complete classical case are produced than a naive classical analysis of the rounded data, with more additional information leading to an improvement in results. Unlike the other glm models previously investigated, the variances of each parameter were lower than that of the classical case, with convergence to the classical results from below with decreasing degrees of rounding R. The standard error plots on the bottom row of Figure 6.8 show that the Godambe information matrix provides good estimates for the standard error of each parameter from the estimating functions 6.8, as the mean estimated standard errors are comparable to the observed sample standard deviations. Figure 6.9 illustrates that that the estimating functions (6.8) produce closer estimates to the complete case when additional information is utilised, and perform better than the naive classical analysis of the rounded data for all predictor means, although this

improvement gets comparably smaller as the magnitude of the department sizes, numbers of females and numbers of males increases.

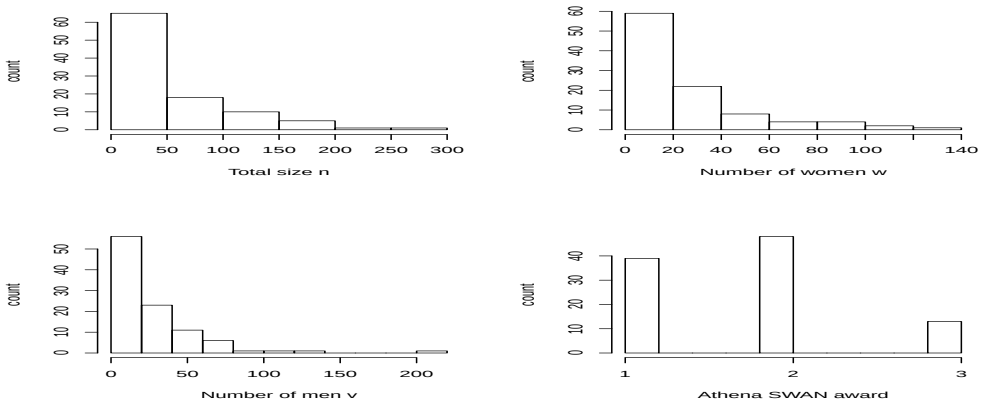

Figure 6.7: Histograms from one replication of the ordinal logistic synthetic example of the total size (top left), number of females (top right), number of males (bottom left) and the ordinal response variable (bottom right)

6.5.3 Analyses of the Athena SWAN dataset

We now estimate the glm parameters for the poisson, binomial and ordinal logistic regression models applied to the datasets described in Section 6.2 using both a naive classical analysis of the rounded datasets and the estimating function methodology outlined in Section 6.4. For the estimating function methodology, $\Phi(w_n, v_n, n_n|\{w_{ni}, v_{ni}, p_{ni}\}, p_n)$ (shown in (6.12)) was used to restrict the domain of the underlying dataset. For the ordinal logistic regression analyses, we set $\alpha_1 = 0$ and estimate α_2 to avoid the identifiability issues associated with completely unrestricted threshold parameters (McCullagh, 1980). Denote the estimates obtained via a naive classical analysis of the rounded data as $\hat{\beta}_r^{year}$, and the estimates obtained via the estimating function methodology shown in (6.8) by $\hat{\beta}_s^{year}$ for each year. Standard deviations for each parameter are estimated using the estimated Godambe information matrix obtained from (6.10) and (6.11). Tables 6.1, 6.2 and 6.3 respectively show the MLEs and estimated standard deviations for the parameters of the poisson, binomial and ordinal logistic regression analyses of the Psychology (top) and Physics (bottom) data. Important predictors (large estimates compared to small standard deviations) are indexed by *. When referencing traditional hypothesis testing in the following description of the results, we assume 5% as the significance level

The poisson regression models fit to Psychology datasets for the years 2012 and 2016 yield estimates and standard deviations that suggest that the most important predictor of the number of females w_n for each cost centre is the location l_n. Being located in England ($l_{n1} = 1$) has a significant positive effect on the number of females for the Psychology 2012 dataset, yielding a low p-value in a traditional hypothesis test. The estimating function methodology yields a higher estimate for β_{41} than the naive classical analysis of the rounded data, with a comparable standard deviation. For the Psychology 2016 dataset, the estimates for β_{43} have significantly

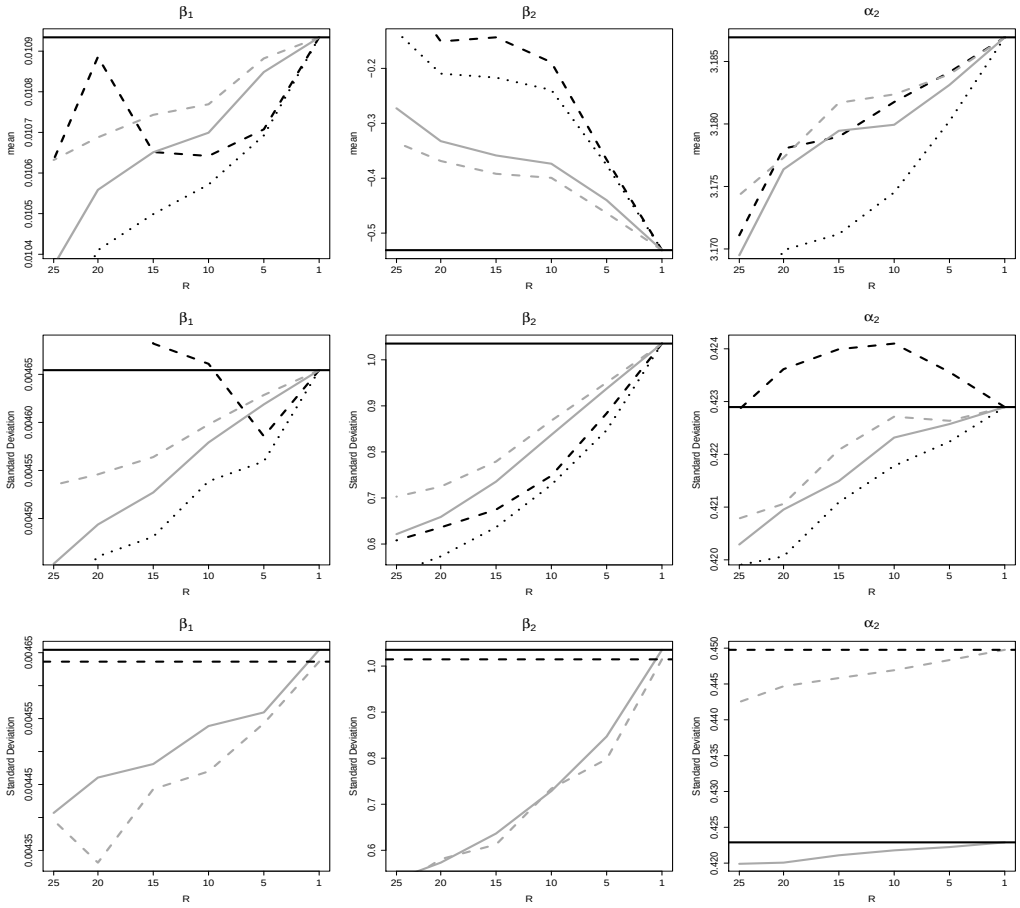

Figure 6.8: Mean estimates (top row), mean estimated standard deviations (middle row) and mean estimated and observed standard deviations (bottom row) for a subset of the parameters (columns) in the ordinal logistic regression synthetic analysis. For the top two rows, Classic results are shown in solid black, classical analysis of the rounded data in dashed black, EE results with no additional information in dotted black (6.15), EE results with additional totals information utilised in solid grey (6.16) and EE results with additional proportions information in dashed grey (6.17).. For the bottom row, observed standard deviations are shown with dashed lines, with mean estimated Godambe standard deviations shown in solid lines.

increased when compared to the 2012 dataset, such that the resultant hypothesis yields suggests that being located in Wales ($l_{n3} = 1$) is an important predictor of the number of females in the traditional hypothesis testing setting.

There are noticeable differences between the MLE's obtained from the poisson estimating function methodology (6.8) and the poisson naive classical analysis. Estimates for β_0 and β_1 obtained from the estimating function analysis (6.8) of the 2012 dataset are noticeably different than that of the naive classical analysis, however in 2016 both methods obtained similar results for these coefficients. Consequently, for the 2012 analysis, a naive classical analysis of the rounded data estimates a significantly different effect of the Athena SWAN Award status on the number of females for each cost centre, which is one of the questions of interest.

The poisson regression analyses of the Physics datasets for 2012 and 2016 suggest that the

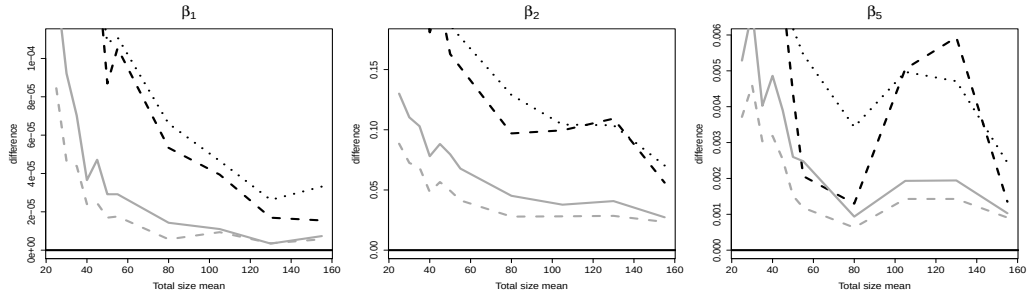

Figure 6.9: Difference between the estimated variances of the complete classical analysis and various models for a subset of the parameters in the ordinal logistic regression synthetic analysis. Classic results are shown in solid black, classical analysis of the rounded data in dashed black, EE results with no additional information in dotted black (6.15), EE results with additional totals information utilised in solid grey (6.16) and EE results with additional proportions information in dashed grey (6.17).

Athena SWAN status a_n and being located in England ($l_{n1} = 1$) or Scotland ($l_{n2} = 2$) are useful predictors of the number of females w_n, yielding comparably large (in magnitude) MLE's compared to their estimated standard deviations for β_1, β_{41} and β_{42}. Furthermore, a higher Athena SWAN status a_n is found to have an important positive effect on the number of females for a given cost centre. The significance and magnitude (in a traditional hypothesis testing setting) of all three country coefficient estimates ($\beta_{41}, \beta_{42}, \beta_{43}$) increases in the 2016 analysis, compared to 2012. Furthermore, the estimating function methodology (6.8) estimates larger coefficients with smaller standard deviations for the location coefficients for 2012 and 2016. Consequently, a naive classical analysis of the rounded data underestimates the importance of each location in the prediction of the number of females for each cost centre.

The binomial regression analyses of the Psychology 2012 dataset yield significant predictors (in the traditional hypothesis testing sense) for β_1, β_3 and β_{51} for both models, suggesting that the cost centre size (n_n), the length of engagement (t_n) and being located in England ($l_{n1} = 1$) are useful in predicting the proportion of females p_n for a given cost centre. The cost centre size is determined to have a important positive effect on the proportion of females. For the analysis of the 2016 version, the binomial regression analyses yield significant estimates for β_1, β_3 and β_4, suggesting the predictive ability of l_n has diminished, while the research intensity r_n now has a significant positive effect on the proportion of females. There are noticeable differences between the estimating function (6.8) and naive classical analyses in the results obtained for β_0 for both Psychology datasets, with the estimating function methodology (6.8) yielding different MLE's with significantly lower standard deviations.

The binomial regression models fit to the Physics datasets yields few significant predictors. For the 2012 dataset, apart from the intercept term β_0, only the estimate for β_{51} yields a significant p-value for the estimating function analysis (6.8), however this significance is lost in the naive classical analysis of the rounded data, with a much smaller estimate with a larger standard deviation being estimated. As a result, the estimating function (6.8) and naive classical methodologies yield different hypothesis test conclusions in determining the importance of being located

	Psychology			
	$\hat{\beta}_r^{2012}$	$\hat{\beta}_s^{2012}$	$\hat{\beta}_r^{2016}$	$\hat{\beta}_s^{2016}$
β_0	2.645 (0.440)*	2.797 (0.445)*	2.935 (0.337)*	2.925 (0.350)*
β_1	0.237 (0.145)	0.129 (0.165)	0.108 (0.151)	0.095 (0.151)
β_2	0.000 (0.003)	0.001 (0.003)	0.000 (0.003)	0.000 (0.003)
β_3	-0.029 (0.207)	-0.057 (0.202)	-0.126 (0.137)	-0.131 (0.137)
β_{41}	0.632 (0.273)*	0.639 (0.274)*	0.830 (0.189)*	0.875 (0.215)*
β_{42}	-0.077 (0.356)	-0.008 (0.347)	0.269 (0.269)	0.288 (0.293)
β_{43}	0.491 (0.493)	0.528 (0.502)	0.838 (0.278)*	0.909 (0.293)*
	Physics			
	$\hat{\beta}_r^{2012}$	$\hat{\beta}_s^{2012}$	$\hat{\beta}_r^{2016}$	$\hat{\beta}_s^{2016}$
β_0	0.063 (0.726)	-0.106 (0.834)	-0.801 (0.601)	-0.717 (0.752)
β_1	0.566 (0.199)*	0.637 (0.207)*	0.727 (0.183)*	0.704 (0.187)*
β_2	0.007 (0.004)	0.007 (0.004)	0.009 (0.004)*	0.009 (0.003)*
β_3	-0.165 (0.310)	-0.300 (0.324)	-0.025 (0.226)	-0.180 (0.293)
β_{41}	1.168 (0.165)*	1.313 (0.158)*	1.528 (0.121)*	1.703 (0.129)*
β_{42}	0.982 (0.267)*	1.161 (0.228)*	1.568 (0.235)*	1.717 (0.222)*
β_{43}	0.624 (0.927)	0.747 (0.683)	1.489 (0.942)	1.753 (0.746)*

Table 6.1: MLEs with standard deviations in parentheses for the Poisson regression analyses.

in England as a predictor of the proportion of females for a given cost centre. The analysis of the 2016 Physics dataset gives larger estimates with smaller standard deviations (compared to 2012) for the location coefficients, again with larger estimates and smaller standard deviations obtained for the estimating function methodology (6.8), compared to the naive classical analysis, meaning a naive classical analysis underestimates the importance of these variables in the prediction of the proportion of women. The significance of the estimates for β_2 and β_4 also increases in 2016 for both methodologies, meaning the Athena SWAN status a_n and research intensity r_n are increasingly predictive of the proportion of females for the Physics dataset.

The ordinal logistic regression models fit to the Psychology datasets for each year illustrate that the length of engagement t_n and research intensity r_n are the most useful predictors of the Athena SWAN status, yielding MLE's with large magnitudes when compared to their standard deviations. Similar estimates are obtained for the threshold parameter α_2 for both the estimating function (6.8) and naive classical analyses, with the estimating function methodology (6.8) yielding lower standard deviations for μ_2 in each year. None of the locations were found to be a significant predictor of the award status in these analyses.

The ordinal logistic analysis of the Physics 2012 dataset yields high MLE's with comparably low standard deviations for β_3, β_{51} and β_{53}, suggesting that the length of engagement t_n and location l_n are the most useful predictors of the Athena SWAN status. The significance of β_3 decreases in 2016, while the results for the other covariates remain largely similar. In both years, the estimating function methodology (6.8) estimates lowest standard deviations for the threshold parameter α_2.

	Psychology			
	$\hat{\beta}_r^{2012}$	$\hat{\beta}_s^{2012}$	$\hat{\beta}_r^{2016}$	$\hat{\beta}_s^{2016}$
β_0	-0.009 (0.205)	-0.134 (0.202)	0.287 (0.210)	0.277 (0.160)
β_1	0.001 (0.000)*	0.001 (0.000)*	0.001 (0.000)*	0.001 (0.000)*
β_2	0.053 (0.065)	0.025 (0.065)	0.020 (0.070)	-0.009 (0.064)
β_3	-0.004 (0.001)*	-0.004 (0.001)*	-0.004 (0.001)*	-0.005 (0.001)*
β_4	0.126 (0.065)	0.119 (0.061)	0.165 (0.053)*	0.127 (0.046)*
β_{51}	0.212 (0.080)*	0.343 (0.089)*	-0.076 (0.140)	0.020 (0.076)
β_{52}	-0.020 (0.145)	0.144 (0.141)	-0.006 (0.176)	-0.001 (0.110)
β_{53}	0.016 (0.270)	0.134 (0.259)	-0.268 (0.213)	-0.081 (0.170)
	Physics			
	$\hat{\beta}_r^{2012}$	$\hat{\beta}_s^{2012}$	$\hat{\beta}_r^{2016}$	$\hat{\beta}_s^{2016}$
β_0	-2.025 (0.478)*	-2.325 (0.397)*	-3.278 (0.398)*	-3.099 (0.364)*
β_1	0.001 (0.001)	0.000 (0.001)	0.000 (0.001)	0.000 (0.001)
β_2	0.049 (0.129)	0.106 (0.108)	0.262 (0.098)*	0.202 (0.079)*
β_3	-0.003 (0.003)	-0.002 (0.003)	0.002 (0.002)	0.001 (0.002)
β_4	0.164 (0.186)	0.198 (0.135)	0.293 (0.142)	0.252 (0.149)
β_{51}	0.176 (0.215)	0.406 (0.185)*	0.681 (0.137)*	0.868 (0.119)*
β_{52}	-0.054 (0.231)	0.171 (0.220)	0.583 (0.216)*	0.665 (0.210)*
β_{53}	-0.481 (0.643)	-0.244 (0.442)	0.561 (0.572)	0.843 (0.373)*

Table 6.2: MLEs with standard deviations in parentheses for the Binomial regression analyses.

6.6 Discussion

In this article, we have developed a methodology for obtaining estimates of glm parameters from datasets with rounded count variables, with the aim of improving upon the results of a naive classical analysis of the rounded dataset. The estimating function methodology (6.8) 'averages' out the contribution of each observation towards the log-likelihood across all possible underlying multivariate values that the observation could have taken before it was subjected to rounding, rather than just treating the rounded observation as the true underlying value, as per the naive classical analysis. Additional information concerning the summation of some variables and the availability of the proportion of females for some observations can then be included to further restrict the domain of the underlying latent dataset, and more accurately approximate the classical log-likelihood. Although they aren't directly included in the underlying classical glm log-likelihood that we want to approximate, the counts of the males for each employment contract v_{ni} can be considered additional information that allows better MLE's to be obtained from the estimating function methodology (6.8).

The estimating function methodology was utilised in the estimation of glm parameters for a poisson, binomial and ordinal logistic analyses of the Athena SWAN data, with the aim of determining the importance of each covariate in the prediction of the Athena SWAN award status, and also the number/proportion of women for a given university cost centre. Each analysis yielded important predictors, with noticeable differences emerging for some parameters between the estimating function methodology and naive classical analysis of the rounded data. Consequently, ignoring the rounding mechanism and performing a naive classical analysis of the

	Psychology			
	$\hat{\beta}_r^{2012}$	$\hat{\beta}_s^{2012}$	$\hat{\beta}_r^{2016}$	$\hat{\beta}_s^{2016}$
β_1	-0.003 (0.004)	-0.002 (0.007)	-0.001 (0.006)	-0.001 (0.008)
β_2	-0.031 (1.527)	2.168 (1.832)	0.373 (1.818)	0.076 (2.132)
β_3	-0.062 (0.011)*	-0.062 (0.014)*	-0.060 (0.010)*	-0.060 (0.013)*
β_4	0.967 (0.371)*	0.665 (0.377)	0.809 (0.410)	0.841 (0.434)
β_{51}	0.793 (0.840)	-0.241 (0.956)	0.569 (0.986)	0.680 (1.218)
β_{52}	-0.262 (0.833)	-0.931 (1.052)	-0.275 (1.047)	-0.151 (1.291)
β_{53}	0.443 (0.904)	-0.194 (1.051)	0.375 (1.085)	0.467 (1.348)
α_2	3.331 (0.641)*	3.247 (0.600)*	3.156 (0.583)*	3.157 (0.533)*
	Physics			
	$\hat{\beta}_r^{2012}$	$\hat{\beta}_s^{2012}$	$\hat{\beta}_r^{2016}$	$\hat{\beta}_s^{2016}$
β_1	-0.030 (0.010)*	-0.027 (0.009)*	-0.030 (0.012)*	-0.028 (0.008)*
β_2	3.411 (4.141)	1.169 (1.760)	-9.936 (5.387)	-1.879 (1.088)
β_3	-0.027 (0.012)*	-0.028 (0.012)*	-0.016 (0.013)	-0.024 (0.013)
β_4	-0.155 (0.487)	0.034 (0.433)	-0.387 (0.669)	-0.006 (0.454)
β_{51}	2.268 (1.117)*	2.320 (1.025)*	4.799 (2.188)*	3.101 (1.063)*
β_{52}	1.953 (1.450)	1.682 (1.361)	4.036 (2.230)	2.585 (1.468)
β_{53}	12.055 (1.055)*	5.668 (0.857)*	13.758 (2.374)*	6.314 (0.986)*
α_2	1.539 (0.449)*	1.568 (0.386)*	1.707 (0.549)*	1.638 (0.445)*

Table 6.3: MLEs with standard deviations in parentheses for the Ordinal logistic regression analyses.

rounded data will lead to innacurate conclusions regarding the importance of various covariates for the research questions posed.

Simulation studies were considered to demonstrate the utility of our estimating function (6.8) construction for datasets of the nature studied in this paper. The simulations showed a convergence of both the estimating function (6.8) and naive classical results towards that of the true underlying dataset with a decreasing degree of rounding. Furthermore, for the poisson, binomial and ordinal logistic regression models investigated, the estimating function methodology (6.8) with the most additional information incorporated provides the best parameter estimates. The ordinal logistic regression synthetic example highlighted a tendency of the estimated parameter standard deviations to converge to the classical results from below, rather than from above as per the other models investigated. As a result, it seems the estimating function (6.8) and naive classical methodologies underestimate the variance of each parameter, with the difference to the true results decreasing with decreasing degrees of rounding.

For classical glm's, the Fisher information matrix is sufficient for obtaining estimates for the variance of each MLE. However, due to the misspecified nature of the estimating function approximation developed in this paper, we obtain variances for both the estimating function (6.8) and naive classical methods through the inverse of the Godambe information matrix, which is equivalent to the Fisher information matrix for correctly specified log-likelihoods. Evaluation of the Godambe information matrix requires the evaluation of two matrix components, the jacobian and the hessian, which are also equal when the log-likelihood is properly specified. The variances

obtained from the Godambe information matrix allow inferences to be made on each parameter, allowing us to determine which covariates are significant predictors of the Athena SWAN award status and the number/proportion of females for a given cost centre.

The function $\frac{\Phi(x_n,y_n|\cdot)}{A_n}$ acts as a weighting function that ensures the sum of the weights of the contributions towards the estimating equation for each possible underlying value for a given observation is equal to 1. In this paper, we focus only on the case where $\Phi(x_n, y_n|\cdot)$ is the product of identity functions, taking the value 1 if the potential underlying value $(x_{n,n})$ is possible for the n^{th} observation, and 0 otherwise. This results in a summation of the estimating equations over all the possible underlying multivariate values w_n, v_n, n_n that were possible, given the rounded observations w_n^*, v_n^*, n_n^*, where each possible underlying value receives an equal weight. If more information is available, this density could be better specified. For example, if it known that some underying values are more likely than others, then the weights can be adjusted accordingly by a different choice of $\Phi(x_n, y_n|\cdot)$.

Chapter 7

Discussion and Future Work

Symbolic data analysis provides tools to deal with data that takes a non-standard (symbolic) form, such as intervals or histograms. Such constructions are often useful in reducing the computational burden associated with the storage, transmission and analysis of the dataset, given the size and dimension of a symbolic dataset is usually less than that of the underlying classical dataset from which it was aggregated. Furthermore, it is often necessary for data to arrive in a non-standard form, due to reasons such as privacy, or a natural aggregation that occurs during observation. Classical methods of analysis are often unsuitable for symbolic datasets, as they fail to take into account the inherent variability that occurs within each symbol, a property not present in classical pointwise observations. Most existing SDA methods focus on an exploratory or descriptive analysis of symbolic datasets, and are often dependant on an assumed uniform within-symbol distribution. These methods are effective if results are desired that are interpretable at the symbolic level, however if a classical framework is required for the output of the analysis then these methods are often unsuitable.

When data arrives in a non-standard form, an analysis is often required that delivers results comparable to that of a classical analysis of the underlying classical dataset from which the symbolic dataset was constructed. There are some examples of methodologies for non-standard data that accomplish this. Heitjan (1989) calculates variances from a binned dataset, and examines the differences between raw midpoint estimates, Sheppard's corrected estimates and estimates obtained from weighted parametric binned model, assuming a normal distribution for the underlying dataset. Beranger et al. (2018) provided a parametric framework in which a classical parametric model can be fit to a symbolic dataset. The results of the symbolic parametric model converge towards that of the corresponding classical analysis with increasing information retention in the data aggregation process. Parzen (1962) constructed a kernel density estimator for histogram data as the classical density estimator for the set of midpoints, each weighted by the counts for their respective bins. There are also parallels that exist between SDA and existing methodologies for missing data, however these methods often require the occurrence of classical-valued observations, from which a density for the missing data can be fit.

In Chapter 3, the parametric framework of Beranger et al. (2018) was extended to the field of composite likelihood (Cox and Reid, 2004, Lindsay, 1982, Varin, 2008) to address the computa-

tional issues associated with the analysis of large-dimensional histograms. A symbolic composite likelihood function was developed, with multivariate histograms used as the motivating example. It was demonstrated that the symbolic composite likelihood function obtains comparable mle's to a classical analysis of the complete microdata, if a certain amount of information is retained during the data aggregation process. Results were also obtained that show the loss in efficiency of these estimates that occurs as less temporal information is retained. Extensive simulation studies were performed to demonstrate the convergence of the symbolic results towards the classical results through the varying of the number of bins and temporal blocks, and the utility of this methodology was demonstrated via the analysis of several datasets consisting of historical and future simulated temperature observations from various climate scenarios.

In Chapter 4 a logistic regression model was developed that can accommodate covariates that take the form of marginal histograms. This analysis was not previously possible. By making certain assumptions about the relationship between the underlying classical covariates, estimates for the complete classical logistic regression model can be obtained from the optimisation of a set of lower-dimensional likelihoods. This method also allows the logistic regression analysis of large-dimensional multivariate covariate histograms, which cannot be analysed using the Beranger et al. (2018) approach due to the computational burden associated with estimating the large-dimensional integrals within the likelihood. Two real datasets were analysed using this construction, for which classical results were previously provided. In each case, the newly developed marginal histogram logistic regression model was able to obtain almost as good predictions as the classical analyses of the complete microdata at vastly cheaper computational cost.

Often a parametric analysis of symbolic datasets is unsuitable for the same reasons as it might be for a classical dataset. That is, the practitioner is unable to assume a parametric density for the classical data. In Chapter 5 we developed a non-parametric framework for the analysis of non-standard data that can be used to obtain non-parametric estimates for various statistics, such as means, variances, correlations and quantiles. Specific methodologies were developed for intervals and histograms, in which additional information available in surrounding symbols is utilised to estimate these quantities for each symbol. Empirical likelihood (Owen, 1988, 1990) has emerged in recent years as a highly effective method of estimating the variance of these quantities, without the need for an underlying parametric assumption. We extended the classical empirical likelihood methodology to the scenario where non-standard data is observed, whereby estimates for variances of parameters can be obtained from symbolic datasets that are comparable to those of the classical empirical likelihood methodology performed on the underlying microdata. The utility of these constructions was demonstrated via simulation studies and analyses of real datasets.

The work in Chapter 6 was motivated by the rounding that occurs in data collection during the creation of the Athena SWAN dataset. Counts for the numbers of men, women and people in each department are rounded to the nearest 5 for privacy reasons, leading to the occurence of a non-standard dataset. An analysis was desired that determined the importance of various variables in the prediction of the number/proportion of women and the Athena award status for each department. Ideally, a classical GLM model would be fit to the data. However, the

rounding mechanism means that this approach will deliver innaccurate results. In Chapter 6 we performed a GLM analysis of this dataset using symbolic methods, and have shown that we are able to improve the results by utilising some key relationships between the rounded covariates, such as the fact that the true underlying number of men and women must add up to the number of people. The subsequent likelihood is misspecified, and so estimates for the variances of the regression parameters are obtained through the estimated Godambe information matrix Godambe (1960).

In this thesis, methods of analysis for non-standard data were developed that obtain results which possess a classical interpretation. Furthermore, if a certain amount of information is retained during the aggregation process, these methods provide results that are comparable to that of the corresponding classical analysis of the latent microdata. The amount of information retention required to obtain comparable results is shown to be reasonable for the examples looked at in this thesis. For example, for the analyses of various histogram datasets, convergence to the classical results tended to occur with less than $B = 30$ marginal bins. Much of the work in this thesis focused on developing methods that utilise available additional information accompanying the symbols to account for the information loss associated with the data aggregation. This leads to each symbolic analysis providing closer parameter estimates to the analysis of the underlying microdata.

Throughout this thesis, the focus has mainly been on data exhibiting the i.i.d. (identically and independently distributed) property, which was appropriate for the various applications investigated here. However, there are instances in which data arises for which the i.i.d. assumption is not valid. In many cases, where the data can be considered exchangeable but not i.i.d., a hierarchical representation along the lines of that in Zhang et al. (2019) would be possible, directly extending the work in this thesis. In other cases, for example, in time series data, the ordering of the underlying data affects the subsequent analyses, and the simple time-aggregation of the data into symbolic objects such as histograms will lead to the loss of this information. There are also instances where the parameter values are expected to change over time. An example of this could be a different analysis of the millennial scale climate extremes to that seen in 3.5, where instead of assuming the marginal GEV parameters vary through space, they could be modelled as linear functions of time. In this model, it would be necessary to retain as much temporal information as possible within the marginal histograms in order to obtain comparable composite MLE's via the symbolic likelihood to the classical case, meaning a different construction may need to be considered. One potential construction that could address these concerns could be the aggregation of the data into separate multi-dimensional symbols for each time point. Of course, this construction would present its own challenges in the derivation of a likelihood function, and would be an interesting area of future work.

An area for expansion for the field of SDA that can further improve the effectiveness of the methods developed in this thesis that is mentioned but not developed is the question of symbol design. If the practitioner is aware of the models they want to fit prior to the aggregation of the classical data into a set of symbols, then an optimal symbol design could potentially be constructed that leads to more accurate results from fewer numbers of symbols, with obvious

benefits in computation. For example, for interval-valued data a better construction could be an optimal choice of end points, instead of the usual min-max construction. For histogram-valued observations, the usual choice is to construct each symbol using B equally-spaced bins. One potential expansion is to allow for uneven bins, and develop methodologies that determine the optimal location of the $B + 1$ histogram break points. The development of such methodologies that determine the optimal symbol construction for specific classical analyses is therefore a potentially highly significant future direction for SDA.

Appendix A

Chapter 4 Supporting information

A.1 Appendices

A.1.1 Proof of Proposition 4.3.1

We utilise the arguments presented in Beranger et al. (2018) to derive of Proposition 4.3.1. Note that if the underlying microdata $\mathbf{X} = (X_1, ..., X_N)$ are i.i.d. then the K subsets $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(K)}$ are similarly i.i.d. For histogram-valued data,

$$f(\mathbf{s}|\boldsymbol{x}, y, \vartheta) = \prod_{k=1}^{K} f(s_k|\boldsymbol{x}^{(k)}, \vartheta),$$

where

$$f(s_k|\boldsymbol{x}^{(k)}, \vartheta) = \prod_{\boldsymbol{b}_k=\mathbf{1}_k}^{\boldsymbol{B}_k} \mathbb{1}\left\{\sum_{n=1}^{N_k} \mathbb{1}\{x_n^{(k)} \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_k}\} = s_{\boldsymbol{b}_k}\right\}.$$

For the symbolic, histogram-based model, we therefore obtain

$$\begin{aligned}
L_{SM}(\mathbf{s};\beta) &= \int_{\mathcal{D}_{\mathbf{X}}} L_M(\boldsymbol{x}, y; \beta) f(\mathbf{s}|\boldsymbol{x}, y, \vartheta) \mathrm{d}\boldsymbol{x} \\
&= \prod_{k\in\Omega} \int_{\mathcal{D}_{\mathbf{X}^{(k)}}} L_M(\boldsymbol{x}^{(k)}, y; \beta) f(s_k|\boldsymbol{x}^{(k)}, \vartheta) \mathrm{d}\boldsymbol{x}^{(k)} \\
&= \prod_{k\in\Omega} \prod_{n=1}^{N_k} \prod_{\boldsymbol{b}_k=\mathbf{1}_k}^{\boldsymbol{B}_k} \left(\int_{\mathcal{D}_{X_n^{(k)}}} L_M(x_n^{(k)}, y_n; \beta) \mathrm{d}x_n^{(k)} \right)^{\mathbb{1}\{x_n^{(k)} \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_k}\}} \\
&\propto \prod_{k\in\Omega} \prod_{\boldsymbol{b}_k=\mathbf{1}_k}^{\boldsymbol{B}_k} \left(\int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}} P_M(y = k|X = x) \mathrm{d}x \right)^{s_{\boldsymbol{b}_k}}.
\end{aligned}$$

Similarly, for the histogram-based OvR model, we obtain

$$\begin{aligned}
L_{SO}(\mathbf{s};\beta) &= \int_{\mathcal{D}_{\mathbf{X}}} L_O(\boldsymbol{x},y;\beta) f(\mathbf{s}|\boldsymbol{x},y,\vartheta)\mathrm{d}\boldsymbol{x} \\
&= \prod_{k\in\Omega}\int_{\mathcal{D}_{\mathbf{X}^{(k)}}} L_O(\boldsymbol{x}^{(k)},y;\beta) f(s_k|\boldsymbol{x}^{(k)},\vartheta)\mathrm{d}\boldsymbol{x}^{(k)} \\
&= \prod_{k\in\Omega}\prod_{n=1}^{N_k}\prod_{\boldsymbol{b}_k=\mathbf{1}_k}^{\boldsymbol{B}_k}\left(\int_{\mathcal{D}_{X_n^{(k)}}} L_O(x_n^{(k)},y_n;\beta)\mathrm{d}x_n^{(k)}\right)^{\mathbb{1}\{x_n^{(k)}\in\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}\}} \\
&\propto \prod_{k\in\Omega}\prod_{\boldsymbol{b}_k=\mathbf{1}_k}^{\boldsymbol{B}_k}\left(\int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}} P_{\mathrm{O}}(Y=k|X=x)\mathrm{d}x \prod_{k'\in\Omega\setminus\{k\}}\int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}} P_{\mathrm{O}}(Y\neq k'|X=x)\mathrm{d}x\right)^{s_{\boldsymbol{b}_k}}.
\end{aligned}$$

A.1.2 Proof of Proposition 4.3.2

We now show that if there is neither complete nor quasi-complete separation of the set of histograms $\boldsymbol{s}$, then $L_{SO}(\mathbf{s};\beta)$ and $L_{SM}(\mathbf{s};\beta)$ have unique global maxima. The following arguments are analogous to those proposed by Albert and Anderson (1984). Suppose that there is complete separation exhibited by the histogram dataset for the k^{th} class, according to Definition 4.3.2.

As a result, the complete separation property holds for all vectors $\beta_k = a_k b_k$ for $a_k > 0$. We now examine the behaviours of the integrals of the $P_{\mathrm{O}}(Y=k|X)$ and $P_{\mathrm{O}}(Y\neq k|X)$ terms in the likelihood functions (4.6) and (4.7). Using the mean value theorem, and given $a_k b_k^\top x^*_{\boldsymbol{b}_j} > 0$ if $j=k$ and $a_k b_k^\top x^*_{\boldsymbol{b}_j} < 0$ if $j\neq k$ for all non-empty bins and $a_k > 0$, we obtain

$$\begin{aligned}
\lim_{a_k\to\infty}\int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}} P_{\mathrm{O}}(Y=k|x)\mathrm{d}x &= \lim_{a_k\to\infty}\int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}} \frac{e^{a_k b_k^\top x}}{1+e^{a_k b_k^\top x}}\mathrm{d}x \propto \lim_{a_k\to\infty}\frac{e^{a_k b_k^\top x^*_{\boldsymbol{b}_k}}}{1+e^{a_k b_k^\top x^*_{\boldsymbol{b}_k}}} = 1 \\
\lim_{a_k\to\infty}\int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}} P_{\mathrm{O}}(Y\neq k|x)\mathrm{d}x &= \lim_{a_k\to\infty}\int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}} \frac{1}{1+e^{a_k b_k^\top x}}\mathrm{d}x \propto \lim_{a_k\to\infty}\frac{1}{1+e^{a_k b_k^\top x^*_{\boldsymbol{b}_k}}} = 1,
\end{aligned}$$

where $x^*_{\boldsymbol{b}_k}$ is some point inside $\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}$. Each integral therefore approaches a constant for all bins as a_k increases, meaning the maximum value of each likelihood function is attained at the boundary of the parameter space, i.e. $\hat{\beta}_k = \infty$.

Now suppose there is quasi-complete separation exhibited by the histogram dataset, according to Definition 4.3.2. Continuing with the previous notation, denote A_k^{D+1} as the set of all vectors b_k that satisfy the complete separation condition, meaning that A_k^{D+1} is a convex set. Denote the parameter vector $\alpha_k(a_k) = c_k + a_k b_k$, where $a_k > 0$ and $a_k \in A^{D+1}$. Consequently,

$$P_{\mathrm{O}}(Y=k|x) = \frac{e^{\alpha_k(a)^\top x}}{1+e^{\alpha_k(a_k)^\top x}}.$$

The log-likelihood for the component of the OvR model (4.7) estimating the parameters for the

k^{th} class, β_k, can therefore be expressed as

$$\begin{aligned}\log L^k_{SO}(\mathbf{s};\beta_k) &= \sum_{\boldsymbol{b}_k=\mathbf{1}_k}^{\boldsymbol{B}_k} s_{\boldsymbol{b}_k} \log \int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}} P_{\mathrm{O}}(Y=k|x)\mathrm{d}x + \sum_{k'\in\Omega\setminus\{k\}} \sum_{\boldsymbol{b}_{k'}=\mathbf{1}_{k'}}^{\boldsymbol{B}_{k'}} s_{\boldsymbol{b}_{k'}} \log \int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_{k'}}} P_{\mathrm{O}}(Y\neq k|x)\mathrm{d}x \\ &= \sum_{\boldsymbol{b}_k=\mathbf{1}_k}^{\boldsymbol{B}_k} s_{\boldsymbol{b}_k} \log \int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}} \frac{e^{c_k^\top x+a_k b_k^\top x}}{1+e^{c_k^\top x+a_k b_k^\top x}}\mathrm{d}x + \sum_{k'\in\Omega\setminus\{k\}} \sum_{\boldsymbol{b}_{k'}=\mathbf{1}_{k'}}^{\boldsymbol{B}_{k'}} s_{\boldsymbol{b}_{k'}} \log \int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_{k'}}} \frac{1}{1+e^{c_k^\top x+a_k b_k^\top x}}\mathrm{d}x.\end{aligned}$$

Given that $b_k^\top x > 0$ for all $x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_k}$ such that $s_{\boldsymbol{b}_k} > 0$, the function $\frac{e^{c_k^\top x+a_k b_k^\top x}}{1+e^{c_k^\top x+a_k b_k^\top x}}$ is monotonically increasing with a_k for all $x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_k}$ such that $s_{\boldsymbol{b}_k} > 0$. Consequently,

$$\sum_{\boldsymbol{b}_k=\mathbf{1}_k}^{\boldsymbol{B}_k} s_{\boldsymbol{b}_k} \log \int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}} \frac{e^{c_k^\top x+a_k b_k^\top x}}{1+e^{c_k^\top x+a_k b_k^\top x}}\mathrm{d}x$$

is monotonically increasing with increasing a_k. Similarly, given that $b_k^\top x < 0$ for all $x \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_{k'}}$ such that $s_{\boldsymbol{b}_{k'}} > 0$ and $k' \neq k$

$$\sum_{k'\in\Omega\setminus\{k\}} \sum_{\boldsymbol{b}_{k'}=\mathbf{1}_{k'}}^{\boldsymbol{B}_{k'}} s_{\boldsymbol{b}_{k'}} \log \int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_{k'}}} \frac{1}{1+e^{c_k^\top x+a_k b_k^\top x}}\mathrm{d}x$$

is monotonically increasing with increasing a_k. Therefore the log-likelihood function $\log L^k_{SO}(\mathbf{s};\beta)$ is monotonically increasing with a_k, and the maximum value is attained at the boundary of the parameter domain, i.e. $\hat{\beta}_k = \infty$.

The above arguments show that if there is complete or quasi-complete separation for any class $k \in \Omega$, then there is no unique MLE for the symbolic OvR model. Using similar arguments, it can be shown that the maximum value for the symbolic multinomial likelihood $\log L_{SM}(\mathbf{s};\boldsymbol{\beta})$ is attained at $\hat{\boldsymbol{\beta}} = (\infty,\ldots,\infty)^\top$ if there is either complete or quasi-complete separation in the data.

By the mean value theorem,

$$\begin{aligned}L_{SM}(\mathbf{s};\boldsymbol{\beta}) &\propto \prod_{k\in\Omega} \prod_{\boldsymbol{b}_k=\mathbf{1}_k}^{\boldsymbol{B}_k} \left(\int_{\boldsymbol{\Upsilon}_{\boldsymbol{b}_k}} P_M(y=k|X=x)\mathrm{d}x\right)^{s_{\boldsymbol{b}_k}} \\ &\propto \prod_{k\in\Omega} \prod_{\boldsymbol{b}_k=\mathbf{1}_k}^{\boldsymbol{B}_k} P_M(y=k|X=x^*_{\boldsymbol{b}_k})^{s_{\boldsymbol{b}_k}},\end{aligned}$$

where $x^*_{\boldsymbol{b}_k} \in \boldsymbol{\Upsilon}_{\boldsymbol{b}_k}$ is some point located inside the $\boldsymbol{b}_k^{th}$ bin. The symbolic multinomial likelihood $L_{SM}(\mathbf{s};\boldsymbol{\beta})$ is therefore proportional to the classical likelihood for some dataset $\boldsymbol{x}^*$, consisting of $\sum_{k\in\Omega}\sum_{\boldsymbol{b}_k=\mathbf{1}_k}^{\boldsymbol{B}_k} \mathbb{1}\{s_{\boldsymbol{b}_k} > 0\}$ distinct values $x^*_{\boldsymbol{b}_k}$, each appearing $s_{\boldsymbol{b}_k}$ times, $\boldsymbol{b}_k = \mathbf{1}_k,\ldots,\boldsymbol{B}_k$, $k \in \Omega$. Albert and Anderson (1984) proved that $L_{\mathrm{M}}(\boldsymbol{x},y;\boldsymbol{\beta})$ is a closed convex function. Therefore, $L_{\mathrm{M}}(\boldsymbol{x}^*,y;\boldsymbol{\beta})$ and subsequently $L_{SM}(\mathbf{s};\boldsymbol{\beta})$ are closed convex functions. Similar arguments can be used to show the closed convex nature of $L_{SO}(\mathbf{s};\boldsymbol{\beta})$. As a result, if the histogram-valued data does not exhibit complete separation or quasi-complete separation (Definition 4.3.2), then

there is a unique global maximum of $L_{SM}(\mathbf{s};\boldsymbol{\beta})$. Similarly, if the histogram-valued data does not exhibit complete separation or quasi-complete separation for any class $k \in \Omega$ (Definition 4.3.1), then there is a unique global maximum of $L_{SO}(\mathbf{s};\boldsymbol{\beta})$.

A.1.3 Proof of Proposition 4.3.3

We first use the arguments in Cramer (2007), Wooldridge (2002) to derive $L_O^{(j)}$ and $L_{SO}^{(j)}$ using the latent variable formulation of the logistic regression model (equivalent to the log odds formulation in Section 4.2). Consider a binary ($K = 2$) logistic regression model, such that the OvR and multinomial model are equivalent. The latent variable formulation for the logistic regression model based on the $\boldsymbol{i}^{th}$ subset of variables can be written as

$$Y_n^* = \beta^{\boldsymbol{i}\top} X_n^{\boldsymbol{i}} + e_n^{\boldsymbol{i}}, \qquad n = 1, \ldots, N, \tag{A.1}$$

where Y_n^* is an unseen latent variable and $e_n^{\boldsymbol{i}}$ is an error term following a logistic distribution. Classification is then achieved by setting $Y_n = 1$ if $Y_n^* < 0$ and $Y_n = 2$ otherwise. In the full model

$$Y_n^* = \beta^\top X_n + u_n,$$

where u_n follows a logistic distribution with zero mean and unit variance. In the smaller model (A.1) indexed by $\boldsymbol{i}$, the omitted terms are absorbed into the error term. That is

$$e_n^{\boldsymbol{i}} = u_n + \sum_{i' \in \mathcal{I}_1^{-\boldsymbol{i}}} \beta^{i'} X_{ni'}.$$

Suppose that there are correlations between included and omitted variables for the model based on subset $\boldsymbol{i}$, and we can express each omitted variable as a linear function of the included variables, i.e. $X^{i'} = \alpha_{\boldsymbol{i}i'}^\top X^{\boldsymbol{i}} + \epsilon_{\boldsymbol{i}i'}$, as described in Section 4.3.3. W.l.o.g. we can assume the i^{th} variable has zero mean and variance given by σ_i^2, and that the covariance between variables i and i' is given by $\sigma_{ii'}$. The error term $e_n^{\boldsymbol{i}}$ can therefore be expressed as

$$e_n^{\boldsymbol{i}} = u_n + \sum_{i' \in \mathcal{I}_1^{-\boldsymbol{i}}} \beta^{i'} \left(\alpha_{\boldsymbol{i}i'}^\top X_n^{\boldsymbol{i}} + \epsilon_{n\boldsymbol{i}i'} \right).$$

We now rewrite (A.1) by absorbing the terms in $e_n^{\boldsymbol{i}}$ that are dependent on the included variables into the model. That is

$$Y_n^* = \sum_{i_1' \in \boldsymbol{i}} \left(\beta^{i_1'} + \sum_{i_2' \notin \boldsymbol{i}} \beta^{i_2'} \alpha_{i_1' i_2'} \right) X_{ni_1'} + \tilde{e}_n^{\boldsymbol{i}},$$

where

$$\tilde{e}_n^{\boldsymbol{i}} = u_n + \sum_{i' \notin \boldsymbol{i}} \beta^{i'} \epsilon_{n\boldsymbol{i}i'}.$$

Pingel (2014) shows that if X is distributed according to a logistic distribution with mean 0 and variance $\frac{\pi^2}{3}$, then a standard normal distribution also fits the distribution of X reasonably well. The similarities between the standard logistic density and a rescaled normal density have also been investigated by Jeffress (1973), Bowling and Khasawneh (2009) and Pingel (2014), whereby different values for the rescaling factor C are proposed based on the criteria used to match the logistic and normal distributions. In practise any of these values can be used here, but we proceed with $\frac{\pi^2}{3}$ in the simulations and real data analyses in Sections 4.4 and 4.5 respectively. As a result, $\sum_{i' \notin \boldsymbol{i}} \beta^{i'} \epsilon_{nii'}$ is approximately normally distributed, and the error term $\tilde{e}_n^{\boldsymbol{i}}$ is approximately logistically distributed with mean zero (i.e. $\mathbb{E}(\tilde{e}_n^{\boldsymbol{i}}) = 0$) and variance given by

$$\text{Var}(\tilde{e}_n^{\boldsymbol{i}}) = 1 + \frac{\sum_{i'_1 \in \mathcal{I}_1^{-\boldsymbol{i}}} \left(\beta^{i'_1 2} \lambda^2_{\boldsymbol{i}i'_1} + 2 \sum_{i'_2 \in \mathcal{I}_1^{-\boldsymbol{i}}, i'_2 \neq i'_1} \beta^{i'_1} \beta^{i'_2} \lambda_{\boldsymbol{i}i'_1 i'_2} \right)}{\pi^2/3}.$$

In the full model,

$$P(Y_n = 2|X_n) = P(Y_n^* > 0|X_n) = P(u_n < \beta^\top X_n) = \frac{\exp\{\beta^\top X_n\}}{1 + \exp\{\beta^\top X_n\}},$$

and we obtain MLE's $\hat{\beta}$ for β by maximising over the sum of this quantity over all observations. Note that $P(u_n < \beta^\top X_n) = P_O(Y_n = 2|X_n)$, resulting in the equivalency between the latent and log odds formulations of the logistic regression model. For the omitted variable model,

$$P(Y_n = 2|X_n^{\boldsymbol{i}}) = P(Y_n^{\boldsymbol{i}*} > 0|X_n^{\boldsymbol{i}}) = P\left(\tilde{e}_n^{\boldsymbol{i}} < \sum_{i'_1 \in \boldsymbol{i}} \left(\beta^{i'_1} + \sum_{i'_2 \notin \boldsymbol{i}} \beta^{i'_2} \alpha_{i'_1 i'_2} \right) X_{ni'_1} \right).$$

Rescaling $\tilde{e}_n^{\boldsymbol{i}}$ by its standard deviation gives us a random variable with approximately the same distribution as u_n, i.e. $\tilde{u}_n = \frac{\tilde{e}_n^{\boldsymbol{i}}}{\sqrt{\text{Var}(e_n^{\boldsymbol{i}})}}$ will approximately follow a logistic distribution with zero mean and unit variance. As a consequence,

$$\begin{aligned} P(Y_n = 2|X_n^{\boldsymbol{i}}) &= P\left(\frac{\tilde{e}_n^{\boldsymbol{i}}}{\sqrt{\text{Var}(\tilde{e}_n^{\boldsymbol{i}})}} < \frac{\sum_{i'_1 \in \boldsymbol{i}} \left(\beta^{i'_1} + \sum_{i'_2 \notin \boldsymbol{i}} \beta^{i'_2} \alpha_{i'_1 i'_2} \right) X_{ni'_1}}{\sqrt{\text{Var}(\tilde{e}_n^{\boldsymbol{i}})}} \right) \\ &\approx P\left(\tilde{u}_n < \frac{\sum_{i'_1 \in \boldsymbol{i}} \left(\beta^{i'_1} + \sum_{i'_2 \notin \boldsymbol{i}} \beta^{i'_2} \alpha_{i'_1 i'_2} \right) X_{ni'_1}}{\sqrt{\text{Var}(\tilde{e}_n^{\boldsymbol{i}})}} \right) \\ &= \frac{\exp\{\widetilde{\beta}^{\boldsymbol{i}\top} X_n^{\boldsymbol{i}}\}}{1 + \exp\{\widetilde{\beta}^{\boldsymbol{i}\top} X_n^{\boldsymbol{i}}\}}, \end{aligned}$$

where $\widetilde{\beta}^{\boldsymbol{i}} = \frac{\beta^{\boldsymbol{i}} + \left[0, \left(\sum_{i' \in \mathcal{I}_1^{-\boldsymbol{i}}} \beta^{i'} \alpha_{ii'} \right)^\top \right]^\top}{\sqrt{\text{Var}(\tilde{e}_n^{\boldsymbol{i}})}} \in \mathbb{R}^{(j+1)}$. Therefore we see that the regression parameters $\widetilde{\beta}^{\boldsymbol{i}}$ for the OvR model fit to the data indexed by $\boldsymbol{i}$ can be expressed as functions of the regression parameters β for the complete $D-$dimensional OvR model. The value for $\widetilde{\beta}^{\boldsymbol{i}}$ that maximises the binary logistic likelihood for the $\boldsymbol{i}^{th}$ subset is therefore a rescaled version of the value for $\beta^{\boldsymbol{i}}$ that maximises the complete binary logistic likelihood.

Albert and Anderson (1984) proved that if the dataset $\boldsymbol{x}$ does not exhibit complete or quasi-complete separation, then there is a unique value for the MLE $\hat{\beta}$ for the complete model regression parameter β. As a result, unique values for the MLE $\hat{\beta}^{\boldsymbol{i}}$ for $\widetilde{\beta}^{\boldsymbol{i}}$ exist if the data does not exhibit complete separation in the variables indexed by $\boldsymbol{i}$. It is trivial to show that if the complete dataset $\boldsymbol{x}$ does not exhibit complete or quasi-complete separation, then there is no $\boldsymbol{i} \in I_j$ for which $\boldsymbol{x}^{\boldsymbol{i}}$ exhibits complete or quasi-complete separation. If there was, then there would exist a $b^{\boldsymbol{i}}$ such that

$$b^{\boldsymbol{i}\top} x_n^{\boldsymbol{i}} > 0 \text{ for all } n \text{ such that } y_n = 2$$
$$b^{\boldsymbol{i}\top} x_n^{\boldsymbol{i}} < 0 \text{ for all } n \text{ such that } y_n = 1.$$

Setting $b_d = b_d^{\boldsymbol{i}}$ if $d \in \boldsymbol{i}$ and 0 otherwise yields the vector $b = (b_1, \ldots, b_D)$ such that

$$b^\top x_n = b^{\boldsymbol{i}\top} x_n^{\boldsymbol{i}} > 0 \text{ for all } n \text{ such that } y_n = 2$$
$$b^\top x_n = b^{\boldsymbol{i}\top} x_n^{\boldsymbol{i}} < 0 \text{ for all } n \text{ such that } y_n = 1,$$

which is a contradiction. As a result, a sufficient condition for the existence and uniqueness of all MLE's $\hat{\beta}^{\boldsymbol{i}}$ for $\widetilde{\beta}^{\boldsymbol{i}}$, $\boldsymbol{i} \in \boldsymbol{i}$, is that the complete data $\boldsymbol{x}$ does not exhibit complete or quasi-complete separation. Now, given $\hat{\beta}^{\boldsymbol{i}}$ is the value for $\widetilde{\beta}^{\boldsymbol{i}}$ that minimises the log-likelihood for the $\boldsymbol{i}^{th}$ model

$$\log L(\boldsymbol{x}^{\boldsymbol{i}}, y; \widetilde{\beta}^{\boldsymbol{i}}) = \sum_{n=1}^{N} \mathbb{1}\{y_n = 1\} \log P(Y = 1 | x_n^{\boldsymbol{i}}, y_n, \widetilde{\beta}^{\boldsymbol{i}}) + \mathbb{1}\{y_n = 2\} \log P(Y = 2 | x_n^{\boldsymbol{i}}, y_n, \widetilde{\beta}^{\boldsymbol{i}}),$$

i.e. $\log L(\boldsymbol{x}^{\boldsymbol{i}}, y; \hat{\beta}^{\boldsymbol{i}}) < \log L(\boldsymbol{x}^{\boldsymbol{i}}, y; \widetilde{\beta}^{\boldsymbol{i}})$ for all $\widetilde{\beta}^{\boldsymbol{i}} \in \mathcal{D}_{\widetilde{\beta}^{\boldsymbol{i}}}$, the parameter $\hat{\boldsymbol{\beta}}^{(j)} = \left\{\hat{\beta}^{\boldsymbol{i}}\right\}_{\boldsymbol{i} \in I_j}$ therefore minimises the log-likelihood

$$\log L_{\mathrm{O}}^{(j)}(\boldsymbol{x}, y; \widetilde{\boldsymbol{\beta}}^{(j)}) = \sum_{\boldsymbol{i} \in I_j} \log L(\boldsymbol{x}^{\boldsymbol{i}}, y; \tilde{\boldsymbol{\beta}}^{\boldsymbol{i}}),$$

i.e. $\log L_{\mathrm{O}}^{(j)}(\boldsymbol{x}, y; \hat{\boldsymbol{\beta}}^{(j)}) < \log L_{\mathrm{O}}^{(j)}(\boldsymbol{x}, y; \widetilde{\boldsymbol{\beta}}^{(j)})$ for all $\widetilde{\boldsymbol{\beta}^{(j)}} \in \mathcal{D}_{\widetilde{\boldsymbol{\beta}^{(j)}}}$. Therefore an estimate $\hat{\beta}$ for β such that

$$\hat{\beta}^{\boldsymbol{i}} = \frac{\beta^{\boldsymbol{i}} + \left[0, \left(\sum_{i' \in \mathcal{I}_1^{-\boldsymbol{i}}} \hat{\beta}^{i'} \alpha_{\boldsymbol{i}i'}\right)^\top\right]^\top}{\sqrt{1 + \frac{\sum_{i'_1 \in \mathcal{I}_1^{-\boldsymbol{i}}} \left(\beta^{i'_1 2} \lambda^2_{\boldsymbol{i}i'_1} + 2 \sum_{i'_2 \in \mathcal{I}_1^{-\boldsymbol{i}}, i'_2 \neq i'_1} \beta^{i'_1} \beta^{i'_2} \lambda_{\boldsymbol{i}i'_1 i'_2}\right)}{\pi^2/3}}},$$

(i.e. an estimate for β that yields the MLE's for each of the smaller-dimensional logistic models based on the variables indexed by $\boldsymbol{i}$) will minimise $\log L_{\mathrm{O}}^{(j)}(\boldsymbol{x}, y; \widetilde{\boldsymbol{\beta}}^{(j)})$. By the definition of the model this $\hat{\beta}$ will exist. Furthermore, given the above equations can be reduced to a polynomial system of equations with quasi-random coefficients and significantly more equations than unknowns, there is only one configuration of $\boldsymbol{\beta}$ that will lead to $\hat{\boldsymbol{\beta}}$, meaning the estimates obtained from the maximisation of the symbolic j-wise likelihood are estimates of the regression parameter for the complete underlying model. Given that a logistic OvR model is just the product of

K binary logistic regression models, the above results hold for the OvR model for $K > 2$ classes. Consequently, the $j-$dimensional OvR model can therefore be written as

$$L_{\mathrm{O}}^{(j)}(\boldsymbol{x}, y; \boldsymbol{\beta}) = \prod_{\boldsymbol{i} \in \mathcal{I}_j} L_{\mathrm{O}}(\boldsymbol{x}^{\boldsymbol{i}}, y; \widetilde{\boldsymbol{\beta}}^{\boldsymbol{i}}),$$

where

$$\widetilde{\beta}_k^{\boldsymbol{i}} = \frac{\beta_k^{\boldsymbol{i}} + \left[0, \left(\sum_{i' \in \mathcal{I}_1^{-\boldsymbol{i}}} \beta_k^{i'} \alpha_{\boldsymbol{i}i'}\right)^\top\right]^\top}{\sqrt{\mathrm{Var}(\tilde{e}_{nk}^{\boldsymbol{i}})}} \in \mathbb{R}^{(j+1)}.$$

Through similar arguments, we can find an expression for the symbolic j-dimensional OvR model as

$$L_{\mathrm{SO}}^{(j)}(\boldsymbol{s}; \boldsymbol{\beta}) = \prod_{\boldsymbol{i} \in \mathcal{I}_j} L_{\mathrm{SO}}(\boldsymbol{s}^{\boldsymbol{i}}, y; \widetilde{\boldsymbol{\beta}}^{\boldsymbol{i}}).$$

Appendix B

Chapter 5 Supporting information

B.1 Appendices

B.1.1 Proof of Symbolic Empirical Likelihood

The following arguments are taken from the classical derivation of EL, and applied to the non-standard data setting. Given the constraints (C3), (C4) and (C5) defined in Section 5.3, the solution for $\hat{q}_N^{(c)}$ can be obtained using Lagrangian multipliers, with the solution obtained by minimising the Lagrangian function

$$Q = \sum_{c=1}^{C} n_c \log q_n^{(c)} - \zeta \left(\sum_{c=1}^{C} n_c q_n^{(c)} - 1 \right) - \lambda^\top \left(\sum_{c=1}^{C} n_c q_n^{(c)} g'(s_c;\theta) \right)$$

with respect to $\boldsymbol{q} = (q_1, \ldots, q_C)$, ζ and λ. Equating the partial derivatives to zero yields the following:

$$\left\{ \begin{array}{lll} \frac{\partial Q}{\partial q_n^{(c)}} & = \frac{n_c}{q_n^{(c)}} - \zeta n_c - n_c \lambda^\top g'(s_c;\theta) & = 0 \\ \nabla_\lambda Q & = -(\sum_{c=1}^{C} n_c q_n^{(c)} g'(s_c;\theta)) & = 0 \\ \frac{\partial Q}{\partial \zeta} & = -(\sum_{c=1}^{C} n_c q_n^{(c)} - 1) & = 0 \end{array} \right. \implies \left\{ \begin{array}{ll} q_n^{(c)} & = \frac{1}{\zeta + N\lambda^\top g'(s_c;\theta)} \\ 0 & = -\sum_{c=1}^{C} \frac{n_c g'(s_c;\theta)}{N\{1+\lambda^\top g'(s_c;\theta)\}} \\ \zeta & = N \end{array} \right. , \tag{B.1}$$

from which the solution follows. Now

$$\nabla_\lambda^2 Q = \sum_{c=1}^{C} \frac{n_c g'(s_c;\theta) g'(s_c;\theta)^\top}{(N - \lambda^\top g'(s_c;\theta))^2}.$$

Clearly the denominator is greater than 0, and each numerator term is a positive definite matrix. Consequently there exists a unique solution for λ.

B.1.2 Evaluating m_c for intervals

If we utilise the information contained in ϑ as described in Section 5.4.1 to evaluate the following integrals required to evaluate the m_c function:

$$\int_{\Upsilon_{c_d}} x_{[d]}^k \mathbb{1}\{x_{[d]} \in \upsilon_{b_d}\} \mathrm{d}x_{[d]} = \frac{(z_{b_d}^d)^{k+1} - (z_{b_d-1}^d)^{k+1}}{k+1} \mathbb{1}\{\upsilon_{b_d}^d \subset \Upsilon_{c_d}^d\},$$

so that

$$\begin{aligned}
\int_{\Upsilon_c} x_{[d]}^k \mathbb{1}\{x \in \upsilon_{\boldsymbol{b}}\} \mathrm{d}x &= \int_{\Upsilon_c} x_{[d]}^k \prod_{d'=1}^{D} \mathbb{1}\{x_{[d']} \in \upsilon_{b_{d'}}^{d'}\} \mathrm{d}x \\
&= \left\{ \int_{\Upsilon_{c_d}^d} x_{[d]}^k \mathbb{1}\{x_{[d]} \in \upsilon_{b_d}^d\} \mathrm{d}x_{[d]} \right\} \times \left\{ \prod_{d' \neq d} \int_{\Upsilon_{b_{d'}}^{d'}} \mathbb{1}\{x_{[d']} \in \upsilon_{b_{d'}}^{d'}\} \mathrm{d}x_{[d]} \right\} \\
\int_{\Upsilon_c} x_{[d]} x_{[e]} \mathbb{1}\{x \in \Upsilon_{c'}\} \mathrm{d}x &= \left\{ \int_{\Upsilon_{c_d}^d} x_{[d]} \mathbb{1}\{x_{[d]} \in \upsilon_{c_d}^d\} \mathrm{d}x_{[d]} \int_{\Upsilon_{c_e}^e} x_{[e]} \mathbb{1}\{x_{[e]} \in \upsilon_{b_e}^e\} \mathrm{d}x_{[e]} \right\} \\
&\quad \times \left\{ \prod_{d' \neq d,e} \int_{\Upsilon_{c'_d}^{d'}} \mathbb{1}\{x_{[d']} \in \upsilon_{b_{d'}}^{d'}\} \mathrm{d}x_{[d']} \right\}
\end{aligned}$$

for any $k \in \mathbb{N}$, $d, e = 1, \ldots, D$, $d \neq e$, $c, c' = 1, \ldots, C$.

B.1.3 Derivation of $\phi_c(x)$ for histograms

When constructing a histogram such as that described in Section 5.4.2, the location of the multivariate break points $\Upsilon = ((y_1^d, \ldots, y_1^{C^1}), \ldots, (y_1^D, \ldots, y_1^{C^D}))$ is often an arbitrary choice and a unique histogram $\boldsymbol{s}' = (s'_{\boldsymbol{1}}, \ldots, s'_{\boldsymbol{C}})$ can be contructed for any choice of $\Upsilon' = \Upsilon + \boldsymbol{u}$, $\boldsymbol{u} = (u_1, \ldots, u_D)$, where $\boldsymbol{u} \in \left(-\frac{\delta}{2}, \frac{\delta}{2}\right)$ and $\boldsymbol{C}$ is constant. The bin locations $\Upsilon'_{\boldsymbol{c}}$ could therefore be considered a realisation $\Upsilon'_{\boldsymbol{c}} = \Upsilon_{\boldsymbol{c}} + \boldsymbol{u}$ of a random variable $\hat{\Upsilon}_{\boldsymbol{c}} = \Upsilon_{\boldsymbol{c}} + \mathbf{U}$, such that $\mathbf{U} \sim U\left(-\frac{\delta}{2}, \frac{\delta}{2}\right)$, and

$$H_x(\Upsilon'_{\boldsymbol{c}}) = f_{\mathbf{U}}(\boldsymbol{u}) = \frac{1}{\prod_{d=1}^D \delta_d} \mathbb{1}\{x \in \Upsilon'_{\boldsymbol{c}}\}.$$

Note that

$$\mathbb{P}\left(X \in \Upsilon'_{\boldsymbol{c}'} | X \in \Upsilon_{\boldsymbol{c}''}, \boldsymbol{s}\right) = \frac{|\Upsilon'_{\boldsymbol{c}'} \cap \Upsilon_{\boldsymbol{c}''}|}{|\Upsilon_{\boldsymbol{c}''}|},$$

and that each bin is disjoint, meaning a classical observation x can only fall in one bin region if the break points are fixed. Then let $\mathbf{A}_{\boldsymbol{c}''}(\Upsilon'_{\boldsymbol{c}'}) = \left(A_{c''_1}(\Upsilon^{1'}_{c'_1}), \ldots, A_{c''_D}(\Upsilon^{D'}_{c'_D})\right)$, $\boldsymbol{B}_{\boldsymbol{c}''}(\Upsilon'_{\boldsymbol{c}'}) =$

$\left(B_{c_1''}(\Upsilon_{c_1'}^{1'}),\ldots,B_{c_D''}(\Upsilon_{c_D'}^{D'})\right)$ and $\mathbf{C}_{\boldsymbol{c}''}(\Upsilon'_{\boldsymbol{c}'}) = \left(C_{c_1''}(\Upsilon_{c_1'}^{1'}),\ldots,C_{c_D''}(\Upsilon_{c_D'}^{D'})\right)$ such that

$$\begin{aligned}
A_{c_d''}\left(\Upsilon_{c_d'}^{d'}\right) &= \mathbb{1}\left\{y_{c_d'-1}^{d'} \in \left(y_{c_d''-1}^{d}, y_{c_d''}^{d}\right)\right\}\left(y_{c_d''}^{d} - y_{c_d'-1}^{d'}\right)\\
B_{c_d''}\left(\Upsilon_{c_d'}^{d'}\right) &= \mathbb{1}\left\{y_{c_d'}^{d'} \in \left(y_{c_d''-1}^{d}, y_{c_d''}^{d}\right)\right\}\left(y_{c_d'}^{d'} - y_{c_d''-1}^{d}\right)\\
C_{c_d''}\left(\Upsilon_{c_d'}^{d'}\right) &= A_{c_d''}\left(\Upsilon_{c_d'}^{d'}\right) + B_{c_d''}\left(\Upsilon_{c_d'}^{d'}\right),
\end{aligned}$$

for $d = 1,\ldots,D$, and $|\Upsilon'_{\boldsymbol{c}'} \cap \Upsilon_{\boldsymbol{c}''}| = \prod_{d=1}^{D} C_{c_d''}\left(\Upsilon_{c_d'}^{d'}\right)$. If we then use the fact that $\Upsilon' = \Upsilon + \boldsymbol{u}$, $y_{c_d}^{d} = y_{1_d}^{d} + (c_d - 1)\delta_d$, $y_{c_d}^{d} - y_{c_d-1}^{d} = \delta_d$ and $u_d \in (-\frac{\delta}{2}, \frac{\delta}{2})$, we obtain

$$\begin{aligned}
A_{c_d''}\left(\Upsilon_{c_d'}^{d'}\right) =& \mathbb{1}\left\{y_{c_d'}^{d'} - y_{c_d''}^{d} + u_d - \delta_d \in (-\delta_d, 0)\right\}(y_{c_d''}^{d} - y_{c_d'}^{d} - u_d + \delta_d)\\
=& \mathbb{1}\left\{(b_d' - c_d'')\delta_d + u_d \in (0,\delta_d)\right\}((c_d'' - c_d' + 1)\delta_d - u_d)\\
=& \mathbb{1}\left\{c_d' = c_d''\right\}\mathbb{1}\left\{u_d \in \left(0, \frac{\delta_d}{2}\right)\right\}(\delta_d - u_d) + \mathbb{1}\left\{c_d' = c_d'' + 1\right\}\mathbb{1}\left\{u_d \in \left(-\frac{\delta_d}{2}, 0\right)\right\}(-u_d)\\
=& (\delta_d - u_d)\mathbb{1}\left\{c_d' = c_d''\right\}\mathbb{1}\left\{u_d \in \left(0, \frac{\delta_d}{2}\right)\right\} - u_d\mathbb{1}\left\{c_d' = c_d'' + 1\right\}\mathbb{1}\left\{u_d \in \left(-\frac{\delta_d}{2}, 0\right)\right\},\\
B_{c_d''}\left(\Upsilon_{c_d'}^{d'}\right) =& (\delta_d + u_d)\mathbb{1}\{c_d' = c_d''\}\mathbb{1}\left\{u_d \in \left(-\frac{\delta_d}{2}, 0\right)\right\} + u_d\mathbb{1}\{c_d' = c_d'' - 1\}\mathbb{1}\left\{u_d \in \left(0, \frac{\delta_d}{2}\right)\right\},\\
C_{c_d''}\left(\Upsilon_{c_d'}^{d'}\right) =& \mathbb{1}\{c_d' = c_d''\}\left((\delta_d - u_d)\mathbb{1}\left\{u_d \in \left(0, \frac{\delta_d}{2}\right)\right\} + (\delta_d + u_d)\mathbb{1}\left\{u_d \in \left(-\frac{\delta_d}{2}, 0\right)\right\}\right)\\
&- u_d\left(\mathbb{1}\{c_d' = c_d'' + 1\}\mathbb{1}\left\{u_d \in \left(-\frac{\delta_d}{2}, 0\right)\right\} - \mathbb{1}\{c_d' = c_d'' - 1\}\mathbb{1}\left\{u_d \in \left(0, \frac{\delta_d}{2}\right)\right\}\right).
\end{aligned}$$

Now rewrite $\mathbb{1}\{x \in \Upsilon'_{\boldsymbol{c}'}\}$ such that

$$\begin{aligned}
\mathbb{1}\{x \in \Upsilon'_{\boldsymbol{c}'}\} &= \prod_{d=1}^{D} \mathbb{1}\{x_{[d]} \in \Upsilon_{c_d'}^{d'}\} = \prod_{d=1}^{D} \mathbb{1}\left\{x_{[d]} \in \left(y_{c_d'-1}^{d'}, y_{c_d'}^{d'}\right)\right\}\\
&= \prod_{d=1}^{D} \mathbb{1}\left\{x_{[d]} \in \left(y_{c_d'}^{d} - \delta_d + u_d, y_{c_d'}^{d} + u_d\right)\right\}\\
&= \prod_{d=1}^{D} \mathbb{1}\left\{u_d \in \left(x_{[d]} - y_{c_d'}^{d}, x_{[d]} - y_{c_d'}^{d} + \delta_d\right)\right\}.
\end{aligned}$$

Let

$$
\begin{aligned}
A_{1c'_d} &= \left(0, \frac{\delta_d}{2}\right) \cap \left(x_{[d]} - y^d_{c'_d}, x_{[d]} - y^d_{c'_d} + \delta_d\right) \\
A_{2c'_d} &= \left(-\frac{\delta_d}{2}, 0\right) \cap \left(x_{[d]} - y^d_{c'_d}, x_{[d]} - y^d_{c'_d} + \delta_d\right) \\
|A_{1c'_d}| &= \left(x_{[d]} - y^d_{c'_d-1}\right) \mathbb{1}\left\{x_{[d]} \in \left(y^d_{c'_d-1}, y^d_{c'_d} - \frac{\delta_d}{2}\right)\right\} + \frac{\delta_d}{2}\mathbb{1}\left\{x_{[d]} \in \left(y^d_{c'_d} - \frac{\delta_d}{2}, y^d_{c'_d}\right)\right\} \\
&\quad + \left(\frac{\delta_d}{2} - x_{[d]} + y^d_{c'_d}\right) \mathbb{1}\left\{x_{[d]} \in \left(y^d_{c'_d}, y^d_{c'_d} + \frac{\delta_d}{2}\right)\right\} \\
|A_{2c'_d}| &= \frac{\delta_d}{2}\mathbb{1}\left\{x_{[d]} \in \left(y^d_{c'_d-1}, y^d_{c'_d} - \frac{\delta_d}{2}\right)\right\} + \left(y^d_{c'_d} - x_{[d]}\right) \mathbb{1}\left\{x_{[d]} \in \left(y^d_{c'_d} - \frac{\delta_d}{2}, y^d_{c'_d}\right)\right\} \\
&\quad + \left(\frac{\delta_d}{2} - y^d_{c'_d-1} + x_{[d]}\right) \mathbb{1}\left\{x_{[d]} \in \left(y^d_{c'_d-1} - \frac{\delta_d}{2}, y^d_{c'_d-1}\right)\right\},
\end{aligned}
$$

and

$$
\begin{aligned}
M_{1c'_d} &= \left(\frac{x_{[d]} - y^d_{c'_d-1}}{2}\right) \mathbb{1}\left\{x_{[d]} \in \left(y^d_{c'_d-1}, y^d_{c'_d} - \frac{\delta_d}{2}\right)\right\} + \frac{\delta_d}{4}\mathbb{1}\left\{x_{[d]} \in \left(y^d_{c'_d} - \frac{\delta_d}{2}, y^d_{c'_d}\right)\right\} \\
&\quad + \left(\frac{x_{[d]} - y^d_{c'_d}}{2} + \frac{\delta_d}{4}\right) \mathbb{1}\left\{x_{[d]} \in \left(y^d_{c'_d}, y^d_{c'_d} + \frac{\delta_d}{2}\right)\right\} \\
M_{2c'_d} &= -\frac{\delta_d}{4}\mathbb{1}\left\{x_{[d]} \in \left(y^d_{c'_d-1}, y^d_{c'_d} - \frac{\delta_d}{2}\right)\right\} + \mathbb{1}\left\{x_{[d]} \in \left(y^d_{c'_d} - \frac{\delta_d}{2}, y^d_{c'_d}\right)\right\} \left(\frac{x_{[d]} - y^d_{c'_d}}{2}\right) \\
&\quad + \mathbb{1}\left\{x_{[d]} \in \left(y^d_{c'_d-1} - \frac{\delta_d}{2}, y^d_{c'_d-1}\right)\right\} \left(\frac{x_{[d]} - y^d_{c'_d-1}}{2} - \frac{\delta_d}{4}\right).
\end{aligned}
$$

Then

$$
\begin{aligned}
&\int_{\mathcal{D}_{\Upsilon'_{c'}}} \mathbb{1}\{x \in \Upsilon'_{\boldsymbol{c}'}\} |\Upsilon'_{\boldsymbol{c}'} \cap \Upsilon_{\boldsymbol{c}''}| \mathrm{d}\Upsilon'_{\boldsymbol{c}'} \\
&\qquad = \int_{-\frac{\delta}{2}}^{\frac{\delta}{2}} \mathbb{1}\{x \in \Upsilon'_{\boldsymbol{c}'}\} \prod_{d=1}^{D} C_{c''_d}\left(\Upsilon^{d'}_{c'_d}\right) \mathrm{d}\boldsymbol{u} \\
&\qquad = \prod_{d=1}^{D} \int_{-\frac{\delta_d}{2}}^{\frac{\delta_d}{2}} \mathbb{1}\left\{x_{[d]} \in \Upsilon'_{c'_d}\right\} C_{c''_d}\left(\Upsilon^{d'}_{c'_d}\right) \mathrm{d}u_d \\
&\qquad = \prod_{d=1}^{D} \left(\mathbb{1}\{c'_d = c''_d\} \int_{A_{1c'_d}} (\delta_d - u_d)\mathrm{d}u_d + \mathbb{1}\{c'_d = c''_d\} \int_{A_{2c'_d}} (\delta_d + u_d)\mathrm{d}u_d \right. \\
&\qquad\qquad \left. -\mathbb{1}\{c'_d = c''_d + 1\} \int_{A_{2c'_d}} u_d \mathrm{d}u_d + \mathbb{1}\{c'_d = c''_d - 1\} \int_{A_{1c'_d}} u_d \mathrm{d}u_d \right) \\
&\qquad = \prod_{d=1}^{D} \left(\mathbb{1}\{c'_d = c''_d\} \left(\delta_d \left(|A_{1c'_d}| + |A_{2c'_d}|\right) - M_{1c'_d}|A_{1c'_d}| + M_{2c'_d}|A_{2c'_d}|\right) \right. \\
&\qquad\qquad \left. -\mathbb{1}\{c'_d = c''_d + 1\} M_{2c'_d}|A_{2c'_d}| + \mathbb{1}\{c'_d = c''_d - 1\} M_{1c'_d}|A_{1c'_d}| \right),
\end{aligned}
$$

and as a result,

$$\begin{aligned}
\phi(x) &= \int_{\mathcal{D}(\Lambda)} f_{X|\boldsymbol{S}=\boldsymbol{s}'}(x) H_x(\lambda) \mathrm{d}\lambda \\
&= \int_{\mathcal{D}(\Upsilon')} \sum_{\boldsymbol{c}'=\mathbf{1}}^{\boldsymbol{C}} \mathbb{1}\{x \in s'_{\boldsymbol{c}'}\} \frac{\mathbb{P}\left(X \in s'_{\boldsymbol{c}'} | \boldsymbol{s}\right)}{|\Upsilon'_{\boldsymbol{c}'}|} H_x\left(\Upsilon'_{\boldsymbol{c}'}\right) \mathrm{d}\Upsilon' \\
&= \sum_{\boldsymbol{c}'=\mathbf{1}}^{\boldsymbol{C}} \int_{\mathcal{D}(\Upsilon')} \frac{\mathbb{1}\left\{x \in \Upsilon'_{\boldsymbol{c}'}\right\}}{|\Upsilon'_{\boldsymbol{c}'}|} \sum_{\boldsymbol{c}''=\mathbf{1}}^{\boldsymbol{C}} \mathbb{P}\left(X \in \Upsilon'_{\boldsymbol{c}'} | X \in \Upsilon_{\boldsymbol{c}''}, \boldsymbol{s}\right) \mathbb{P}\left(X \in \Upsilon_{\boldsymbol{c}''} | \boldsymbol{s}\right) H_x(\Upsilon'_{\boldsymbol{c}'}) \, \mathrm{d}\Upsilon'_{\boldsymbol{c}'} \\
&= \sum_{\boldsymbol{c}'=\mathbf{1}}^{\boldsymbol{C}} \sum_{\boldsymbol{c}''=\mathbf{1}}^{\boldsymbol{C}} \frac{n_{\boldsymbol{c}''}}{N} \int_{\mathcal{D}_{\Upsilon'_{\boldsymbol{c}'}}} \frac{\mathbb{1}\left\{x \in \Upsilon'_{\boldsymbol{c}'}\right\}}{|\Upsilon'_{\boldsymbol{c}'}|} \mathbb{P}\left(X \in \Upsilon'_{\boldsymbol{c}'} | X \in \Upsilon_{\boldsymbol{c}''}, \boldsymbol{s}\right) H\left(\Upsilon'_{\boldsymbol{c}'}\right) \mathrm{d}\Upsilon'_{\boldsymbol{c}'} \\
&= \frac{1}{\prod_{d=1}^{D} \delta_d^3} \sum_{\boldsymbol{c}'=\mathbf{1}}^{\boldsymbol{C}} \sum_{\boldsymbol{c}''=\boldsymbol{c}'-\mathbf{1}}^{\boldsymbol{c}'+\mathbf{1}} \frac{n_{\boldsymbol{c}''}}{N} \int_{\mathcal{D}_{\Upsilon'_{\boldsymbol{c}'}}} \mathbb{1}\left\{x \in \Upsilon'_{\boldsymbol{c}'}\right\} |\Upsilon'_{\boldsymbol{c}'} \cap \Upsilon_{\boldsymbol{c}''}| \mathrm{d}\Upsilon'_{\boldsymbol{c}'} \\
&= \frac{1}{\prod_{d=1}^{D} \delta_d^3} \sum_{\boldsymbol{c}'=\mathbf{1}}^{\boldsymbol{C}} \sum_{\boldsymbol{c}''=\boldsymbol{c}'-\mathbf{1}}^{\boldsymbol{c}'+\mathbf{1}} \frac{n_{\boldsymbol{c}''}}{N} \prod_{d=1}^{D} \Big(\mathbb{1}\{c'_d = c''_d\} \Big(\delta_d \left(|A_{1c'_d}| + |A_{2c'_d}| \right) - M_{1c'_d} |A_{1c'_d}| + M_{2c'_d} |A_{2c'_d}| \Big) \\
&\qquad\qquad - \mathbb{1}\{c'_d = c''_d + 1\} M_{2c'_d} |A_{2c'_d}| + \mathbb{1}\{c'_d = c''_d - 1\} M_{1c'_d} |A_{1c'_d}| \Big) .
\end{aligned}$$

As a consequence we also have an expression for $\phi_{\boldsymbol{c}}$ and setting

$$\begin{aligned}
J_{\boldsymbol{c}',\boldsymbol{c}''}(x) =& \mathbb{1}\left\{c'_d = c''_d\right\} \left(\delta_d \left(|A_{1c'_d}| + |A_{2c'_d}| \right) - M_{1c'_d} |A_{1c'_d}| + M_{2c'_d} |A_{2c'_d}| \right) \\
& - \mathbb{1}\left\{c'_d = c''_d + 1\right\} M_{2c'_d} |A_{2c'_d}| + \mathbb{1}\left\{c'_d = c''_d - 1\right\} M_{1c'_d} |A_{1c'_d}|
\end{aligned}$$

then yields the estimates shown in Section 5.4.2. The following integrations can then be used to compute $P_H^{s_c}$:

$$\begin{aligned}
\int_{\Upsilon^d_{c_d}} \left(|A_{1c'_d}| + |A_{2c'_d}| \right) \mathrm{d}x_{[d]} &= \frac{\delta_d^2}{4} \left(3\mathbb{1}\{c'_d = c_d\} + \frac{1}{2}\mathbb{1}\left\{c'_d = c_d - 1\right\} + \frac{1}{2}\mathbb{1}\left\{c'_d = c_d + 1\right\} \right), \\
\int_{\Upsilon^d_{c_d}} M_{1c'_d} |A_{1c'_d}| \mathrm{d}x_{[d]} &= \frac{\delta_d^3}{12} \left(\mathbb{1}\left\{c'_d = c_d\right\} + \frac{1}{2}\mathbb{1}\left\{c'_d = c_d - 1\right\} \right), \\
\int_{\Upsilon^d_{c_d}} M_{2c'_d} |A_{2c'_d}| \mathrm{d}x_{[d]} &= -\frac{\delta_d^3}{12} \left(\mathbb{1}\left\{c'_d = c_d\right\} + \frac{1}{2}\mathbb{1}\left\{c'_d = c_d + 1\right\} \right).
\end{aligned}$$

B.1.4 Additional graphical information

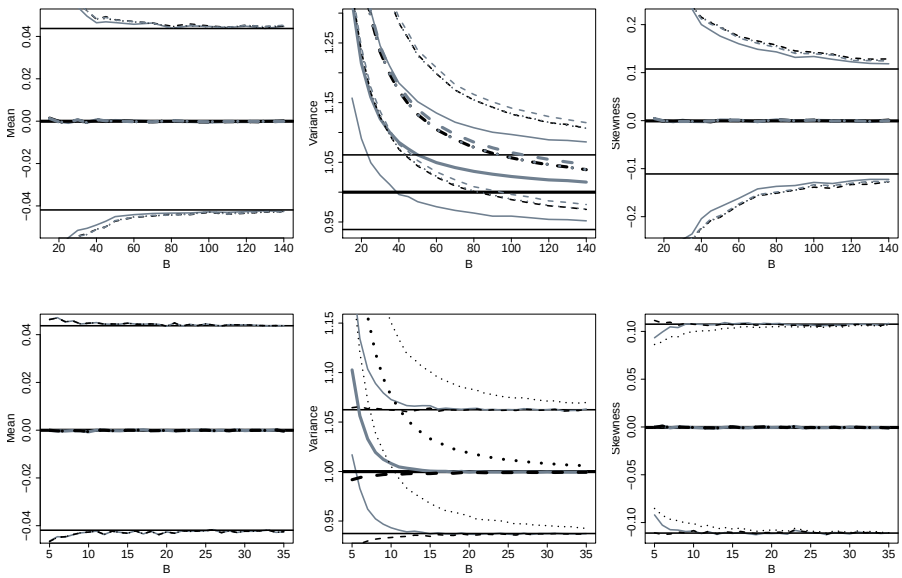

Figure B.1: Estimates and 95% confidence intervals of the mean (left), variance (centre) and skewness (right) as function of C the number of symbols when the aggregates take the form of intervals (top) or histograms (bottom). Classical estimates are given by black horizontal lines, SIU estimates by dotted black lines, SDD estimates by solid grey lines and rounded analysis with Sheppard's correction by dashed black lines. Estimates obtained from the estimated histogram (5.13) are given in the top row respectively by dashed and dotted grey lines for the SIU and SDD approaches. Original data was simulated from the standard normal distribution.

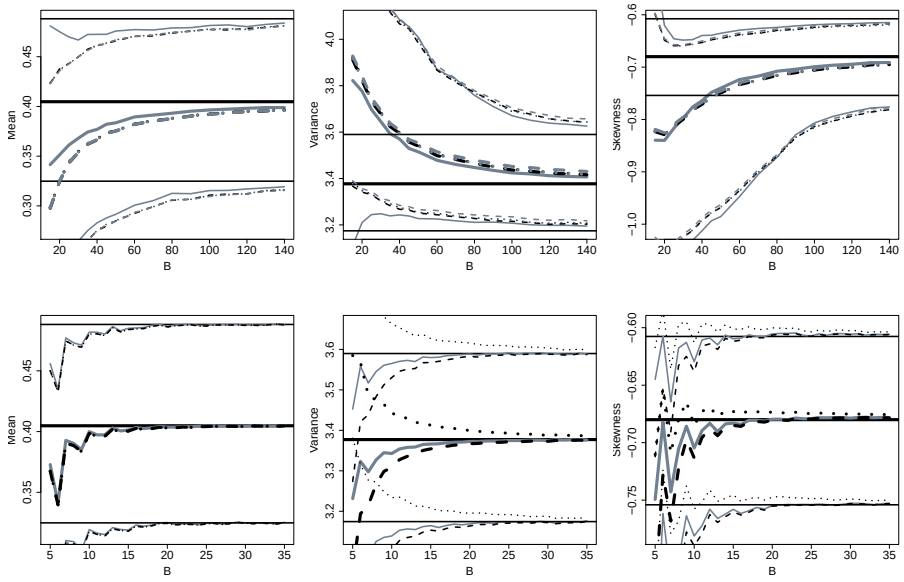

Figure B.2: Estimates and 95% confidence intervals of the mean (left), variance (centre) and skewness (right) as function of C the number of symbols when the aggregates take the form of intervals (top) or histograms (bottom). Classical estimates are given by black horizontal lines, SIU estimates by dotted black lines, SDD estimates by solid grey lines and rounded analysis with Sheppard's correction by dashed black lines. Estimates obtained from the estimated histogram (5.13) are given in the top row respectively by dashed and dotted grey lines for the SIU and SDD approaches. Original data was simulated from the mixture of skew-normal distributions.

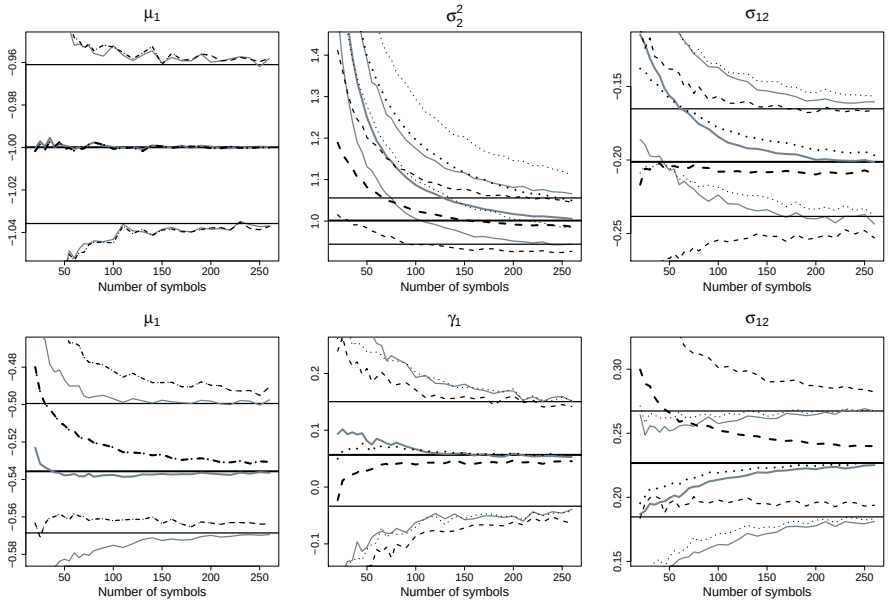

Figure B.3: Estimates and 95% confidence intervals for some of the statistics of interest as a function of C the number of rectangles. Classical estimates are given by black horizontal lines, SIU estimates by dotted black lines and SDD estimates by solid grey lines. Original data were simulated from a normal distribution with correlation $\rho = -0.2$ (top row) and skew-normal distribution with $\alpha = (1, 0.5)$.

Appendix C

Chapter 6 Supporting information

C.1 Complete results for the synthetic examples

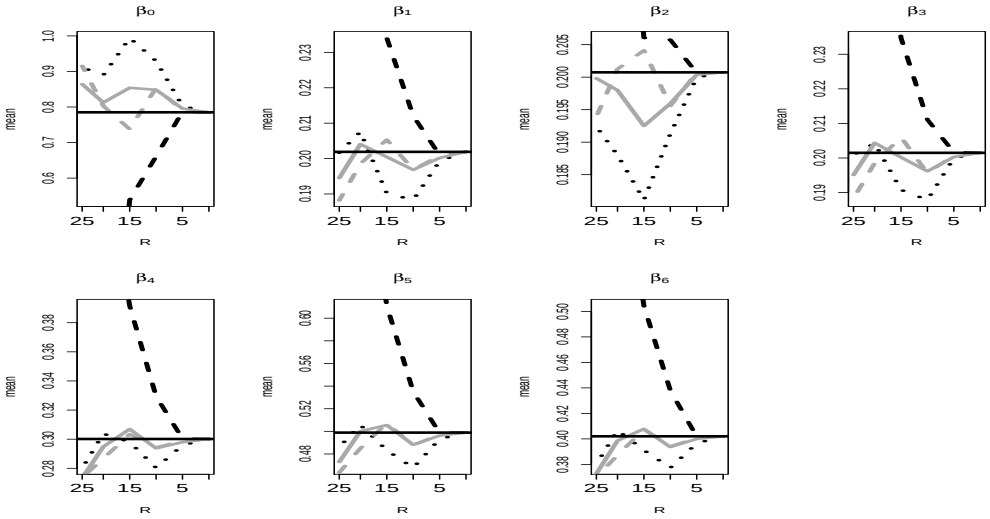

Figure C.1: Mean estimates for each parameter in the poisson synthetic analysis. Classic results are shown in solid black, classical analysis of the rounded data in dashed black, EE results with no additional information in dotted black (6.15), EE results with additional totals information utilised in solid grey (6.16) and EE results with additional proportions information in dashed grey (6.17).

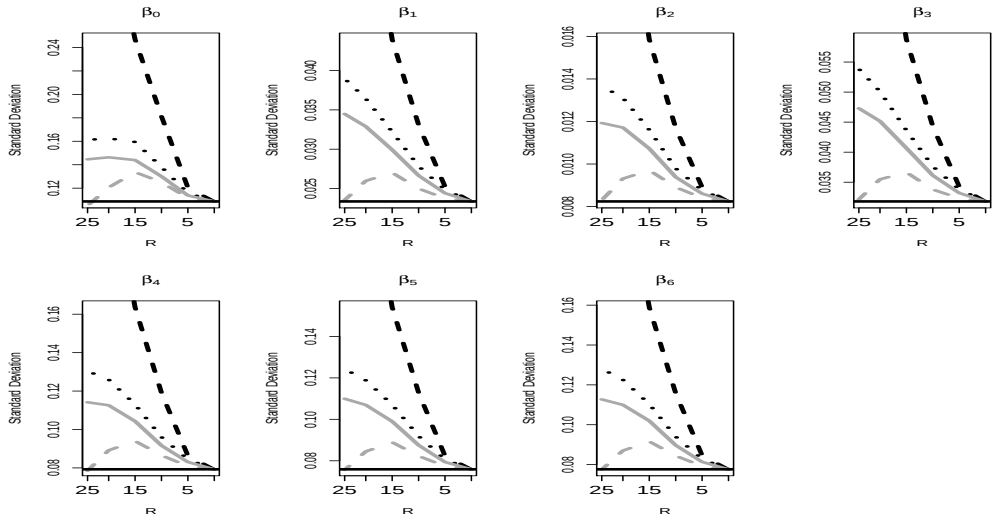

Figure C.2: Mean estimated standard deviations for each parameter in the poisson synthetic analysis. Classic results are shown in solid black, classical analysis of the rounded data in dashed black, EE results with no additional information in dotted black (6.15), EE results with additional totals information utilised in solid grey (6.16) and EE results with additional proportions information in dashed grey (6.17).

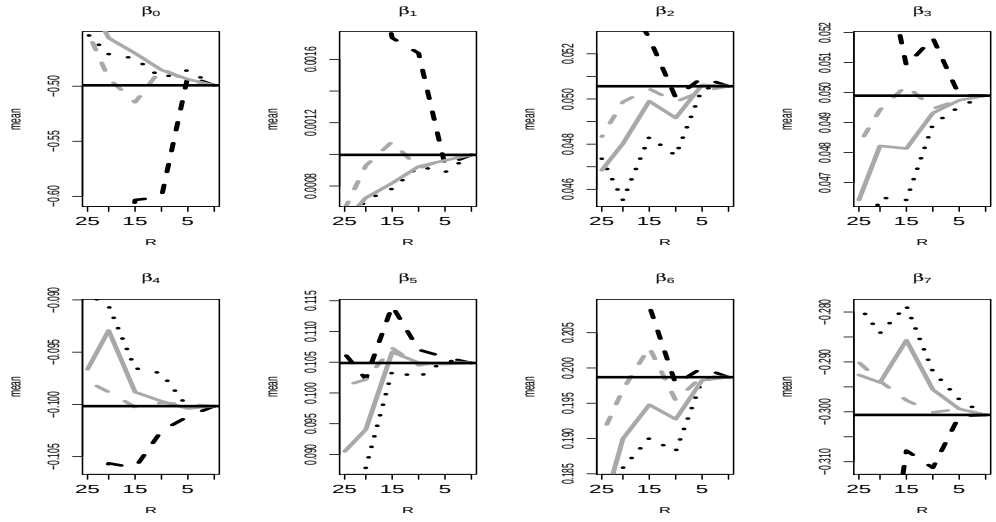

Figure C.3: Mean estimates each parameter in the binomial regression synthetic analysis. Classic results are shown in solid black, classical analysis of the rounded data in dashed black, EE results with no additional information in dotted black (6.15), EE results with additional totals information utilised in solid grey (6.16) and EE results with additional proportions information in dashed grey (6.17).

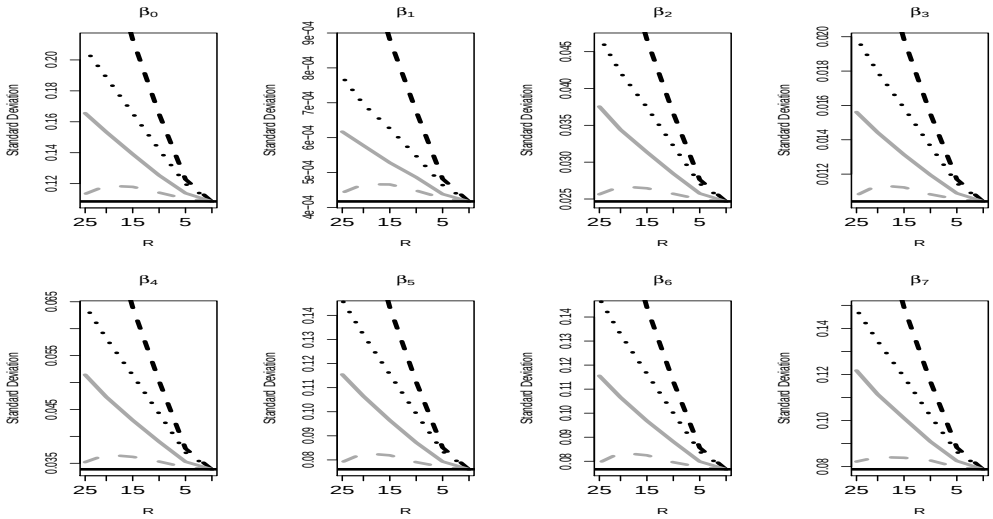

Figure C.4: Mean estimated standard deviations for each parameter in the binomial regression synthetic analysis. Classic results are shown in solid black, classical analysis of the rounded data in dashed black, EE results with no additional information in dotted black (6.15), EE results with additional totals information utilised in solid grey (6.16) and EE results with additional proportions information in dashed grey (6.17).

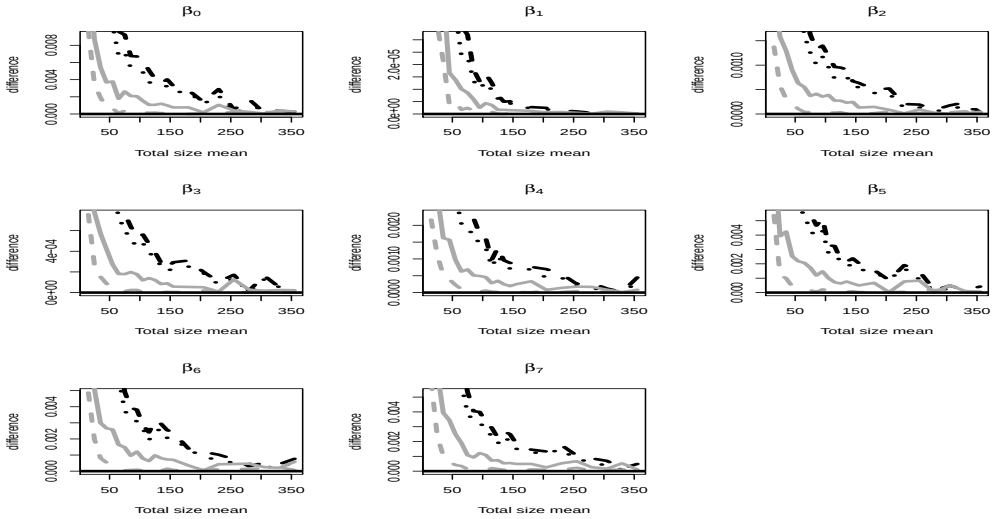

Figure C.5: Difference between the estimated variances of the complete classical analysis and various models. Classic results are shown in solid black, classical analysis of the rounded data in dashed black, EE results with no additional information in dotted black (6.15), EE results with additional totals information utilised in solid grey (6.16) and EE results with additional proportions information in dashed grey (6.17).

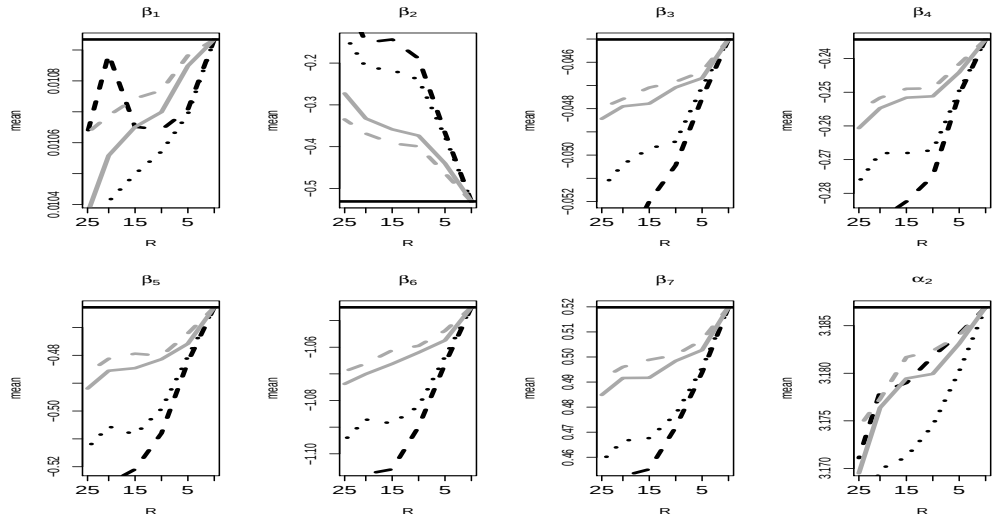

Figure C.6: Mean estimates for each parameter in the ordinal logistic regression synthetic analysis. Classic results are shown in solid black, classical analysis of the rounded data in dashed black, EE results with no additional information in dotted black (6.15), EE results with additional totals information utilised in solid grey (6.16) and EE results with additional proportions information in dashed grey (6.17).

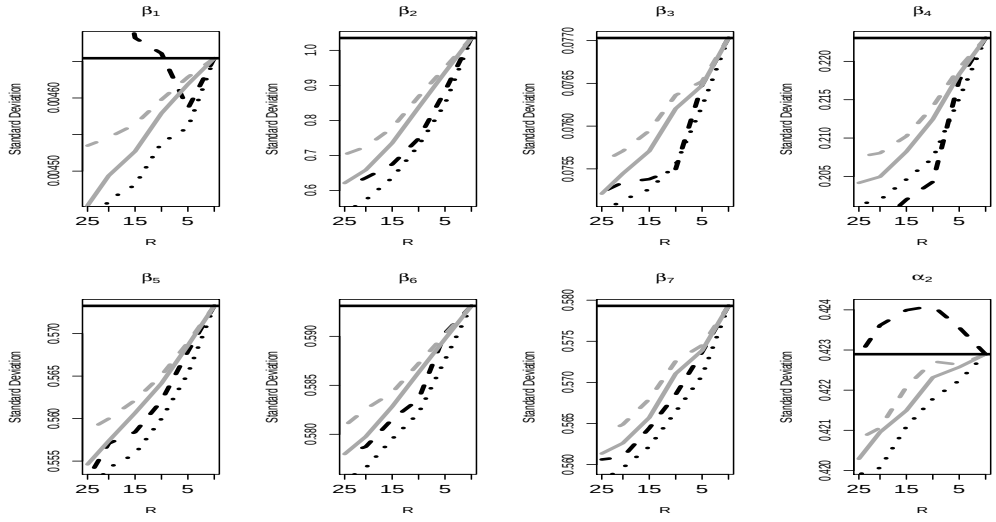

Figure C.7: Mean estimated variances for each parameter in the ordinal logistic regression synthetic analysis. Classic results are shown in solid black, classical analysis of the rounded data in dashed black, EE results with no additional information in dotted black (6.15), EE results with additional totals information utilised in solid grey (6.16) and EE results with additional proportions information in dashed grey (6.17).

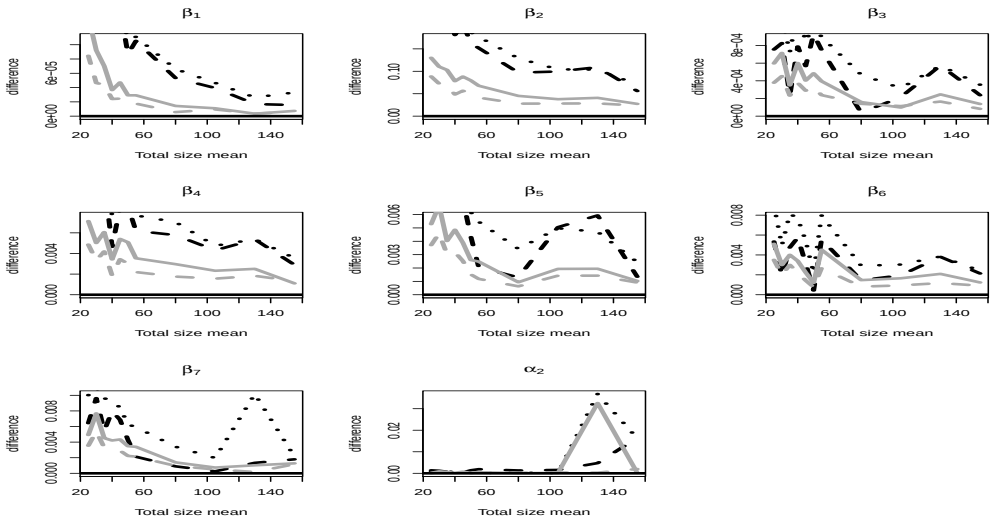

Figure C.8: Difference between the estimated variances of the complete classical analysis and various models for each parameter in the ordinal logistic regression synthetic analysis. Classic results are shown in solid black, classical analysis of the rounded data in dashed black, EE results with no additional information in dotted black (6.15), EE results with additional totals information utilised in solid grey (6.16) and EE results with additional proportions information in dashed grey (6.17).

List of Figures

List of Tables

Bibliography

Albert, A. and J. A. Anderson (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika 71*(1).

Armstrong, B. (1985). Measurement error in the generalized linear model. *Communication in Statistics - Simulation and Computation 14*(3), 529–544.

Baldi, P., P. Sadowski, and D. Whiteson (2014). Searching for exotic particles in high-energy physics with deep learning. *Nature Communications 5*.

Bardenet, R., A. Doucet, and C. Holmes (2014). Towards scaling up mcmc: an adaptive subsampling approach. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Bardenet, R., A. Doucet, and C. Holmes (2017). On markov chain monte carlo methods for tall data. *Journal of Machine Learning Research 18*(1), 1–43.

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

Beranger, B., S. A. Padoan, and S. A. Sisson (2017). Models for extremal dependence derived from skew-symmetric families. *Scandinavian Journal of Statistics 44*(1), 21–45. 10.1111/sjos.12240.

Beranger, B., A. Stephenson, and S. A. Sisson (2019). High-dimensional inference using the extremal skew-t process. *https://arxiv.org/abs/1907.10187*.

Bertrand, P. and F. Goupil (2000). *Descriptive statistics for symbolic data. In: Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer.

Bhowmik, A., J. Ghosh, and O. Koyejo (2016). Generalized linear models for aggregated data. *Arxiv preprint*.

Bierkens, J., P. Fearnhead, and G. Roberts (2019). The zig-zag process and super-efficient sampling for bayesian analysis of big data. *Annals of Statistics 47*(3), 1288–1320.

Billard, L. (2003). Symbolic data analysis: Definitions and examples. Technical report, 62 pages.

Billard, L. (2007). *Selected Contributions in Data Analysis and Classification*, Chapter Dependencies and Variation Components of Symbolic Interval-Valued Data. Springer, Berlin.

Billard, L. (2008). Sample covariance functions for complex quantitative data. *Proceedings of World IASC Conference*, 157–163.

Billard, L. (2011). Brief overview of symbolic data and analytic issues. *Statistical Analysis and Data Mining 4*(2), 149–156.

Billard, L. and E. Diday (2000). *Data analysis, Classificaiton, and Related Methods*, Chapter Regression Analysis for Interval-Valued Data, pp. 369–374. Springer-Verlag, Berlin.

Billard, L. and E. Diday (2003). From the statistics of data to the statistics of knowledge. *Journal of the American Statistical Association 98*, 470–487.

Billard, L. and E. Diday (2006). *Symbolic data analysis*. Wiley Series in Computational Statistics. John Wiley & Sons, Ltd., Chichester.

Billard, L. and J. Le Rademacher (2012). Principal component analysis for interval data. *WIREs Computational Statistics*.

Blanchet, J. and A. C. Davison (2011). Spatial modeling of extreme snow depth. *The Annals of Applied Statistics 5*(3), 1699–1725.

Blower, G. and J. E. Kelsall (2002). Nonlinear kernel density estimation for binned data: convergence in entropy. *Bernoulli 8*(4), 423–449.

Bock, H.-H. and E. Diday (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.

Bowling, S. R. and M. T. Khasawneh (2009). A logistic approximation to the cumulative normal distribution. *Journal of Industrial Engineering and Management 2*(1), 114–127.

Bravo, F. (2004). Empirical likelihood based inference with applications to some econometric models. *Econometric Theory 20*(2), 231–264.

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

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

Cadez, I. V., P. Smyth, G. J. McLachlan, and C. E. McLaren (2002). Maximum likelihood estimation of mixture densities for binned and truncated multivariate data. *Machine Learning 47*(1), 7–34.

Castruccio, S., R. Huser, and M. G. Genton (2016). High-order composite likelihood inference for max-stable distributions and processes. *Journal of Computational and Graphical Statistics 24*(4).

Cigsar, B. and D. Unal (2019). Comparison of data mining classification algorithms determining the default risk. *Scientific Programming*.

Cox, D. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society*, 215–242.

Cox, D. R. and N. Reid (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika 91*(3), 729–737.

Cramer, J. S. (2007). Robustness of logist analysis: Unobserved heterogeneity and misspecified disturbances. *Oxford Bulletin of Economics and Statistics 69*, 545–555.

Davison, A. C., S. A. Padoan, and M. Ribatet (2012). Statistical modelling of spatial extremes. *Statistical Science 27*, 161–186.

de A Lima Neto, E., G. M. Cordeiro, and F. de Carvalho (2011). Bivariate symbolic regression models for interval- valued variables. *Journal of Statistical Computation and Simulation 81*(11), 1727–1744.

de Haan, L. (1984). A spectral representation for max-stable processes. *Ann. Probab. 12*(4), 1194–1204.

de Haan, L. and A. Ferreira (2006). *Extreme value theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York. An introduction.

de Souza, R. M., F. J. A. Cysneiros, D. C. Queiroz, and R. A. de A. Fagundes (2008). A multi-class logistic regression model for interval data. In *Conference: Systems, Man and Cybernetics,*.

de Souza, R. M. C. R., D. C. F. Queiroz, and F. J. A. Cysneiros (2011). Logistic regression-based pattern classifiers for symbolic interval data. *Pattern Analysis and Applications 14*, 273–282.

Dedduwakumara, D. S. and L. A. Prendergast (2018). Confidence intervals for quantiles from histograms and other grouped data. *Communications in Statistics - Simulation and Computation*.

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

Diday, E. (1989). Introduction a l'approche symbolique en analyse des données. *RAIRO Rech. Opér. 23*(2), 193–236.

Douzal-Chouakria, A., L. Billard, and E. Diday (2011). Principal component analysis for interval-valued observations. *Statistical Analysis and Data Mining 4*(2), 229–246.

Dua, D. and C. Graff (2017). UCI machine learning repository.

Dunteman, G. H. and R. H. Moon-Ho (2006). *An Introduction to Generalized Linear Models*, Volume 1 of *Quantitative Applications in the Social Sciences*. Sage Publications.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics 7*(1), 1–26.

Eichelberger, R. K. and V. S. Sheng (2013). An empirical study of reducing multiclass classification methodologies. In P. P. (Ed.), *Machine Learning and Data Mining in Pattern Recognition*, Volume 7988 of *MLDM 2013. Lecture Notes in Computer Science*, pp. 505–519. Springer, Berlin, Heidelberg.

Elashoff, M. and L. Ryan (2004). An em algorithm for estimating equations. *Journal of Computational and Graphical Statistics 13*(1).

Fahrmeir, L. and H. Kaufmann (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics 13*(1), 342–368.

Faraway, J. J. (2010). *International Encyclopedia of Education*, pp. 178–183. Elsevier Inc.

Fithian, W. and H. B. Hasan (2014). Local case-control sampling: Efficient subsampling in imbalanced data sets. *The Annals of Statistics 42*(5), 1693–1724.

Genton, M. G., Y. Ma, and S. H (2011). On the likelihood function of gaussian max-stable processes. *Biometrika 98*(2), 481–488.

Gioia, F. (2006). Principal component analysis on interval data. *Computational Statistics 12*(2).

Gladence, M. L., M. Anu, and M. Karthi (2015, August). A statistical comparison of logistic regression and different bayes classification methods for machine learning. *Journal of Engineering and Applied Sciences 10*(14).

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

Godambe, V. P. (1990). *Estimating Functions*. Oxford : Clarendon Press ; New York : Oxford University Press.

Gu, C. (1993). Smoothing spline density estimation: A dimensionless automatic algorithm. *Journal of the American Statistical Association 88*(422), 495–504.

Gunawan, D., K. Dang, M. Quiroz, R. Kohn, and M. Tran (2019). Subsampling sequential monte carlo for static bayesian models. *Arxiv preprint*.

Haitovsky, Y. (1983). *Grouped Data*, pp. 527–536. NY: Wiley.

Hall, P. (1982, 399). The influence of rounding errors on some nonparametric estimators of a density and its derivatives. *SIAM Journal on Applied Mathematics 42*(2), 390.

Hall, P. (1996). On the accuracy of binned kernel density estimators. *Journal of Multivariate Analysis 56*, 165–184.

Hastie, T., R. Tibshirani, and J. Friedman (2008). *The Elements of Statistical Learning 2nd edition*. Springer.

Hauser, R. P. and D. Booth (2011). Predicting bankruptcy with robust logistic regression. *Journal of Data Science 9*, 565–584.

Heitjan, D. F. (1989). Inference from grouped continuous data: A review. *Statistical Science 4*(2), 164–183.

Heitjan, D. F. and D. B. Rubin (1991). Ignorability and coarse data. *The Annals of Statistics 19*(4), 2244–2253.

Hosmer, D. W., S. Lemeshow, and R. X. Sturdivant (2013). *Applied Logistic Regression 3rd edition*. John Wiley and Sons.

Huang, W. K., M. L. Stein, D. J. McInerney, S. Sun, and E. J. Moyer (2016). Estimating changes in temperature extremes from millennial scale climate simulations using generalized extreme value (gev) distributions. *Arxiv preprint*.

Hyunjoon, K. and G. Zheng (2010). A logistic regression analysis for predicting bankruptcy in the hospitality industry. *Journal of Hospitality Financial Management 14*(1).

Jeffress, L. A. (1973). The logistic distribution as an approximation to the normal curve. *The Journal of the Acoustical Society of America 53*(1), 1296.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society 81*, 158–171.

Jennrich, R. I. and P. F. Sampson (1976). Newton-raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics 18*(1), 11–17.

Johnson, T. R. (2006). Generalized linear models with ordinally-observed covariates. *British Journal of Mathematical and Statistical Psychology 59*(2), 275–300.

Johnson, T. R. and M. M. Wiest (2014). Generalized linear models with coarsened covariates: A practical bayesian approach. *Psychological Methods 19*(2), 281–299.

Jones, M. C. (1998). The edge frequency polygon. *Biometrika 85*(1), 235–239.

Kiang, M. Y. (2003). A comparative assessment of classification methods. *Decision Support Systems 35*, 441–454.

Kim, H. S. (2004, August). *Topics in ordinal logistic regression and its applications*. Ph. D. thesis, Office of Graduate Studies of Texas A and M University.

Knottnerus, J. A. (1992). Application of logistic regression to the analysis of diagnostic data: exact modeling of a probability tree of multiple binary variables. *Med Decision Making 12*(2), 93–108.

Koo, J. Y. and C. Kooperberg (2000). Logspline density estimation for binned data. *Statistics and Probability Letters 46*, 133–147.

Kosmelj, K. and L. Billard (2014). Symbolic covariance matrix for interval-valued variables and its application to principal component analysis: A case study. *Metodoloski Zvezki 11*(1), 1–20.

Lauro, C. N. and F. Palumbo (2000). Principal component analysis of interval data: a symbolic data analysis approach. *Computational Statistics 15*(1), 73–87.

Le Rademacher, J. (2008). *Principal Component Analysis for Interval-Valued and Histogram-Valued Data and Likelihood Functions and some Maximum Likelihood Estimators for symbolic data.* Ph. D. thesis, The University of Georgia.

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

Le Rademacher, J. and L. Billard (2013). Principal component analysis for histogram-valued data. *Advances in Data Analysis and Classification 11*(2), 327–351.

Lee, W., S. K. Sinha, T. E. Arbuckle, and M. Fisher (2018). Estimation in generalized linear models under censored covariates with an application to mirec data. *Statistics in Medicine 37*(1), 4539–4556.

Lee, Y., S. Yoon, S. Murshed, M.-K. Kim, C. Cho, H.-J. Baek, and J.-S. Park (2013). Spatial modeling of the highest daily maximum temperature in korea via max-stable processes. *Advances in Atmospheric Sciences 30*(6), 160–1620.

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

Lindsay, B. G. (1982). Conditional score functions: some optimality results. *Biometrika 69*, 503–512.

Lindsay, B. G. (1988). Composite likelihood methods. In *Statistical inference from stochastic processes (Ithaca, NY, 1987)*, Volume 80 of *Contemp. Math.*, pp. 221–239. Amer. Math. Soc., Providence, RI.

Lipsitz, S., M. Parzen, S. Natarajan, J. Ibrahim, and G. Fitzmaurice (2004). Generalized linear models with a coarsened covariate. *Journal of the Royal Statistical Society*, 279–292.

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics 9*(2), 407–426.

Liu, Y. Y., M. Yang, M. Ramsay, S. X. Li, and J. W. Coid (2011, April). A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. *Journal of Quantitative Criminology*.

MacLaurin, D. and R. P. Adams (2014). Firefly monte carlo: Exact mcmc with subsets of data. In *Proceedings of the conference on Uncertainty in Artificial Intelligence (UAI)*.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological) 42*(2), 109–142.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. Monographs on Statisics and Applied Probability. Chapman and Hall/CRC Press.

Merlo, J., B. Chaix, H. Ohlsson, A. Beekman, K. Johnell, P. Hjerpe, L. Rstam, and K. Larsen (2006). A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of Epidemiology and Community Health 60*(4), 290–297.

Min, H. (2013). Ordered logit regression modeling of the self-rated health in hawaii, with comparisons to the ols model. *The Journal of Modern Applied Statistical Methods 12*(2), 371–380.

Minnotte, M. C. (1996). The bias-optimized frequency polygon. *Computational Statistics 11*(1).

Minnotte, M. C. (1998). Achieving higher-order convergence rates for density estimation with binned data. *Journal of the American Statistical Association 93*.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society 135*(3), 370–384.

Niwa, T., B. Ying, K. Saito, W. Jin, S. Takada, T. Ueda, and H. Taguchi (2009). Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of escherichia coli proteins. *PNAS 106*(11), 4201–4206.

Nordman, D. J. and S. N. Lahiri (2014). A review of empirical likelihood methods for time series. *Journal of Statistical Planning and Inference 155*, 1–18.

Oliveira, M. R., M. Azeitona, A. Pacheco, and R. Valadas (2018). Population symbolic covariance matrices for interval data. *arXiv:1810.06474*.

Oliver, C. I. (2014). *Fundamentals of Applied Probability and Random Processes*. Elsevier Inc.

Owen, A. B. (1988, June). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika 75*(2), 237–249.

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics 18*(1), 90–120.

Owen, A. B. (2007). Infinitely imbalanced logistic regression. *Journal of Machine Learning Research 8*, 761–773.

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

Pampel, F. C. (2000). *Logistic Regression: A Primer*, Volume 1. Sage Publications.

Parzen, M. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics 1*, 1065–1076.

Pavlopoulos, D., R. Muffels, and J. Vermunt (2010). Wage mobility in europe. a comparative analysis using restricted multinomial logit regression. *Quality and Quantity 44*(1), 115–129.

Pingel, R. (2014). Some approximations of the logistic distribution with application to the covariance matrix of logistic regression. *Statistics and Probability Letters 85*(1), 63–68.

Piyadi, R. D., N. Wei, and A. K. Gupta (2017). Adjusted empirical likelihood for time series models. *The Indian Journal of Statistics 79*(2), 336–360.

Qin, J. (2017). *Biased Sampling, Over-identified Parameter Problems and Beyond*, Chapter Empirical Likelihood with Applications. ICSA Book Series in Statistics. Springer.

Quiroz, M., R. Kohn, M. Villani, and T. Minh-Ngoc (2019). Speeding up mcmc by efficient data subsampling. *Journal of the American Statistical Association 114*(526), 831–843.

QUT (2016). Technical report on classification methods for crop types. Technical report, Queensland University of Technology.

Resnick, S. I. (1987). *Extreme values, regular variation, and point processes*, Volume 4 of *Applied Probability. A Series of the Applied Probability Trust*. New York: Springer-Verlag.

Ribatet, M. (2015). Spatialextremes: Modelling spatial extremes - r package version 2.0-2.

Rizzi, S., M. Thinnggaard, and G. Engholm (2016). Comparison of non-parametric methods for ungrouping coarsely aggregated data. *BMC Medical Research Methodology 16*(59).

Schlather, M. (2002). Models for stationary max-stable random fields. *Extemes 5*(1), 33–44.

Schneeweiss, H., J. Komlos, and A. S. Ahmad (2010). Symmetric and asymmetric rounding: a review and some new results. *Advances in Statistical Analysis 94*(3), 247–271.

Scott, D. W. (1985). Frequency polygons: Theory and application. *Journal of the American Statistical Association 80*(390), 348–354.

Scott, D. W. and S. J. Sheather (1985). Kernel density estimation with binned data. *Communication in Statistics - Theory and Methods 14*(6), 1353–1359.

Sheppard, W. F. (1897). On the calculation of the most probable values of frequency constants, for data arranged according to equidistant division of a scale. *Proceedings of the London Mathematical Society 1*(1).

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

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC Press.

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

Smith, L., R. J. Hydman, and S. Wood (2004). Spline interpolation for demographic variables: the monotonicity problem. *Journal Of Population Research 21*.

Smith, R. L. (1990). Max-stable processes and spatial extemes. *Unpublished manuscript*.

Stocker, T., D. Qin, G. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. Midgley (2013). Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change. Technical report, Intergovernmental Panel on Climate Change.

Tranmer, M. H. and D. Steel (1997). Logistic regression analysis with aggregate data: tackling the ecological fallacy. *Paper presented at the American Statistical Association Conference*.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis 91*, 1–28.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica 21*, 5–42.

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

Wand, M. P. (1997). Data-based choice of histogram bin width. *The American Statistician 51*(1), 59–64.

Wang, B. and W. Wertelecki (2013). Density estimation for data with rounding errors. *Computational Statistics and Data Analysis 65*, 4–12.

Wang, C. Y. and M. S. Pepe (2000). Expected estimating equations to accomodate covariate measurement error. *Journal of the Royal Statistical Society 62*, 509–524.

Wang, C. Y., E. C. Yijian Huang, and M. K. J. (2008). Expected estimating equations for missing data, measurement error, and misclassification, with application to longitudinal nonignorable missing data. *Biometrics 64*(1), 85–95.

Wang, H., R. Zhu, and P. Ma (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*.

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

Whitaker, T., B. Beranger, and S. A. Sisson (2019, Aug). Composite likelihood methods for histogram-valued random variables. *arXiv e-prints*, arXiv:1908.11548.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*(1), 1–25.

Willenborg, L. and T. de Waal (1996). *Statistical disclosure control in practise*. Lecture Notes in Statistics. Springer.

Willenborg, L. and T. de Waal (2001). *Elements of Statistical Disclosure Control*. Lecture Notes in Statistics. Springer, New York.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge.

Yongho, J., A. Jeongyoun, and P. Cheolwoo (2015). A nonparametric kernel approach to interval-valued data analysis. *Technometrics 57*(4), 566–575.

Zhang, X. (2017). *Probabilistic modelling of symbolic data and blocking collapsed Gibbs samplers for topic models*. Ph. D. thesis, UNSW Sydney.

Zhang, X., B. Beranger, and S. A. Sisson (2019). Constructing likelihood functions for interval-valued random variables. *Scandinavian Journal of Statistics 47*(1), 1–35.

Zhou, M. (2015). *Empirical Likelihood Method in Survival Analysis*, Volume 1 of *CRC Biostatistics Series*. Chapman and Hall.

Zhou, M. and G. Li (2008). Empirical likelihood analysis of the buckley–james estimator. *Journal of Multivariate Analysis 99*, 649–664.

Zhou, Y., A. T. K. Wan, and X. Wang (2008, September). Estimating equations inference with missing data. *Journal of the American Statistical Association 103*(483), 1187–1199.