Statistics and Computing (2020) 30:1459–1477
https://doi.org/10.1007/s11222-020-09955-5

Composite likelihood methods for histogram-valued random variables

T. Whitaker[1] · B. Beranger[1] · S. A. Sisson[1]

Received: 5 November 2019 / Accepted: 28 May 2020 / Published online: 11 June 2020

Abstract

Symbolic data analysis has been proposed as a technique for summarising large and complex datasets into a much smaller and tractable number of distributions—such as random rectangles or histograms—each describing a portion of the larger dataset. Recent work has developed likelihood-based methods that permit fitting models for the underlying data while only observing the distributional summaries. However, while powerful, when working with random histograms this approach rapidly becomes computationally intractable as the dimension of the underlying data increases. We introduce a composite-likelihood variation of this likelihood-based approach for the analysis of random histograms in K dimensions, through the construction of lower-dimensional marginal histograms. The performance of this approach is examined through simulated and real data analysis of max-stable models for spatial extremes using millions of observed datapoints in more than $K = 100$ dimensions. Large computational savings are available compared to existing model fitting approaches.

Keywords Climate models · Composite likelihoods · Random histograms · Spatial extremes · Symbolic data analysis

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 in current statistical practise, 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), Bayesian hierarchical modelling (Lin et al. 2017) and logistic regression (Whitaker et al. 2019). Likelihood-based methods for distributional data were introduced by Le Rademacher and Billard (2011) and Brito and Silva (2012) for direct modelling at the level of the distributional summary.

More recently, Zhang et al. (2020) 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 stan-

✉ T. Whitaker
t.whitaker@unsw.edu.au

B. Beranger
B.beranger@unsw.edu.au

S. A. Sisson
Scott.sisson@unsw.edu.au

[1] UNSW Data Science Hub and School of Mathematics and Statistics, University of New South Wales, Sydney 2052, Australia

dard 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. The approach of Zhang et al. (2020) and Beranger et al. (2018), and the one taken here, differs from the standard usage of SDA in that it is primarily focused on fitting models for the micro-data, given the observed symbols, rather than on fitting models for the symbols themselves.

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 (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 process models to very large temporal datasets.

This article is structured as follows: In Sect. 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 Sect. 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 Sect. 4 we explore the performance of the histogram-based composite likelihood function through simulation studies using max-stable processes, and in Sect. 5 we analyse real and future-simulated datasets comprising daily maxima temperature data from 105 locations across Australia. We conclude with a Discussion.

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.

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. 2020). 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

$$L(s;\theta,\phi) \propto \int_{D_{\boldsymbol{X}}} f_{S|\boldsymbol{X}=\boldsymbol{x}}(s|\boldsymbol{x},\phi) g_{\boldsymbol{X}}(\boldsymbol{x};\theta) d\boldsymbol{x}. \qquad (1)$$

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 (1) then becomes

$$L(\boldsymbol{s};\theta) \propto \frac{N!}{s_{\boldsymbol{1}}! \ldots s_{\boldsymbol{B}}!} \prod_{\boldsymbol{b}=\boldsymbol{1}}^{\boldsymbol{B}} P_{\boldsymbol{b}}(\theta)^{s_{\boldsymbol{b}}}, \qquad (2)$$

where $\boldsymbol{s} = (s_{\boldsymbol{1}}, \ldots, s_{\boldsymbol{B}})$ is the observed value of $\boldsymbol{S}$, and where $P_{\boldsymbol{b}}(\theta) = \int_{\Upsilon_{\boldsymbol{b}}} g_X(z;\theta) dz$ 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 (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 Sect. 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 $\boldsymbol{b}^{\boldsymbol{i}}$ 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 $\boldsymbol{b}^{\boldsymbol{i}}$, we may construct an I-dimensional random *marginal* histogram $\boldsymbol{S}^{\boldsymbol{i}} = (S^{\boldsymbol{i}}_{\boldsymbol{1}^i}, \ldots, S^{\boldsymbol{i}}_{\boldsymbol{B}^i})$ as the associated vector of random counts from the first bin $\boldsymbol{1}^{\boldsymbol{i}} = (1, \ldots, 1)$ to the last bin $\boldsymbol{B}^{\boldsymbol{i}} = (B^{i_1}, \ldots, B^{i_I})$. The vector $\boldsymbol{S}^{\boldsymbol{i}}$ has length $B^{i_1} \times \ldots \times B^{i_I}$ and satisfies $\sum_{\boldsymbol{b}^i} S^{\boldsymbol{i}}_{\boldsymbol{b}^i} = N$.

Note that we can write $\boldsymbol{S}^{\boldsymbol{i}}_{\boldsymbol{b}^i} = \sum_{\tilde{\boldsymbol{b}}:\tilde{\boldsymbol{b}}^i=\boldsymbol{b}^i} \boldsymbol{S}_{\tilde{\boldsymbol{b}}}$ 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, $\boldsymbol{S}^{\boldsymbol{i}}$ is truly a marginal histogram of $\boldsymbol{S}$ in the usual sense of the term.

Similarly to (2), the resulting symbolic likelihood function for the marginal histogram $\boldsymbol{S}^{\boldsymbol{i}}$ is then given by

$$L(\boldsymbol{S}^{\boldsymbol{i}};\theta) \propto \frac{N!}{s^{\boldsymbol{i}}_{\boldsymbol{1}^i}! \cdots s^{\boldsymbol{i}}_{\boldsymbol{B}^i}!} \prod_{\boldsymbol{b}^i=\boldsymbol{1}^i}^{\boldsymbol{B}^i} P_{\boldsymbol{b}^i}(\theta)^{s^{\boldsymbol{i}}_{\boldsymbol{b}^i}}, \qquad (3)$$

where $\boldsymbol{s}^{\boldsymbol{i}} = (s^{\boldsymbol{i}}_{\boldsymbol{1}^i}, \ldots, s^{\boldsymbol{i}}_{\boldsymbol{B}^i})$ denotes the observed value of $\boldsymbol{S}^{\boldsymbol{i}}$ and $P_{\boldsymbol{b}^i}(\theta) = \int_{\Upsilon_{\boldsymbol{b}^i}} g^{\boldsymbol{i}}_{\boldsymbol{X}^i}(z^{\boldsymbol{i}};\theta) dz^{\boldsymbol{i}}$ is the probability of observing a datapoint within the I-dimensional marginal bin $\Upsilon_{\boldsymbol{b}^i}$ under the marginal model

$$g^{\boldsymbol{i}}_{\boldsymbol{X}^{\boldsymbol{i}}}(\boldsymbol{x}^{\boldsymbol{i}};\theta) = \int g_{\boldsymbol{X}}(z;\theta)dz^{-\boldsymbol{i}},$$

where $z^{-\boldsymbol{i}}$ is the vector of elements of z that are not in $z^{\boldsymbol{i}}$. In the case where $I = \{1,\dots,K\}$ then (3) is equal to (2).

Following similar arguments to Beranger et al. (2018), the symbolic likelihood $L(\boldsymbol{S}^{\boldsymbol{i}};\theta)$ approaches the equivalent classical data likelihood $L(\boldsymbol{X}^{\boldsymbol{i}};\theta) = g^{\boldsymbol{i}}_{\boldsymbol{X}^{\boldsymbol{i}}}(\boldsymbol{X}^{\boldsymbol{i}};\theta)$ 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 $k = 1,\dots,K$, 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

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

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 $x^{\boldsymbol{i}}$ being contained in exactly one bin. For the symbolic likelihood (3), this means that empty bins ($\boldsymbol{s}^{\boldsymbol{i}}_{\boldsymbol{b}^{\boldsymbol{i}}} = 0$) will not contribute to the likelihood, and the N non-empty bins ($\boldsymbol{s}^{\boldsymbol{i}}_{\boldsymbol{b}^{\boldsymbol{i}}} = 1$) will contribute the term $g^{\boldsymbol{i}}_{\boldsymbol{X}^{\boldsymbol{i}}}(\boldsymbol{x}^{\boldsymbol{i}};\theta) = g_{X^{\boldsymbol{i}}}(x^{\boldsymbol{i}};\theta)$, which is the equivalent term contributed to the classical likelihood function $L(\boldsymbol{X}^{\boldsymbol{i}};\theta)$.

As a result, this means that taking more bins will allow $L(\boldsymbol{S}^{\boldsymbol{i}};\theta)$, taken as an approximation to $L(\boldsymbol{X}^{\boldsymbol{i}};\theta)$, 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 $L(\boldsymbol{X}^{\boldsymbol{i}};\theta) = \prod_{k=1}^N g_{X^{\boldsymbol{i}}}(x^{\boldsymbol{i}}_k;\theta)$ 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

$$\ell(\boldsymbol{S}^{\boldsymbol{i}_2};\theta) \propto \sum_{b_{i_1}=1}^{B^{i_1}} \sum_{b_{i_2}=1}^{B^{i_2}} s^{\boldsymbol{i}_2}_{(b_{i_1},b_{i_2})} \log P_{(b_{i_1},b_{i_2})}(\theta). \tag{4}$$

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

$$\ell(\boldsymbol{S}^{\boldsymbol{i}_3};\theta) \propto \sum_{b_{i_1}=1}^{B^{i_1}} \sum_{b_{i_2}=1}^{B^{i_2}} \sum_{b_{i_3}=1}^{B^{i_3}} s^{\boldsymbol{i}_3}_{(b_{i_1},b_{i_2},b_{i_3})} \log P_{(b_{i_1},b_{i_2},b_{i_3})}(\theta). \tag{5}$$

Clearly the number of terms in the full symbolic likelihood (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, $k = 1,\dots,K$, 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

$$P_{\boldsymbol{b}}(\theta) = \int_{\Upsilon_{\boldsymbol{b}}} g_{\boldsymbol{X}}(z;\theta)dz$$

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 (4) and (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.

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 Sect. 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 Sect. 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

Sect. 2.1, and defining $\mathcal{I}_j = \{\boldsymbol{i} : \boldsymbol{i} \subseteq \{1, \dots, K\}, |\boldsymbol{i}| = j\}$ to be the set of all j-dimensional subsets of $\{1, \dots, K\}$, the j-wise composite likelihood function can be written as

$$L_{CL}^{(j)}(\boldsymbol{x}; \theta) \propto \prod_{\boldsymbol{i} \in \mathcal{I}_j} g_{X^{\boldsymbol{i}}}^{\boldsymbol{i}}(\boldsymbol{x}^{\boldsymbol{i}}; \theta), \tag{6}$$

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 (4) and (5), when $j = 2$ the pairwise composite log-likelihood function, $\ell_{CL}^{(2)}$, is given by

$$\ell_{CL}^{(2)}(\boldsymbol{x}; \theta) \propto \sum_{i_1=1}^{K-1} \sum_{i_2=i_1+1}^{K} \log g_{X^{i_1}, X^{i_2}}(\boldsymbol{x}^{i_1}, \boldsymbol{x}^{i_2}; \theta), \tag{7}$$

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

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

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 $\mathcal{I} = \{(1, \dots, K)\}$, 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 composite MLE is used in the place of the standard MLE (Varin et al. 2011).

The number of terms in the $j-$wise composite likelihood function (6) increases exponentially with an increasing number of dimensions K. Padoan et al. (2010) empirically demonstrated that for a max-stable process model with K spatial dimensions (which we use here in Sects. 4 and 5), the trace of the asymptotic covariance matrix of the estimate is minimised if most pairs are excluded from the likelihood, leaving only pairs within a low taper distance and gains in both efficiency and computation. Sang and Genton (2014) consider the use of tapering to reduce this computational burden, giving each term in the likelihood (6) a weight 1 for all sets of pairs or triples located closer than a specified taper distance, and 0 otherwise. The optimal value for the taper distance is obtained via the minimisation of two different characteristics of the covariance matrix; the trace and the determinant. The computational cost of obtaining the optimal taper distance can be minimised by using a subsample of the data.

Bevilacqua et al. (2012) propose two approaches to estimate covariance functions for space and space-time data. The first bases the weights on the distance between the pairs/triples, with a maximal value of 1, and 0 for all locations further apart than a given taper distance. The second method chooses weights by minimising the asymptotic covariance matrix of the model parameters. Both methods are empirically shown to significantly outperform the equal-weights approach in terms of efficiency and computational burden. Li and Sang (2018) derive optimal weights for each tuple in the $j-$wise likelihood of a Gaussian process by choosing them to obtain the optimal estimating equations for the model. Tapering and the assumption of a block diagonal form for the sets of locations are then used to reduce computational costs.

Such weighted composite likelihood methods reduce the computational burden associated with increasing dimension K, but can still be computationally infeasible for datasets with large numbers of observations N. In the following section, we introduce a histogram-based approach to reduce the computational burden associated with the composite likelihood analysis of large datasets. While we do not pursue it here, the tapering and related methods described above can be implemented in this setting to further improve computational efficiency.

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.

3.1 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 $t = 1, \ldots, T$, 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 $g_X(\boldsymbol{x};\theta) = \prod_{i=1}^N g_X(x_i;\theta)$ 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). We can now define the j-wise symbolic composite likelihood for all j-dimensional marginal histograms $\boldsymbol{S}_t^{\boldsymbol{i}}$ of $\boldsymbol{S}_t$, $i \in \mathcal{I}_j$, $t = 1, \ldots, T$ as follows.

Proposition 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

$$L^{(j)}_{SCL}(\boldsymbol{S}_{1:T};\theta) = \prod_{t=1}^{T}\prod_{\boldsymbol{i}\in\mathcal{I}_j} L(\boldsymbol{S}_t^{\boldsymbol{i}};\theta), \tag{8}$$

where $L(\boldsymbol{S}_t^{\boldsymbol{i}};\theta)$ is defined in (3). Defining the maximum j-wise symbolic composite likelihood estimator as $\hat{\theta}^{(j)}_{SCL} = \arg\max_\theta L^{(j)}_{SCL}(\boldsymbol{S}_{1:T};\theta)$, following standard composite likelihood construction arguments (Lindsay 1988) we have

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

as $T \to \infty$ 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

$$\hat{H}(\hat{\theta}^{(j)}_{SCL}) = -\sum_{t=1}^{T}\sum_{\boldsymbol{i}\in\mathcal{I}_j} \nabla^2 \ell(\boldsymbol{S}_t^{\boldsymbol{i}};\theta) \tag{9}$$

$$= -\sum_{t=1}^{T}\sum_{\boldsymbol{i}\in\mathcal{I}_j}\sum_{\boldsymbol{b}^{\boldsymbol{i}}=\boldsymbol{1}^{\boldsymbol{i}}}^{\boldsymbol{B}^{\boldsymbol{i}}} s^{\boldsymbol{i}}_{t,\boldsymbol{b}^{\boldsymbol{i}}} \nabla^2 \log P_{t,\boldsymbol{b}^{\boldsymbol{i}}}(\hat{\theta}^{(j)}_{SCL}) \tag{10}$$

$$\hat{J}(\hat{\theta}^{(j)}_{SCL}) = \sum_{t=1}^{T}\left(\sum_{\boldsymbol{i}\in\mathcal{I}_j}\nabla\ell(\boldsymbol{S}_t^{\boldsymbol{i}};\theta)\right)\left(\sum_{\boldsymbol{i}\in\mathcal{I}_j}\nabla\ell(\boldsymbol{S}_t^{\boldsymbol{i}};\theta)\right)^{\top}$$

$$= \sum_{t=1}^{T}\left(\sum_{\boldsymbol{i}\in\mathcal{I}_j}\sum_{\boldsymbol{b}^{\boldsymbol{i}}=\boldsymbol{1}^{\boldsymbol{i}}}^{\boldsymbol{B}^{\boldsymbol{i}}} s^{\boldsymbol{i}}_{t,\boldsymbol{b}^{\boldsymbol{i}}}\nabla\log P_{t,\boldsymbol{b}^{\boldsymbol{i}}}(\hat{\theta}^{(j)}_{SCL})\right)\times \tag{11}$$

$$\left(\sum_{\boldsymbol{i}\in\mathcal{I}_j}\sum_{\boldsymbol{b}^{\boldsymbol{i}}=\boldsymbol{1}^{\boldsymbol{i}}}^{\boldsymbol{B}^{\boldsymbol{i}}} s^{\boldsymbol{i}}_{t,\boldsymbol{b}^{\boldsymbol{i}}}\nabla\log P_{t,\boldsymbol{b}^{\boldsymbol{i}}}(\hat{\theta}^{(j)}_{SCL})\right)^{\top}, \tag{12}$$

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

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

$$\ell^{(2)}_{SCL}(\boldsymbol{S}_{1:T};\theta) = \sum_{t=1}^{T}\sum_{i_1=1}^{K-1}\sum_{i_2=i_1+1}^{K}\ell(\boldsymbol{S}_t^{(i_1,i_2)};\theta) \tag{13}$$

where $\ell(\boldsymbol{S}_t^{(i_1,i_2)};\theta)$ is given by (4), and the triple-wise ($j = 3$) symbolic composite log-likelihood function is given by

$$\ell^{(3)}_{SCL}(\boldsymbol{S}_{1:T};\theta) = \sum_{t=1}^{T}\sum_{i_1=1}^{K-2}\sum_{i_2=i_1+1}^{K-1}\sum_{i_3=i_2+1}^{K}\ell(\boldsymbol{S}_t^{(i_1,i_2,i_3)};\theta) \tag{14}$$

where $\ell(\boldsymbol{S}_t^{(i_1,i_2,i_3)};\theta)$ is given by (5).

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 (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 (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 (13). 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 $\boldsymbol{b}^{\boldsymbol{i}} = (b_{i_1}, b_{i_2})$ 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

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

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

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

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

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

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

Consequently, the pairwise symbolic composite log likelihood is given as

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

Now, letting the number of bins $B^{i_1}, B^{i_2} \to \infty$ such that each bin's volume → 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 (in the number of bins) consistent approximation of the standard composite likelihood, which itself provides an asymptotically (in the number of datapoints) consistent estimator of the true parameter. That is, as the number of bins increases (and the bin volume decreases), then the symbolic composite MLE approaches the standard composite MLE $\hat{\theta}^{(j)}_{SCL} \to \hat{\theta}^{(j)}_{CL}$. Further, if in addition the amount of data N increases, then as $\hat{\theta}^{(j)}_{CL} \to \theta_0$ (where θ_0 is the true parameter value) then $\hat{\theta}^{(j)}_{SCL} \to \theta_0$. The symbolic composite approximation can be arbitrarily close to the classical composite equivalent (although 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 (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

$$\sum_{t=1}^{T} s^{\boldsymbol{i}}_{t,\boldsymbol{b}^{\boldsymbol{i}}} = s^{\boldsymbol{i}}_{\boldsymbol{b}^{\boldsymbol{i}}}, \ \forall \boldsymbol{b}, \boldsymbol{i}, \tag{15}$$

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 (15).) However, $T = 1$ will not be the optimal choice if interest is also in computing $\mathrm{Var}(\hat{\theta}^{(j)}_{SCL})$—see the following section.

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 $k = 1, \dots, K$) 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

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

Following the same arguments as in Sect. 3.2 it is straightforward to show that

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

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 (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 (11). 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

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

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

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

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

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

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 (11)

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

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

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

using similar arguments to the above.

Compared to the standard composite likelihood sensitivity matrix $\hat{J}(\hat{\theta}^{(j)}_{CL})$, (16) 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) $t = 1, \dots, T$ 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 (16) 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 Sect. 4.2.5.) 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 $\mathrm{Var}(\hat{\theta}^{(j)}_{SCL})$ 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 Sect. 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 micro-data are available, then $T = N$ histograms (again, with the same bins as for computing the symbolic composite MLE) can be used to determine the lowest possible variance of the MLE. This is viable as the computational cost of evaluating the variance, given the MLE, is only a small fraction of the total computation required to optimise the likelihood (with $T = 1$).

In terms of identifying the number of bins within each histogram, here we follow the strategy of increasing the value of B 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 previous value of B 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. While this approach is slightly inefficient as it requires repeated likelihood optimisation to compute multiple symbolic composite likelihood estimators $\hat{\theta}^{(j)}_{SCL}$ for varying values of B, the simulations in Sect. 4 (e.g. Table 7) 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 histogram-based approach. This simple approach is used in both the simulation studies in Sect. 4, and the real data analysis in Sect. 5.

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.

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

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

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 $K \ge 2$ spatial locations with coordinates $t_k \in \mathcal{T} = \mathbb{R}^d$, $k = 1, \ldots, K$, where $K \le d + 1$. 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 $c^{(j)}(\mathbf{v}) = \left(c^{(j)}_1(\mathbf{v}), \ldots, c^{(j)}_{j-1}(\mathbf{v}), c^{(j)}_{j+1}(\mathbf{v}), \ldots, c^{(j)}_K(\mathbf{v})\right)^\top \in \mathbb{R}^{K-1}$, where, for $k = 1, \ldots, K$, $v_k = v(y_k; \mu, \sigma, \xi)^{-1}$ and $c^{(j)}_k(\mathbf{v}) = (t_j - t_k)^\top \Sigma^{-1}(t_j - t_k)/2 - \log\left(\frac{v_j}{v_k}\right)$. Then, writing $\Sigma^{(j)} = (t_j 1^\top_{K-1} - \tilde{T}_{-j})^\top \Sigma^{-1}(t_j 1^\top_{K-1} - \tilde{T}_{-j})$, 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

$$P(Y_1(t) \le y_1, \ldots, Y_K(t) \le y_K) \quad (16)$$

Table 1 Spatial dependence parameter specifications for the Gaussian max-stable model, following Padoan et al. (2010)

Model	σ_{11}	σ_{12}	σ_{22}
Σ_1	300	0	300
Σ_2	300	150	300
Σ_3	300	150	200
Σ_4	3000	1500	3000
Σ_5	30	15	30

$$= \exp\left\{-\sum_{j=1}^{K}\frac{1}{v_j}\Phi_{K-1}\left(c^{(j)}(\mathbf{v});\, \Sigma^{(j)}\right)\right\}, \tag{17}$$

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 Sect. 5. The complexity of the density function associated with (17) and its computational burden explodes with an increasing number of locations K, and consequently the complete likelihood isn't feasible in the analysis of such datasets. 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 ($\hat{\theta}^{(j)}_{SCL}$ and $\hat{\theta}^{(j)}_{CL}$ 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. $(\mu, \sigma, \xi) = (0, 1, 0)$).

4.2 Comparisons with composite likelihoods

4.2.1 Varying the number of bins, B

We generate $N = 1000$ realisations for $K = 15$ locations and 5 different configurations of the covariance matrix Σ, with true values given in Table 1, which represent a range of dependence 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 1 shows the resulting bivariate histograms for two locations with $\Sigma = \Sigma_3$.

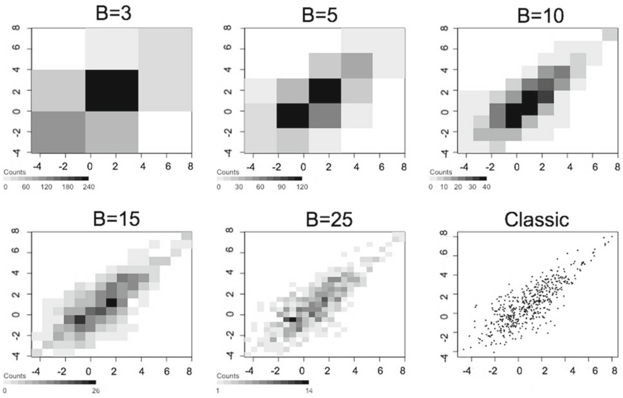

Fig. 1 $B \times B$ bivariate histograms for different values of B for the same classical dataset (bottom right panel) of size $N = 1000$, generated at two spatial locations under the Gaussian max-stable model with $\Sigma = \Sigma_3$ (Table 1)

Tables 2 and 3 report the resulting mean symbolic composite and composite MLEs, $\hat{\theta}^{(2)}_{SCL}$ and $\hat{\theta}^{(2)}_{CL}$, with standard errors in parentheses, based on 1000 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 ($B \approx 10$).

4.2.2 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 4 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

Table 2 Mean (and standard errors) for the dependence parameters 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

Model	B	σ_{11}	σ_{12}	σ_{22}
Σ_1	2	335.5 (585.5)	5.7 (232.2)	317.2 (125.1)
	3	301.0 (34.5)	− 0.1 (16.9)	301.9 (33.5)
	5	299.1 (23.1)	− 0.9 (13.2)	299.9 (24.1)
	10	299.8 (20.2)	− 0.5 (11.1)	300.0 (20.9)
	15	299.8 (18.9)	− 0.3 (10.4)	300.0 (19.5)
	25	299.7 (18.0)	− 0.3 (10.0)	300.2 (18.9)
	Classic	300.76 (17.1)	−0.4 (9.7)	301.02 (18.1)
Σ_2	2	316.59 (149.1)	165.1 (246.8)	332.9 (153.5)
	3	299.6 (35.0)	149.7 (24.9)	300.8 (33.7)
	5	298.9 (23.4)	149.2 (16.7)	299.9 (23.4)
	10	299.3 (20.2)	149.6 (13.9)	300.3 (19.9)
	15	299.4 (19.2)	149.7 (13.2)	300.5 (19.0)
	25	299.7 (18.3)	149.9 (12.5)	300.5 (18.1)
	Classic	300.7 (17.0)	150.4 (11.6)	301.53 (17.0)
Σ_3	2	321.6 (360.0)	162.3 (210.6)	210.8 (131.2)
	3	296.1 (30.6)	147.4 (20.1)	197.9 (19.9)
	5	298.8 (23.3)	149.4 (15.3)	199.6 (15.4)
	10	299.0 (19.3)	149.6 (12.3)	199.7 (12.9)
	15	299.5 (18.7)	149.8 (11.6)	199.8 (12.1)
	25	299.7 (17.8)	150.0 (11.2)	200.0 (11.8)
	Classic	300.7 (16.4)	150.6 (10.2)	200.6 (10.9)
Σ_4	2	3554 (2071)	1848 (1319)	3473 (1839)
	3	2954 (435)	1453 (294)	2952 (405)
	5	3003 (345)	1500 (244)	2996 (337))
	10	3002 (249)	1506 (169)	2997 (239)
	15	2992 (217)	1498 (148)	2988 (211)
	25	2992 (199)	1499 (136)	2991 (200))
	Classic	3002 (190)	1503 (124)	2999 (189)
Σ_5	2	30.97 (3.57)	15.53 (2.81)	30.98 (3.86)
	3	29.83 (2.04)	14.89 (1.58)	29.82 (2.18)
	5	29.86 (1.54)	14.85 (1.17)	29.82 (1.71)
	10	29.93 (1.27)	14.92 (0.95)	29.91 (1.45)
	15	29.96 (1.20)	14.93 (0.91)	29.91 (1.33)
	25	29.97 (1.13)	14.95 (0.86)	29.94 (1.28)
	Classic	30.10 (0.94)	15.06 (0.66)	30.06 (1.03)

Results based on $N = 1000$ observations at $K = 15$ spatial locations and $T = 1$ random histogram

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.

4.2.3 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 5 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 repli-

Table 3 Mean (and standard errors) of the GEV parameters (μ, σ, ξ) 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

Model	B	μ	σ	ξ
Σ_1	2	0.0383 (0.1639)	0.8687 (0.0061)	−0.0194 (0.0301)
	3	0.0812 (0.0550)	0.9195 (0.0342)	0.0182 (0.0210)
	5	0.0067 (0.0295)	0.9666 (0.0285)	0.0136 (0.0194)
	10	−0.0015 (0.0276)	0.9898 (0.0186)	0.0039 (0.0120)
	15	−0.0017 (0.0272)	0.9929 (0.0179)	0.0027 (0.0110)
	25	−0.0016 (0.0272)	0.9954 (0.0179)	0.0013 (0.0102)
	Classic	−0.0019 (0.0262)	0.9986 (0.0173)	0.0007 (0.0084)
Σ_2	2	0.3763 (0.1448)	0.8671 (0.0632)	−0.0163 (0.0284)
	3	0.0755 (0.0439)	0.9258 (0.0284)	0.0151 (0.0192)
	5	0.0077 (0.0280)	0.9705 (0.0266)	0.0114 (0.0182)
	10	0.0002 (0.0267)	0.9912 (0.0182)	0.0023 (0.0118)
	15	−0.0001 (0.0265)	0.9941 (0.0179)	0.0021 (0.0108)
	25	0.0001 (0.0265)	0.9964 (0.0176)	0.0009 (0.0100)
	Classic	−0.0002 (0.0258)	0.9997 (0.0172)	0.0004 (0.0081)
Σ_3	2	0.3596 (0.1310)	0.8671 (0.0586)	−0.0150 (0.0271)
	3	0.0723 (0.0422)	0.9302 (0.0280)	0.0113 (0.0174)
	5	0.0065 (0.0263)	0.9713 (0.0237)	0.0102 (0.0170)
	10	−0.0001 (0.0252)	0.9908 (0.0174)	0.0031 (0.0114)
	15	−0.0009 (0.0249)	0.9942 (0.0170)	0.0021 (0.0105)
	25	−0.0009 (0.0251)	0.9963 (0.0168)	0.0009 (0.0096)
	Classic	−0.0013 (0.0243)	0.9993 (0.0164)	0.0004 (0.0079)
Σ_4	2	0.4337 (0.2211)	0.8691 (0.0847)	−0.0393 (0.0342)
	3	0.0857 (0.0729)	0.9132 (0.0418)	0.0202 (0.0250)
	5	0.0071 (0.0355)	0.9626 (0.0366)	0.0156 (0.0258)
	10	−0.0004 (0.0323)	0.9891 (0.0233)	0.0030 (0.0172)
	15	−0.0009 (0.0318)	0.9930 (0.0224)	0.0009 (0.0147)
	25	−0.0010 (0.0318)	0.9953 (0.0222)	−0.0001 (0.0128)
	Classic	−0.0001 (0.0308)	0.9988 (0.0217)	−0.0025 (0.0113)
Σ_5	2	0.3356 (0.1003)	0.8662 (0.0456)	−0.0002 (0.0093)
	3	0.0633 (0.0246)	0.9452 (0.0184)	0.0032 (0.0099)
	5	0.0071 (0.0157)	0.9821 (0.0140)	0.0021 (0.0076)
	10	0.0012 (0.0149)	0.9928 (0.0111)	0.0009 (0.0046)
	15	0.0004 (0.0146)	0.9952 (0.0108)	0.0007 (0.0038)
	25	0.0001 (0.0145)	0.9970 (0.0106)	0.0003 (0.0031)
	Classic	−0.0004 (0.0144)	0.9997 (0.0104)	0.0000 (0.0004)

Results based on $N = 1000$ observations at $K = 15$ spatial locations and $T = 1$ random histogram

cate 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).

4.2.4 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 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 B_2, B_3. Rows correspond to $B_2^2 \approx B_3^3$ to compare approximately equal numbers of histogram bins

$B_2^2 \vert B_3^2$	σ_{11}		σ_{12}	
	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)
$5^2 \vert 3^3$	300.55 (0.95)	300.23 (2.44)	150.40 (0.66)	150.09 (1.66)
$8^2 \vert 4^3$	300.45 (0.80)	300.21 (1.28)	150.31 (0.54)	150.16 (0.86)
$11^2 \vert 5^3$	300.57 (0.72)	300.42 (0.91)	150.39 (0.46)	150.30 (0.62)

$B_2^2 \vert B_3^2$	σ_{22}		μ	
	Pair	Triple	Pair	Triple
$3^2 \vert 2^3$	200.14 (1.74)	199.68 (5.46)	0.0426 (0.0217)	0.1515 (0.0494)
$5^2 \vert 3^3$	200.26 (0.55)	200.02 (1.50)	0.0016 (0.0025)	0.0411 (0.0209)
$8^2 \vert 4^3$	200.20 (0.50)	200.07 (0.82)	0.0001 (0.0007)	0.0093 (0.0079)
$11^2 \vert 5^3$	200.22 (0.38)	200.19 (0.56)	0.0000 (0.0008)	0.0015 (0.0023)

$B_2^2 \vert B_3^2$	σ		ξ	
	Pair	Triple	Pair	Triple
$3^2 \vert 2^3$	0.9803 (0.0094)	0.9718 (0.0112)	0.0039 (0.0023)	0.0004 (0.0055)
$5^2 \vert 3^3$	0.9978 (0.0033)	0.9807 (0.0092)	0.0008 (0.0013)	0.0037 (0.0023)
$8^2 \vert 4^3$	0.9999 (0.0007)	0.9926 (0.0056)	0.0001 (0.0001)	0.0020 (0.0016)
$11^2 \vert 5^3$	0.9999 (0.0001)	0.9978 (0.0029)	0.0000 (0.0001)	0.0008 (0.0011)

Results based on $N = 10^6$ observations at $K = 10$ spatial locations, $T = 1$ random histogram and $\Sigma = \Sigma_3$

Table 5 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

K	σ_{11}		σ_{12}	
	Pair	Triple	Pair	Triple
3	300.44 (5.80)	299.24 (13.37)	150.30 (2.41)	150.02 (6.75)
5	300.35 (1.53)	299.95 (2.37)	150.28 (1.10)	150.02 (1.99)
10	300.21 (0.88)	299.95 (1.22)	150.15 (0.59)	149.99 (0.83)
15	300.19 (0.71)	299.93 (1.12)	150.12 (0.48)	150.00 (0.73)
20	300.20 (0.78)	299.99 (0.99)	150.14 (0.47)	150.02 (0.70)

K	σ_{22}		μ	
	Pair	Triple	Pair	Triple
3	201.55 (11.12)	200.12 (7.84)	– 0.00003 (0.0011)	0.00727 (0.0102)
5	200.22 (1.00)	199.98 (1.89)	– 0.00002 (0.0010)	0.00671 (0.0088)
10	200.10 (0.53)	199.94 (0.77)	– 0.00006 (0.0009)	0.00595 (0.0068)
15	200.06 (0.46)	200.00 (0.72)	– 0.00004 (0.0001)	0.00553 (0.0054)
20	200.08 (0.44)	199.99(0.69)	– 0.00005 (0.0001)	0.00524 (0.0053)

K	σ		ξ	
	Pair	Triple	Pair	Triple
3	0.9999 (0.0009)	0.9947 (0.0069)	0.00006 (0.00062)	0.00121 (0.00193)
5	0.9999 (0.0008)	0.9950 (0.0064)	0.00008 (0.00059)	0.00111 (0.00187)
10	0.9999 (0.0007)	0.9956 (0.0047)	0.00009 (0.00048)	0.00093 (0.00133)
15	0.9999 (0.0007)	0.9958 (0.0042)	0.00007 (0.00042)	0.00092 (0.00131)
20	0.9999 (0.0007)	0.9961 (0.0039)	0.00006 (0.00048)	0.00080 (0.00121)

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 $\Sigma = \Sigma_3$

Table 6 Mean (and standard errors) of the standard pairwise composite ($\hat{\theta}_{CL}^{(2)}$) and symbolic pairwise composite ($\hat{\theta}_{SCL}^{(2)}$) MLEs from 100 replications of the Gaussian max-stable process model with $B_2 \times B_2$ histograms with $B_2 = 25$

N	σ_{11}		σ_{12}	
	Classic	Pair	Classic	Pair
10^3	299.48 (17.09)	298.11 (17.24)	149.90 (10.37)	148.84 (11.05)
10^4	299.07 (5.76)	298.56 (6.07)	149.65 (3.26)	149.09 (3.63)
10^5	300.56 (1.56)	300.49 (2.07)	150.42 (0.98)	150.32 (1.27)
10^6	–	300.21 (0.61)	–	150.18 (0.45)
10^7	–	300.13 (0.23)	–	150.06 (0.17)

N	σ_{22}		μ	
	Classic	Pair	Classic	Pair
10^3	200.45 (11.05)	200.11 (11.69)	−0.0074 (0.0280)	−0.0077 (0.0286)
10^4	199.92 (3.32)	199.39 (3.70)	−0.0017 (0.0074)	−0.0013 (0.0076)
10^5	200.28 (1.14)	200.18 (1.49)	−0.0002 (0.0021)	−0.0002 (0.0025)
10^6	–	200.14 (0.43)	–	0.0000 (0.0007)
10^7	–	200.02 (0.18)	–	−0.0001 (0.0002)

N	σ		ξ	
	Classic	Pair	Classic	
10^3	0.9972 (0.0169)	0.9964 (0.0170)	0.0016 (0.0115)	0.0024 (0.0123)
10^4	0.9989 (0.0051)	0.9988 (0.0052)	−0.0002 (0.0039)	−0.0002 (0.0040)
10^5	1.0000 (0.0014)	1.0000 (0.0015)	0.0001 (0.0010)	0.0001 (0.0013)
10^6	–	1.0000 (0.0004)	–	0.0000 (0.0004)
10^7	–	1.0000 (0.0001)	–	0.0000 (0.0001)

Results are based on $K = 10$ spatial locations, $T = 1$ random histogram and $\Sigma = \Sigma_3$

Table 6 reports the resulting means and standard errors of $\hat{\theta}_{CL}^{(2)}$ and $\hat{\theta}_{SCL}^{(2)}$ 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 $\hat{\theta}_{CL}^{(2)}$ for $N \geq 10^6$. To explore this in more detail, these simulations were repeated for $K = 20, 50, 100$ spatial locations, and a slightly smaller range of observed data ($N = 1000$ to $500,000$) to provide a better comparison with the standard composite MLEs.

Table 7 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 $\hat{\theta}_{CL}^{(2)}$ becomes largely impractical when either becomes moderately large. Clearly computation of $\hat{\theta}_{SCL}^{(2)}$ would take similar times to those in Table 7 for considerably larger N.

An additional step in computing $\hat{\theta}_{SCL}^{(2)}$ 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.

4.2.5 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

Table 7 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

N	$K = 10$				$K = 20$			
	t_c	t_s	t_{histDR}	t_{histR}	t_c	t_s	t_{histDR}	t_{histR}
1000	71.9	22.5	0.8	0.1	383.4	79.6	1.8	0.4
5000	291.8	19.0	0.8	0.3	1578.2	99.3	2.1	1.0
10, 000	591.7	23.8	0.9	0.5	3125.4	103.2	2.4	1.8
50, 000	2626.8	24.2	1.7	2.1	20,459.4	107.3	4.5	7.6
100, 000	5610.7	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}
1000	7333.9	528.5	9.3	3.0	–	2238.0	78.8	12.0
5000	27,616.5	665.1	10.6	7.7	–	2650.2	81.7	30.9
10, 000	–	696.3	12.4	13.5	–	2356.6	85.8	54.1
50, 000	–	744.8	24.8	59.0	–	2300.6	131.6	237.0
100, 000	–	768.1	41.3	115.7	–	2766.9	188.2	461.8
500, 000	–	802.9	156.1	561.3	–	3111.5	627.1	2243.5

Results are based on $T = 1$ random histogram and $\Sigma = \Sigma_3$

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 Sect. 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 Sect. 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 $N = 1000$ 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 $T = 1000$ 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 8, based on 1000 replicate analyses. This procedure is then repeated 100 times while varying the number of marginal histogram bins (B^2), with the results illustrated in Fig. 2.

From Table 8, 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 $T = N$. Figure 2 illustrates how the mean Godambe standard errors, for fixed T, approach the (square root of the) appropriate diagonal term of the limit (16) of the variability matrix $\hat{J}(\hat{\theta}^{(2)}_{SCL})$, as the number of histogram bins becomes large. As $T \rightarrow N$ 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.

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° × 1.875° grid cells over Australia (Fig. 3). Two different scenarios (RCP4.5 and RCP8.5) are used to

Table 8 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 1000 replicate analyses

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

Results are based on $N = 1000$ observations with $B = 25$ and $\Sigma = \Sigma_3$

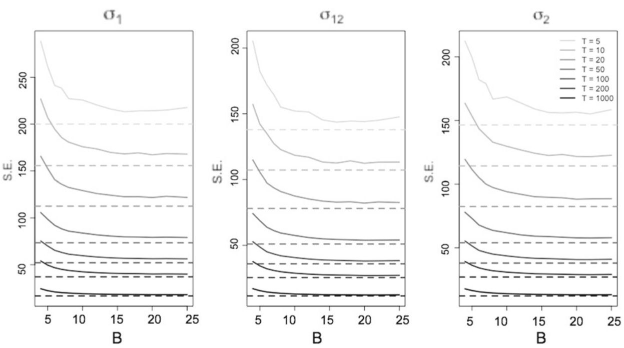

Fig. 2 Godambe standard errors (solid lines) for the dependance parameters $(\sigma_{11}, \sigma_{12}, \sigma_{22})$ of $\hat{\theta}^{(2)}_{SCL}$ for varying number of random histograms T, and number of marginal histogram bins B^2. Dashed horizontal lines denote the appropriate term of the limit (16) of the variability matrix $\hat{J}(\hat{\theta}^{(2)}_{SCL})$. Results are based on $N = 1000$ observations with $\Sigma = \Sigma_3$

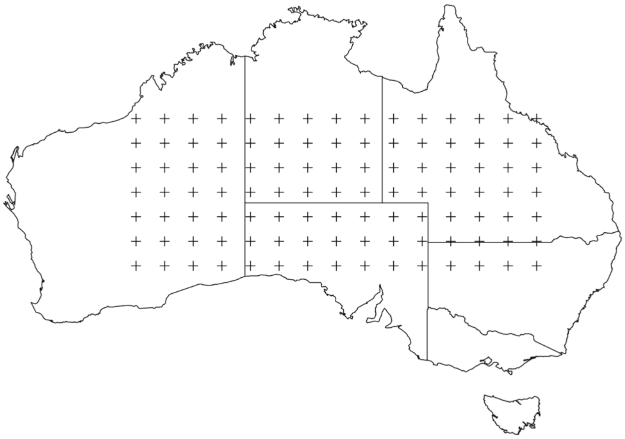

Fig. 3 $K = 105$ spatial locations for the historical and future-simulated temperature data over Australia. Each cross represents the midpoint of a 1.875° × 1.875° box in a spatial grid

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. For this block size, we examined qq-plots of empirical versus theoretical GEV quantiles for each spatial location and each data scenario (historical or future). The fidelity of the GEV limit was very good for both future scenarios, and less good, but reasonable, for the historical data (results not shown).

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

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

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.

For this dataset, the computation times presented in Table 7 (with $K = 100$ and $N = 1000$) indicate that the symbolic composite pairwise approach can be considerably more efficient than the classical composite likelihood function. Further, Table 9 lists the total number of terms in the standard pairwise composite likelihood, $\ell^{(2)}_{CL}(\theta)$, and the symbolic composite likelihood, $\ell^{(2)}_{CL}(\theta)$, 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 9) 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 Sect. 3, the symbolic composite MLE $(\hat{\theta}^{(2)}_{SCL})$ can be computed exactly with $T = 1$ random histogram, and so this optimisation (which evaluates

Table 9 Total number of terms in each pairwise composite likelihood function for $N = 936, 570$ block maxima over $K = 105$ spatial locations

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

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

Table 10 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 $B = 15, 20, 25, 30$

B	σ_{11}	σ_{12}	σ_{22}	ξ
Historical Data				
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				
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				
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)

the target function many times) can be very efficient. In contrast, $T = N$ histograms are required for the best variance estimates (see Table 8), 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).

Table 10 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 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 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.

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

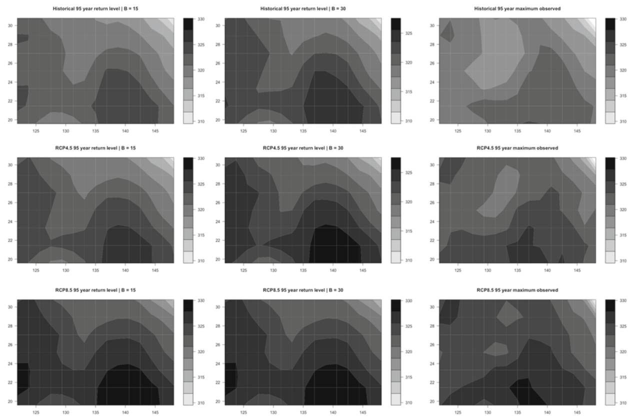
Historical 95 year return level | B = 15
Historical 95 year return level | B = 30
Historical 95 year maximum observed
RCP4.5 95 year return level | B = 15
RCP4.5 95 year return level | B = 30
RCP4.5 95 year maximum observed
RCP8.5 95 year return level | B = 15
RCP8.5 95 year return level | B = 30
RCP8.5 95 year maximum observed
20
22
24
26
28
30
125
130
135
140
145
310
315
320
325
330

Fig. 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)

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 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. (2020) 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 suitable number of bins B to ensure a good approximation of the classical composite likelihood (and MLE), we used a naive approach Sect. 5 in which B was increased until there was only minor improvement in the inference. While this approach is practically viable (in that computational overheads can still be much lower than implementing the standard classical analysis), ideally a method would be constructed to identify a priori the fixed number of bins required to optimise some criterion. This could be e.g. a binwidth selection algorithm, or a loss-function based method etc.

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 at each spatial location, with very little work done to analyse the spatial dependence (Huang et al. 2016). In Sect. 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 Sect. 5. This research is supported by the UNSW Data Science Hub, and the Australian Research Council through the Australian Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS; CE140100049), and the Discovery Project scheme (FT170100079).

References

Beranger, B., Padoan, S.A., Sisson, S.A.: Models for extremal dependence derived from skew-symmetric families. Scand. J. Stat. **44**(1), 21–45 (2017). https://doi.org/10.1111/sjos.12240

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

Beranger, B., Stephenson, A., Sisson, S. A.: High-dimensional inference using the extremal skew-t process. (2019) arXiv:1907.10187

Bertrand, P., Goupil, F.: Descriptive statistics for symbolic data. In: Bock, H. H., Diday, E. (eds.) Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data. Springer (2000)

Bevilacqua, M., Gaetan, C., Mateu, J., Porcu, E.: Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach. J. Am. Stat. Assoc. **107**(497), 268–280 (2012)

Billard, L.: Brief overview of symbolic data and analytic issues. Stat. Anal. Data Min. **4**(2), 149–156 (2011)

Billard, L., Diday, E.: From the statistics of data to the statistics of knowledge. J. Am. Stat. Assoc. **98**, 470–487 (2003)

Billard, L., Diday, E.: Symbolic Data Analysis. Wiley Series in Computational Statistics. Wiley, Chichester (2006)

Blanchet, J., Davison, A.C.: Spatial modeling of extreme snow depth. Ann. Appl. Stat. **5**(3), 1699–1725 (2011)

Bock, H.H., Diday, E. (eds.): Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin (2000)

Brito, P., Silva, A.P.D.: Modelling interval data with normal and skew-normal distributions. J. Appl. Stat. **39**, 3–20 (2012)

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

Castruccio, S., Huser, R., Genton, M.G.: High-order composite likelihood inference for max-stable distributions and processes. J. Comput. Graph. Stat. **24**(4), 1212–1229 (2016)

Davison, A.C., Padoan, S.A., Ribatet, M.: Statistical modelling of spatial extremes. Stat. Sci. **27**, 161–186 (2012)

de Haan, L.: A spectral representation for max-stable processes. Ann. Probab. **12**(4), 1194–1204 (1984)

de Haan, L., Ferreira, A.: Extreme Value Theory. Springer Series in Operations Research and Financial Engineering: An Introduction. Springer, New York (2006)

Dias, S., Brito, P.: Linear regression model with histogram-valued variables. Stat. Anal. Data Min. **8**, 75–113 (2015)

Diday, E.: Introduction a l'approche symbolique en analyse des données. RAIRO Rech. Opér. **23**(2), 193–236 (1989)

Genton, M.G., Ma, Y., Sang, H.: On the likelihood function of Gaussian max-stable processes. Biometrika **98**(2), 481–488 (2011)

Godambe, V.P.: An optimum property of regular maximum likelihood estimation. Ann. Math. Stat. **31**, 1208–1211 (1960)

Huang, W. K., Stein, M. L., McInerney, D. J., Sun, S., Moyer, E. J.: Estimating changes in temperature extremes from millennial scale climate simulations using generalized extreme value (GEV) distributions. arXiv:1512.08775 (2016)

Jenkinson, A.F.: The frequency distribution of the annual maximum (or minimum) values of meteorological elements. Q. J. R. Meteorolog. Soc. **81**, 158–171 (1955)

Kosmelj, K., Billard, L.: Symbolic covariance matrix for interval-valued variables and its application to principal component analysis: a case study. Metodoloski Zvezki **11**(1), 1–20 (2014)

Le Rademacher, J., Billard, L.: Likelihood functions and some maximum likelihood estimators for symbolic data. J. Stat. Plan. Inference **141**, 1593–1602 (2011)

Le Rademacher, J., Billard, L.: Principal component analysis for histogram-valued data. Adv. Data Anal. Classif. **11**(2), 327–351 (2013)

Lee, Y., Yoon, S., Murshed, S., Kim, M.-K., Cho, C., Baek, H.-J., Park, J.-S.: Spatial modeling of the highest daily maximum temperature in korea via max-stable processes. Adv. Atmos. Sci. **30**(6), 160–1620 (2013)

Li, F., Sang, H.: On approximating optimal weighted composite likelihood method for spatial models. J. Rapid Dissem. Stat. Res. **7**(1), (2018)

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

Lindsay, B. G.: Composite likelihood methods. In: Prabhu, N.U. (ed.) Statistical Inference from Stochastic Processes (Ithaca, NY, 1987), Volume 80 of Contemp. Math. pp. 221–239. American Mathematical Society, Providence, RI

Padoan, S.A., Ribatet, M., Sisson, S.: Likelihood-based inference for max-stable processes. J. Am. Stat. Assoc. **105**, 263–277 (2010)

Resnick, S.I.: Extreme Values, Regular Variation, and Point Processes. Volume 4 of Applied Probability. A Series of the Applied Probability Trust. Springer, New York (1987)

Ribatet, M.: Spatialextremes: Modelling spatial extremes - r package version 2.0-2 (2015)

Sang, H., Genton, M.G.: Tapered composite likelihood for spatial max-stable models. Spat. Stat. **8**(1), 86–103 (2014)

Schlather, M.: Models for stationary max-stable random fields. Extemes **5**(1), 33–44 (2002)

Scott, D.W., Sheather, S.J.: Kernel density estimation with binned data. Commun. Stat. Theory Methods **14**(6), 1353–1359 (1985)

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

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

Smith, R. L.: Max-stable processes and spatial extemes. *Unpublished manuscript* (1990)

Stocker, T., Qin, D., Plattner, G., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.: 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 (2013)

Varin, C., Vidoni, P.: A note on composite likelihood inference and model selection. Biometrika **92**, 519–528 (2005)

Varin, C., Reid, N., Firth, D.: An overview of composite likelihood methods. Stat. Sin. **21**, 5–42 (2011)

Wand, M.P.: Data-based choice of histogram bin width. Am. Stat. **51**(1), 59–64 (1997)

Wang, X., Zhang, Z., Li, S.: Set-valued and interval-valued stationary time series. J. Multivar. Anal. **145**, 208–223 (2016)

Whitaker, T., Beranger, B., Sisson, S. A.: Logistic regression models for aggregated data. arXiv:1912.03805 (2019)

Zhang, X.: Probabilistic modelling of symbolic data and blocking collapsed Gibbs samplers for topic models. Ph. D. thesis, UNSW Sydney (2017)

Zhang, X., Beranger, B., Sisson, S.A.: Constructing likelihood functions for interval-valued random variables. Scand. J. Stat. **47**(1), 1–35 (2020)