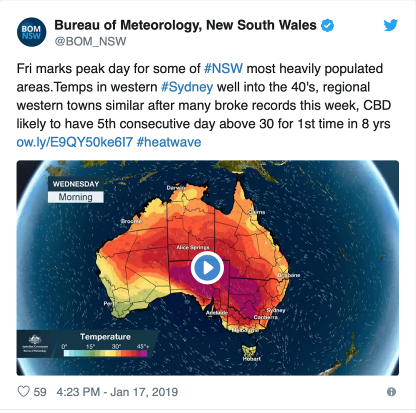
High-dimensional inference for max-stable processes

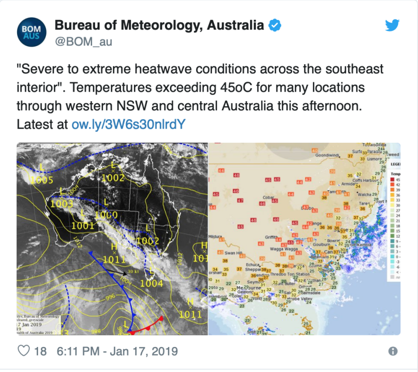
Boris Beranger
Alec Stephenson, Scott A. Sisson & Tom Whitaker

UNSW & ACEMS

Extremes Webinar, University of Lisbon, October 15th

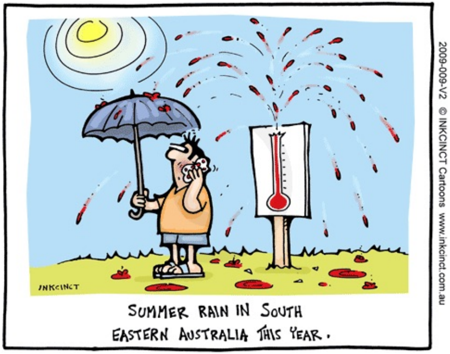
Motivation





- What is the maximum value that a process (Temperature) is expected to reach over some region of interest (NSW/Australia) within the next 20, 50 years?

Talk Outline



1. **Max-stable processes**
   - Construction
   - Inference and limitations
2. Solution #1: Combining methodologies
   - Cdf approximations
   - Simulated examples
3. Solution #2: aggregating data
   - Methodology
   - Simulated and real examples
4. Discussion

Max-stable processes

- Max-stable processes are a useful tool to analyse **spatial extremes**.

$X_1, X_2, \ldots,$ be i.i.d replicates of $X(s), s \in \mathcal{S} \subset \mathbb{R}^k$,

$$\left\{\max_{i=1,\ldots,n} \frac{X_i(s) - b_n(s)}{a_n(s)}\right\}_{s \in \mathcal{S}} \xrightarrow{d} \{Y(s)\}_{s \in \mathcal{S}}$$

for some continuous functions $a_n(s) > 0$ and $b_n(s)$.

- $Y_0(s)$ be the limiting process with unit Fréchet margins

$$P\{Y_0(s_j) \leq y(s_j), j \in I\} = \exp\{-V_0(y(s_j), j \in I)\}$$

where

$$V_0\{y(s_j), j \in I\} = d \int_{\mathbb{W}_d} \max_{j \in I} \left(\frac{w_j}{y(s_j)}\right) \mathrm{d}H(\mathrm{w}).$$

Max-stable processes (2)

Spectral representation (e.g. Schlather, 2002)

Let $\{R_i\}_{i\geq 1}$ be the points of a Poisson process on $\mathbb{R}^+$ with intensity $\xi r^{-(\xi+1)}$, $\xi > 0$.

$X^+ = \max_s(0, X(s))$, $\mu^+(s) = \mathbb{E}[\{X^+(s)\}^\xi] < \infty$

$X_i^+, i = 1, 2, \ldots$ be i.i.d copies of X^+.

Then

$$Y(s) = \max_{i=1,2,\ldots}\{R_i X_i^+(s)\}/\{\mu^+(s)\}^{1/\xi}, \quad s \in \mathcal{S},$$

is a **max-stable process** with ξ-Fréchet 1-d distributions.

The **exponent function** is

$$V\{y(s_j), j \in I\} = \mathbb{E}\left[\max_{j\in I}\left\{\frac{X^+(s_j)^\xi}{\mu^+(s_j)y(s_j)^\xi}\right\}\right].$$

Max-stable models

1. Smith model (Smith, 1990); Schlather model (Schlather, 2002); Brown-Resnick model (Kabluchko et al., 2009);

2. Extremal-t (Opitz, 2013) $X_i(s)$ are i.i.d. copies of a weakly stationary GP with isotropic correlation function $\rho(h)$;

3. Extremal skew-t (Beranger et al., 2017) $X_i(s)$ are i.i.d. copies of a (non-strictly stationary) skew-Normal process;

The exponent function of the extremal Skew-t model is

$$V\{y(s_j), j \in I\} = \sum_{j=1}^{d} \frac{1}{y(s_j)^{\xi}} \Psi_{d-1}\left[\{q_i, i \in I_j\}^{\top}; \bar{\Sigma}_j, \alpha_j^*, \tau_j^*, \nu + 1\right],$$

where Ψ_{d-1} is a $d-1$-dimensional extended skew-t cdf.

Inference

- Consider some locations $z_1, \ldots, z_d \in \mathcal{S}$

The **full likelihood function** is given by

$$\mathrm{L}(\mathbf{z};\theta) = \exp\{-V(\mathbf{z};\theta)\} \sum_{\Pi \in \mathcal{P}_d} \prod_{k=1}^{|\Pi|} -V_{\pi_k}(\mathbf{z};\theta),$$

where:

$\mathcal{P}_d$: set of all possible partitions Π of $\{1, \ldots, d\}$

Π: has elements π_k

$|\mathcal{P}_d|$: cardinality of $\mathcal{P}_d$ corresponds to the **d-th Bell number**

$V_{\pi_k}(\cdot)$: partial derivatives of $V(\cdot)$ w.r.t π_k.

$\Rightarrow$ INTRACTABLE, even for moderate d.

Inference (2)

Composite likelihood (Padoan et al.,2010):

$$\mathrm{CL}_j(\mathbf{z};\theta) = \prod_{q \in \mathcal{Q}_d^{(j)}} \left(\exp\{-V(\mathbf{z}_q;\theta)\} \times \sum_{\Pi \in \mathcal{P}_q} \prod_{k=1}^{|\Pi|} -V_{\pi_k}(\mathbf{z}_q;\theta) \right)^{w_q},$$

$\mathcal{Q}_d^{(j)}$: set of all possible subset of size j of $\{1,\ldots,d\}$
$\mathbf{z}_q$: j-dimensional subvector of $\mathbf{z} \in \mathbb{R}_+^d$
$\mathcal{P}_q$: set of all possible partitions of q where each partition Π has elements π_k
Π: has elements π_k
$V_{\pi_k}(\cdot)$: partial derivatives of $V(\cdot)$ w.r.t π_k.

$j = 3$: Genton et al. (2011), Huser and Davison. (2013)

Higher-order are more efficient but limited to $d = 13$: Castruccio et al. (2016)

Composite likelihoods properties

Behaviour of composite MLE

$\hat{\theta}_{CL}^{(j)}$ is asymptotically ($N \to \infty$) consistent and distributed as

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

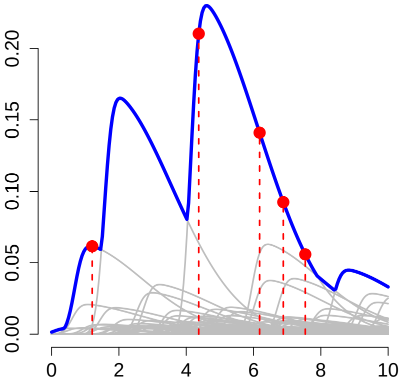
where

- $G^{(j)}(\theta)^{-1} = H^{(j)}(\theta) J^{(j)}(\theta)^{-1} H^{(j)}(\theta)$ is **Godambe** information matrix
- $H^{(j)}(\theta) = -\mathbb{E}(\nabla^2 \ell_{CL}^{(j)}(\theta; \boldsymbol{x}))$ is the **sensitivity** matrix
- $J^{(j)}(\theta) = \mathbb{V}(\nabla \ell_{CL}^{(j)}(\theta; \boldsymbol{x}))$ is the **variability** matrix.

- For standard likelihoods $j = d$ and $H(\theta) = J(\theta)$ and so $G(\theta) = H(\theta) = I(\theta)$ is the Fisher information matrix.

Stephenson & Tawn likelihood

Time occurrences of each block maxima assumed known

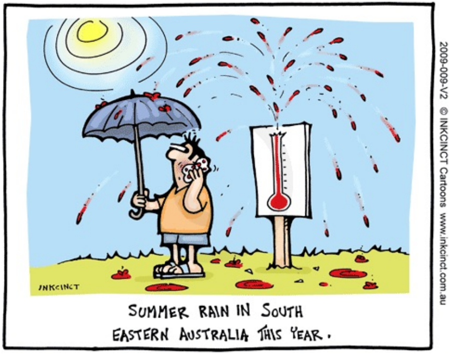


ST likelihood (Stephenson and Tawn,2005):

For each block i given by say $\mathbf{z}^i$, an observed partition Π^i is associated

$$\mathrm{ST}(\mathbf{z};\theta) = \exp\left\{-V(\mathbf{z};\theta)\right\} \times \prod_{k=1}^{|\Pi|} -V_{\pi_k}(\mathbf{z};\theta).$$

Talk Outline



1. **Max-stable processes**
   - Construction
   - Inference and limitations
2. Solution #1: Combining methodologies
   - Cdf approximations
   - Simulated examples
3. Solution #2: aggregating data
   - Methodology
   - Simulated and real examples
4. Discussion

Solution #1: Combining methodologies

Simple idea: Use the knowledge of time occurrences within the composite likelihood framework.

Why would it work?

- Wadsworth (2015): second order bias correction $\Rightarrow$ Requires $n > d(d-1)/2$.
- Huser et al. (2016): both methods can be highly biased in high dimensions.

Bonus: Additional computational improvement

Fast(er) cumulative distribution function evaluations

A **necessity** already highlighted by Wadsworth and Tawn (2014), Castruccio et al. (2016), de Fondeville and Davison (2018).

Skew-t cdf is a function of t cdf $\Rightarrow$ quasi-Monte Carlo approximations

Idea:

- Control the error on the log-scale $\Rightarrow$ **fewer Monte Carlo simulations**
- Evaluations of $\Psi_{d-m}(\cdot)$ in $V_{\pi_k}(z;\theta)$ are relatively more important than those of $\Psi_{d-1}(\cdot)$ in $V(z;\theta)$.
- Set N_{min}: minimum number of simulations
- Set N_{max}: maximum number of simulations

Simulation setup

- $d = 20, 50, 100$ locations on region $\mathcal{S} = [-5, 5] \times [-5, 5]$
- **Extremal skew-*t*** with $\nu = 1$ and $\alpha_i \equiv \alpha(s_i) = \beta_1 s_{i1} + \beta_2 s_{i2}$
- $n = 50$ temporal replicates
- **Power exponential correlation function**

$$\rho(h) = \exp\{-(\|h\|/r)^s\}, \quad r > 0, 0 < s \leq 2$$

  Smoothness $s = 1, 1.5, 1.95$ and range $r = 1.5, 3, 4.5$ (spatial dependence)
- $j = 2, 3, 4, 5, 10, d$
- log-error $= 0.0001$
- 500 replicates, run in parallel using 16 CPUs.

j	$2, 3$	$4, 5$	10	d (Type I)	d (Type II)
$\Psi_{j-m}(\cdot)$	$100, 1000$	$50, 500$	$20, 200$	$50, 500$	$20, 200$
$\Psi_{j-1}(\cdot)$	$10, 100$	$5, 50$	$2, 20$	$5, 50$	$2, 20$

Table: Number of quasi-Monte Carlo simulations N_{min}, N_{max} to compute each $\Psi_{j-m}(\cdot)$ and $\Psi_{j-1}(\cdot)$ terms in $V_{\pi_k}(\mathbf{z}; \theta)$ for each j-wise composite likelihood.

Approximation of the (full) ST likelihood

Measure: $\mathrm{RMSE}(\hat{\theta}) = \sqrt{\mathrm{b}(\hat{\theta})^2 + \mathrm{sd}(\hat{\theta})^2}$

		Type	$\hat{\eta}_j$	$\hat{r}_j$	$\hat{\beta}_{1j}$	$\hat{\beta}_{2j}$
$d = 50$	$\eta = 1.00$	I	0.034	0.211	0.216	0.176
		II	0.042	0.266	0.189	0.196
	$\eta = 1.50$	I	0.024	0.190	0.112	0.104
		II	0.029	0.185	0.145	0.349
	$\eta = 1.95$	I	0.003	0.081	0.215	0.214
		II	0.004	0.095	0.282	0.269
$d = 100$	$\eta = 1.00$	I	0.031	0.203	0.090	0.085
		II	0.035	0.312	0.111	0.131
	$\eta = 1.50$	I	0.019	0.122	0.051	0.045
		II	0.034	0.272	0.203	0.227
	$\eta = 1.95$	I	0.002	0.072	0.070	0.059
		II	0.004	0.102	0.274	0.274

Table: RMSEs when $r = 3.0$, $\beta_1 = 5$ and $\beta_2 = 5$.

Approximation of the (full) ST likelihood (2)

Measure: Time (minutes)

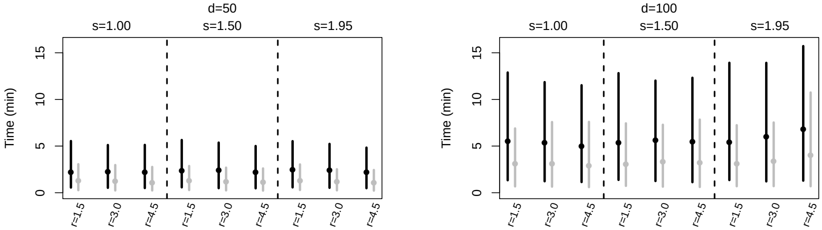


Figure: Mean time (in minutes) and 95% confidence region for the maximisation of the extremal skew-t likelihood function, using the Type I (black) and Type II (grey) approximations.

Performance of composite j-wise likelihoods

Focus on $d = 20$ case.

For $j \in \{1, \ldots, d\}$ and some $q \in \mathcal{Q}_d^{(j)}$ we define the **weights** as

$$w_q = \begin{cases} 1 & \text{if } \max_{i,k \in q; i \neq k} \|s_i - s_k\| < u \\ 0 & \text{otherwise} \end{cases}, \quad u > 0.$$

Evaluate statistical and computational efficiency via the **Time Root Relative Efficiency (TRRE)** criterion:

$$\mathrm{TRRE}(\theta_j) = \frac{\mathrm{RMSE}(\hat{\theta}_d)}{\mathrm{RMSE}(\hat{\theta}_j)} \times \frac{\mathrm{time}(\hat{\theta}_d)}{\mathrm{time}(\hat{\theta}_j)}.$$

$\Longrightarrow$ Values close to 1 indicate good performance of the j-wise likelihood.

Performance of composite j-wise likelihoods (2)

	$\eta = 1.00$	$\eta = 1.50$	$\eta = 1.95$
$j = 2$	04/06/05/04	05/03/04/03	04/02/04/03
$j = 3$	09/04/12/08	10/03/09/06	05/03/09/06
$j = 4$	21/15/21/10	12/12/13/08	07/17/14/10
$j = 5$	10/04/07/04	09/02/06/05	14/25/11/08
$j = 10$	15/19/10/12	18/26/19/14	16/26/10/10

Table: Time root relative efficiency (TRRE) of $\hat{\eta}_j/\hat{r}_j/\hat{\beta}_{1j}/\hat{\beta}_{2j}$ when $r = 3.0$.

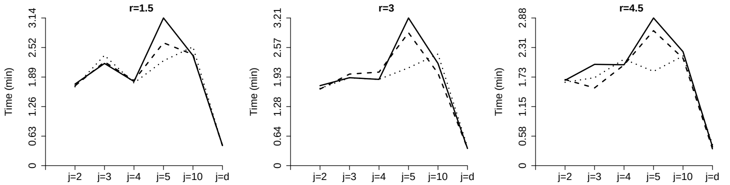
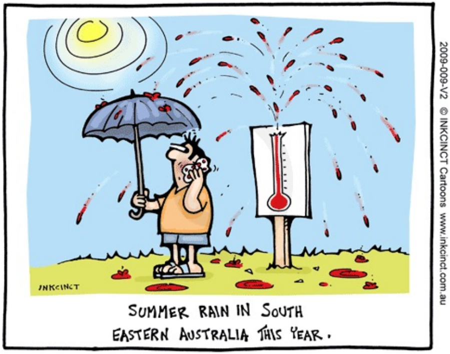


Figure: Average maximisation time (in mins) for the j-wise composite likelihood function. Smoothness values $\eta = 1, 1.5$ and 1.95 are represented by solid, dashed and dotted lines.

Talk Outline



1. **Max-stable processes**
   - Construction
   - Inference and limitations
2. Solution #1: Combining methodologies
   - Cdf approximations
   - Simulated examples
3. Solution #2: aggregating data
   - Methodology
   - Simulated and real examples
4. Discussion

Aggregating data

Main idea borrowed from **Symbolic Data Analysis**

- Summarise a complex & very large dataset in a compact manner.

$$S = \pi(X_{1:N}) : [\mathbb{X}]^N \to \mathbb{S} \text{ such that } x_{1:N} \mapsto \pi(x_{1:N})$$

- Collapse over data not needed in detail for analysis.

A likelihood-based approach: (Beranger, Lin & Sisson, 2018)

$$L(S|\theta,\phi) \propto \int_x g(S|x,\phi)L(x|\theta)dx$$

where

- $L(x|\theta)$ – standard, classical data likelihood
- $g(S|x,\phi)$ – probability of obtaining S given classical data x
- $L(S|\theta)$ – new symbolic likelihood for parameters of classical model

Gist: Fitting the standard classical model $L(x|\theta)$, when the data are viewed only through symbols S as summaries.

Specific case: Random histograms

Underlying data $X_1, \ldots, X_N \in \mathbb{R}^d \sim g(x|\theta)$ collected into random counts histogram, with fixed bins $\mathcal{B}_\mathbf{1}, \ldots, \mathcal{B}_\mathbf{B}$.

Aggregation:

$S = \pi(X_{1:N}) : \mathbb{R}^{d \times N} \to \mathbb{S} = \{0, \ldots, N\}^{B^1 \times \cdots \times B^d}$ such that
$x_{1:N} \mapsto \left(\sum_{i=1}^n \mathbb{I}\{x_i \in \mathcal{B}_\mathbf{1}\}, \ldots, \sum_{i=1}^n \mathbb{I}\{x_i \in \mathcal{B}_\mathbf{B}\}\right)$.

$$g(S|x, \phi) = \begin{cases} 1 & \text{if } s_b \text{ observations in bin } b\text{; for each } b = 1, \ldots, B \\ 0 & \text{else} \end{cases}$$

The symbolic likelihood is then (multinomial):

$$L(S|\theta) \propto \int_x g(S|x) \prod_{k=1}^n g(x_k|\theta) dx \propto \prod_{b=1}^B \left(\int_{B_b} g(z|\theta) dz\right)^{s_b}$$

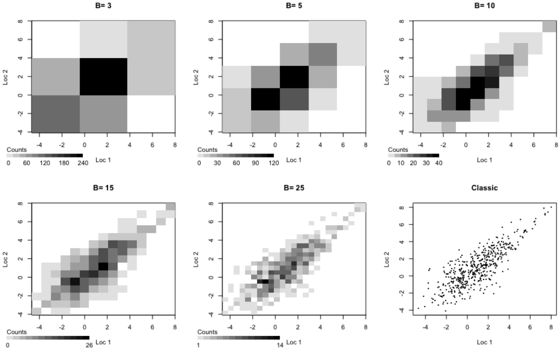
$\Rightarrow$ generalises univariate result of McLachlan & Jones (1988). ✓

Specific case: Random histograms

Can recover classical likelihood as $B \to \infty$

$$\lim_{B\to\infty} L(S|\theta) \propto \lim_{B\to\infty} \prod_{b=1}^{B} \left[\int_{B_b} g(z|\theta)dz \right]^{s_b} = L(X_1, \ldots, X_n|\theta)$$

$\Longrightarrow$ recover classical analysis as we approach classical data. ✓



Consistency: Can show that with a sufficient number of histogram bins can perform analysis arbitrarily close to analysis with full dataset.

Composite symbolic likelihoods

Limitations:

- Multivariate histograms become inefficient as d gets large – number of bins to cover d dimensions accurately gets large fast.
- Calculating $\int_{B_b} g(z|\theta)dz$: has 2^d components – viable for low d.

$\Longrightarrow$ **One option:** Composite likelihoods.

Consider $j = 2$, i.e. pairwise composite likelihood, we have

$$L_{CL}^{(2)}(\boldsymbol{S}|\theta) \propto \prod_i \prod_{j>i} L(S_{ij}|\theta)$$

where S_{ij} is the bivariate *marginal* histogram for dimensions (i, j) and

$$L(S_{ij}|\theta) \propto \prod_b \left(\int_{B_b} g(z_1, z_2|\theta) dz_1 dz_2 \right)^{s_b}.$$

Composite symbolic likelihoods

From $L(S|\theta)$ we have (for a single histogram):

$\hat{\theta}$ is **asymptotically consistent and distributed as**

$$\sqrt{N}\left(\hat{\theta}-\theta\right) \to N\left(0,\, I(\theta)^{-1}\right)$$

when

- $N \to \infty$
- Number of bins $\to \infty$ and volume of each bin $\to 0$ (because then $L(S|\theta) \to L(x|\theta)$)

But when the bins (number and volume) are fixed then

$$\sqrt{N}\left(\hat{\theta}-\theta\right) \to N\left(??(\theta, bins),\, ??(\theta, Bins)^{-1}\right).$$

- Currently working on non-asymptotic (in bins) distribution of MLE

Composite symbolic likelihoods

From $L_{SCL}^{(j)}(\boldsymbol{S}|\theta)$ we have (for a single histogram):

$\hat{\theta}_{SCL}^{(j)}$ is **asymptotically consistent and distributed as**

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

when

- $N \to \infty$
- Number of bins $\to \infty$ and volume of each bin $\to 0$ (because then $L_{SCL}^{(j)}(\boldsymbol{S}|\theta) \to L_{CL}^{(j)}(\boldsymbol{x}|\theta)$)

But when the bins (number and volume) are fixed then, as before

$$\sqrt{N}\left(\hat{\theta}_{SCL}^{(j)} - \theta\right) \to N\left(??(\theta, bins),\ ??(\theta, Bins)^{-1}\right).$$

- Similarly work in progress.

Simulated spatial extremes

(Mean) Pairwise symbolic composite likelihood estimates ($\hat{\theta}^{(2)}_{SCL}$):

- Consider $N = 1\,000$ observations at $K = 15$ spatial locations and $T = 1$ random histogram
- Spatial dependence of Smith model is $\sigma_{11} = 300$, $\sigma_{12} = 150$ and $\sigma_{22} = 200$

B	σ_{11}	σ_{12}	σ_{22}
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)**

Table: Mean (and standard errors) of the symbolic composite MLE $\hat{\theta}^{(2)}_{SCL}$ and composite MLE $\hat{\theta}^{(2)}_{CL}$ (Classic) from 1000 replications of the Gaussian max-stable process model, for $B \times B$ histograms for varying values of B.

- As "bins $\to \infty$" performance approaches classical composite likelihood (also estimated the marginal parameters).
- "Acceptable" results for $B = 10$

Simulated spatial extremes

(Mean) Time comparisons for increasing N

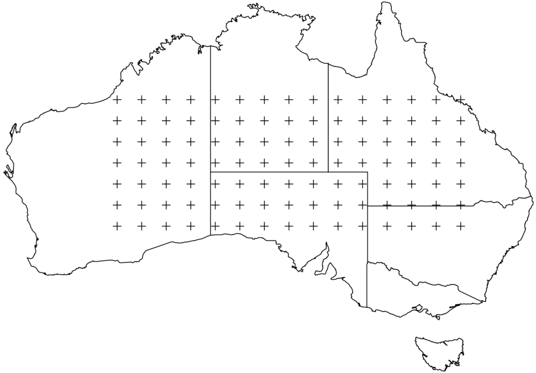
- Consider $B = 25$ bins, $K = 10, 100$ spatial locations and $T = 1$ random histogram. Repetitions $= 10$

N	$K = 10$				$K = 100$			
	t_c	t_s	t_{histDR}	t_{histR}	t_c	t_s	t_{histDR}	t_{histR}
1 000	71.9	22.5	0.8	0.1	–	2 238.0	78.8	12.0
5 000	291.8	19.0	0.8	0.3	–	2 650.2	81.7	30.9
10 000	591.7	23.8	0.9	0.5	–	2 356.6	85.8	54.1
50 000	2 626.8	24.2	1.7	2.1	–	2 300.6	131.6	237.0
100 000	5 610.7	25.4	2.4	4.2	–	2 766.9	188.2	461.8
500 000	31 083.1	23.2	7.5	20.6	–	3 111.5	627.1	2 243.5

Table: Mean computation times (seconds) for different components involved in computing $\hat{\theta}^{(2)}_{CL}$ and $\hat{\theta}^{(2)}_{SCL}$.

- Classical composite likelihood rapidly not feasible as spatial dimensions increases ($K = 20$)
- Symbolic approach much more efficient

Modelling Australian maximum temperature



- 105 spatial locations with temperature observation, over time
- Want to fit spatial model to temperature extremes.
- **Lots of data:**
  - Can't fit using $L(X|\theta)$ or $L_{\mathrm{CL}}^{(j)}(X|\theta)$
  - Can form 105-dimensional histogram(!)
  - $L(S|\theta)$ is completely infeasible
  - Solution **$105 \times 104/2$ bivariate histograms**

Modelling Australian maximum temperatures (2)

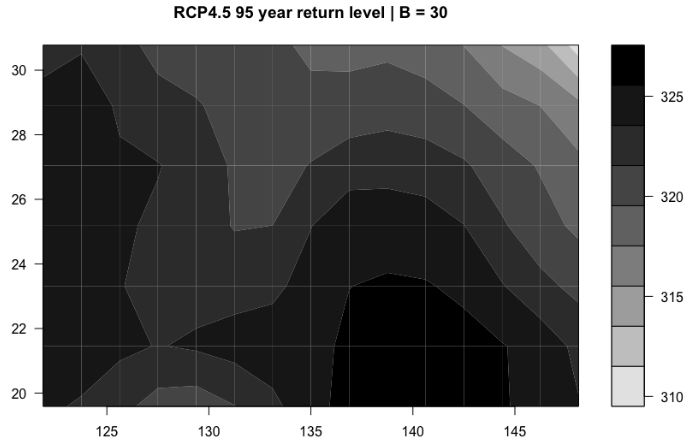
The data:

- Historical observations (1850 – 2006)
- Simulated observations (2006 – 2100) from CSIRO Mk3.6 model with 2 scenarios: RCP4.5 and RCP8.5
- 90 days across summer months (DJF)
- 15-day blocks (6 obs per year)
- μ and σ modelled as functions of space

B	σ_{11}	σ_{12}	σ_{22}	ξ
		Historical Data		
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		
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		
20	128.0 (6.30)	-19.6 (1.29)	66.6 (3.32)	-0.231 (0.059)
25	136.0 (3.95)	-15.1 (0.93)	59.4 (3.17)	-0.234 (0.060)
30	129.9 (4.01)	-13.6 (0.83)	56.4 (2.94)	-0.233 (0.055)

Table: Means and standard errors of the composite MLEs for the Smith model.

Modelling Australian maximum temperatures (3)



Summary

2 solutions to fit max-stable models in high-dimensions

Solution #1:

- Time of occurrences should be recorded (mix CL and ST likelihoods);
- (Crude) Approximations of cdfs are essential;
- Application to 90-dim temperature data from Inner Melbourne region.

Solution #2:

- Aggregating data into histograms;
- Composite likelihood on histogram likelihood;
- Effect of number of histograms and allocation of micro-data data between them;
- Comparing bivariate SCL and trivariate SCL;
- Application to 105-dim Australian temperature data.

THANK YOU

Relevent Manuscripts:

- *Beranger B., A. G. Stephenson & S. A. Sisson (2020). High-dimensional inference using the extremal skew-t process.* *Extremes, In press.*
- *Whitaker T., B. Beranger & S. A. Sisson (2020). Composite likelihood functions for histogram-valued random variables.* *Stat. Comput., In press.*

Contact:

B.Beranger@unsw.edu.au
www.borisberanger.com