First steps in the analysis of Symbolic Data

B. Beranger

jointly with T. Whitaker, J. Lin and S. A. Sisson

UNSW Sydney & ACEMS

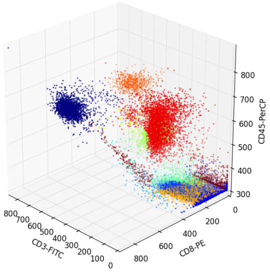
Statistics Seminar, University of Melbourne,
4 October 2017

Outline

Outline

1. What is Symbolic Data Analysis?
2. Existing and new SDA models
3. An example in EVT
4. Discussion

Rise of non-standard data forms



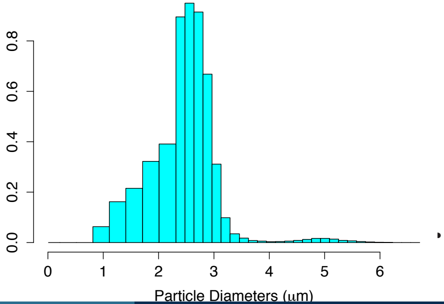
Standard statistical methods analyse classical datasets

E.g. $x_1, \ldots, x_n$ where $x_i \in \mathcal{X} = \mathbb{R}^p$

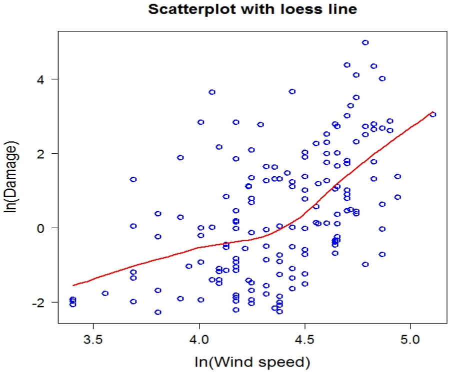
However: Increasingly see non-standard data forms for analysis.

Simple non-standard forms:

- Can arise as result of measurement process
- Blood pressure naturally recorded as (low, high) interval
- Particulate matter directly recorded as counts within particle diameter ranges i.e. histogram

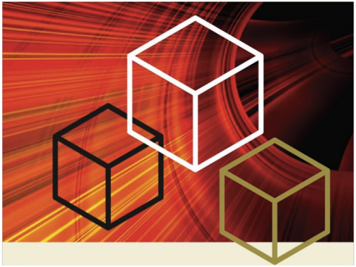


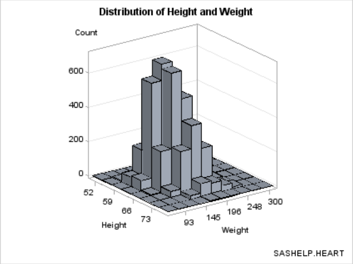
Example: Discretised data = histogram



- E.g. point $(4.0, 0.0)$ actually lies within $[3.95, 4.05) \times [-0.05, 0.05)$
- Strong discretisation could have undesired inferential impact

Symbolic Data Analysis





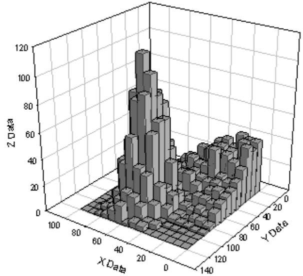
- Established by Diday & coauthors in 1990s.
- Basic unit of data is a distribution rather than usual datapoint.
  - interval (a, b)
  - p-dim hyper-rectangle
  - histogram
  - weighted list etc.
  - can be complicated by "rules"
- Classical data are a special case of symbolic data:

  E.g. symbolic interval $s = (a, b)$ equivalent to classical data point x if $x = a = b$.

  Or histogram $\rightarrow \{x_i\}$ as $\#$ bins $\rightarrow \infty$.

  $\Longrightarrow$ symbolic analyses must reduce to classical methods.

How do symbolic data arise?



Big data → small (symb) data
Easier to analyse (hopefully!)

Possible use in data privacy?
Individual can't be identified.

- Can arise naturally (measurement error):
  E.g. blood pressure, particulate histogram, truncation/rounding.
- 'Big Data' context:
  - Symbolic data points can summarise a complex & very large dataset in a compact manner.
  - Retaining maximal relevant information in original dataset.
  - Collapse over data not needed in detail for analysis.
  - Summarised data have own internal structure, which must be taken into account in any analysis.

Statistical question:
How to do statistical analysis for this form of data?

Outline

How to analyse symbolic data?

State of the art:

Poorly developed in terms of inferential methods.

Current approaches:

- Descriptive statistics (means, covariances)
  $\Rightarrow$ Methods based on $1^{st}/2^{nd}$ moments: clustering, PCA etc.
- Ad-hoc approaches (e.g. regression)
  $\Rightarrow$ Can be plain wrong for inference/prediction.
- Single technique for constructing likelihood functions
  $\Rightarrow$ Limited model-based inferences

Over-prevalence of models for intervals & assuming uniformity
$\Rightarrow$ Need to move beyond uniformity (Lynne Billard)

Current SDA research:

Developing practical model-based (e.g. likelihood-based) procedures for statistical inference using symbolic data for general symbols.

Existing models for symbols (1) (Le Rademacher & Billard, 2011)

Symbol: $S = (S^1, \ldots, S^n)^\top$
E.g. For random intervals $[a_i, b_i]$, $i = 1, \ldots, n$:
- $\circ\, S_i = (a_i, b_i)^\top$
- $\circ\, S_i = (m_i, \log r_i)^\top$

Then specify a standard (classical data) model for $S_1, \ldots, S_n$. E.g.

$$(m_i, \log r_i)^\top \sim N(\mu, \Sigma)$$

Model specification issues:

- Need to find credible models for general S
  - Not always obvious how to do this.
  - Easy to specify models for classical data (e.g. GEV).
  - How to develop models for symbols (with internal variation)?
  - Can't just fit to means. How to account for variation? etc.

Existing models for symbols (2) (Le Rademacher & Billard, 2011)

Inference issues:

- Symbol are summaries of classical data
  - Inference at symbol level only
- Ok but what if interest in modelling underlying data?
  - Want full distributional predictions of x (not just mean/var)

Symbol issues:

- Symbol assumptions are sometimes unrealistic
  - Distribution with the interval $[a, b]$ often assumed uniform.
  - Extremely unlikely and affects inference/prediction.
- Symbol parametrisation are not always stable
  - E.g. $[a, b] = (m, \log r)^\top$, when $a \to b$ then $\log r \to \infty$

Q: How to fit models and make predictions at the level of the classical data, based on observed symbols?

One possible approach (Beranger, Lin & Sisson, 2017, in preparation)

The general approach:

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

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.

- **Limiting case:** as $S_i \to x_i$, then $g(S_i|x,\phi) \to g(x_i|x) = \delta_{x_i}(x)$ and so

$$L(S_i|\theta,\phi) \propto \int_x \delta_{x_i}(x)L(x|\theta)dx = L(x_i|\theta) \qquad \text{(classical likelihood)}$$

- Different symbols give different $g(S|x,\phi)$ (and $\therefore$ $L(S|\theta,\phi)$).

ACEMS

How to construct $g(S|x, \phi)$?

- Typically we can easily describe the distribution of $X|S$:
  - Intervals: $x \sim U(a, b)$ where $S = (a, b)^\top$
  - Histograms: $x \sim \begin{cases} w_i U(b_i, b_{i+1}) & b_i \leq x \leq b_{i+1} \\ 0 & \text{else} \end{cases}$

    for fixed $\{b_i\}$ where $S = (s_1, \ldots, s_B)^\top$, $w_i = s_i / \sum_k s_k$.
  - Gaussian: $x \sim N(\mu, \Sigma)$ where $S = (\mu, \Sigma)^\top$.
- Although $U(a, b)$ specifications are unrealistic (we avoid this later).
- If we specify a prior/marginal on S, we then obtain

$$g(S|x, \phi) = g(S|x) = \frac{f(x|S)f(S)}{f(x)}$$

  where $f(x) = \int f(x|S)f(S)dx$.
- Cute for Bayesians: use a posterior to build a classical likelihood :-)

Specific cases (1)

- **Example (1)**: No specified generative model $L(x|\theta)$

$$\begin{aligned} L(S|\theta,\phi) &\propto \int_x g(S|x,\phi)L(x|\theta)dx \\ \Rightarrow L(S|\phi) &\propto g(S|\phi) \end{aligned}$$

That is:

Directly modelling symbol = existing likelihood approach

(Le Rademacher & Billard, 2011) ✓

Specific cases (2): Random intervals

- **Example (2):** Random intervals: $S = (S_\ell, S_u)^\top$

  Assume:

  - $X_1, \ldots, X_n \sim h(x|\omega)$ for some h (not uniform!) and
  - $S_\ell = X_{(\ell)}$ and $S_u = X_{(u)}$ are lower/upper order statistics.

  Then density of $X|S$ is easily specified as:

$$\mathbf{f(x|s_\ell, s_u)} =$$

$$\prod_{k=1}^{n} \mathbf{h}^{(\mathbf{s}_\ell)}(\mathbf{x_k}|\omega) \prod_{k=1}^{n} \mathbf{h}^{(\mathbf{s}_\ell,\mathbf{s_u})}(\mathbf{x_k}|\omega) \prod_{k=1}^{n} \mathbf{h}^{(\mathbf{s}_\ell)}(\mathbf{x_k}|\omega)\delta_{\mathbf{s}_\ell}(\mathbf{x}_{(\ell)})\delta_{\mathbf{s_u}}(\mathbf{x_{(u)}})$$

  where

  - $x = (x_{(1)}, \ldots, x_{(n)})^\top$
  - $h^{(s_\ell)}(x|\omega) = h(x|\omega)/H(s_\ell|\omega)I(x < s_\ell)$,
    $h^{(s_u)}(x|\omega) = h(x|\omega)/(1 - H(s_u|\omega))I(x > s_u)$,
    $h^{(s_\ell,s_u)}(x|\omega) = h(x|\omega)/(H(s_u|\omega) - H(s_\ell|\omega))I(s_\ell < x < s_u)$.
  - Delta functions enforce $x_{(\ell)} = S_\ell$ and $x_{(u)} = S_u$.

Specific cases (2): Random intervals

- Now, as $X_1, \ldots, X_n \sim h(x|\omega)$, we also have

$$\mathbf{f}(\mathbf{s}_\ell, \mathbf{s}_\mathbf{u}|\omega) = \frac{\mathbf{n}!}{(\ell - \mathbf{1})!(\mathbf{u} - \ell - \mathbf{1})!(\mathbf{n} - \mathbf{u})!} \mathbf{H}(\mathbf{s}_\ell|\omega)^{\ell - \mathbf{1}} \times \left[\mathbf{H}(\mathbf{s}_\mathbf{u}|\omega) - \mathbf{H}(\mathbf{s}_\ell|\omega)\right]^{\mathbf{u} - \ell - \mathbf{1}} \left[\mathbf{1} - \mathbf{H}(\mathbf{s}_\mathbf{u}|\omega)\right]^{\mathbf{n} - \mathbf{u}} \mathbf{h}(\mathbf{s}_\ell|\omega)\mathbf{h}(\mathbf{s}_\mathbf{u}|\omega)$$

where $H(x|\omega) = \int h(z|\omega)dz$.

- And so we have the joint distribution as

$$\mathbf{f}(\mathbf{x}, \mathbf{s}_\ell, \mathbf{s}_\mathbf{u}|\omega) = \frac{\mathbf{n}!}{(\ell - \mathbf{1})!(\mathbf{u} - \ell - \mathbf{1})!(\mathbf{n} - \mathbf{u})!} \prod_{\mathbf{k}=\mathbf{1}}^{\mathbf{n}} \mathbf{h}(\mathbf{x}_\mathbf{k}|\omega)\delta_{\mathbf{s}_\ell}(\mathbf{x}_{(\ell)})\delta_{\mathbf{s}_\mathbf{u}}(\mathbf{x}_{(\mathbf{u})})$$

and finally

$$\mathbf{g}(\mathbf{s}_\ell, \mathbf{s}_\mathbf{u}|\mathbf{x}) = \frac{\mathbf{n}!}{(\ell - \mathbf{1})!(\mathbf{u} - \ell - \mathbf{1})!(\mathbf{n} - \mathbf{u})!} \delta_{\mathbf{s}_\ell}(\mathbf{x}_{(\ell)})\delta_{\mathbf{s}_\mathbf{u}}(\mathbf{x}_{(\mathbf{u})}).$$

- **Note:** This is independent of the form of $h(x|\omega)$!

Specific cases (2): Random intervals

- Now if we want to fit the model $X_1, \ldots, X_n \sim g(x|\theta)$, this gives us

$$\begin{aligned} \mathbf{L}(\mathbf{s}_\ell, \mathbf{s}_\mathbf{u}|\theta) &\propto \int_{\mathbf{x}} \mathbf{g}(\mathbf{s}_\ell, \mathbf{s}_\mathbf{u}|\mathbf{x}, \phi) \prod_{\mathbf{k}=1}^{\mathbf{n}} \mathbf{g}(\mathbf{x}_\mathbf{k}|\theta) \mathbf{dx} \\ &\propto \frac{\mathbf{n}!}{(\ell-\mathbf{1})!(\mathbf{u}-\ell-\mathbf{1})!(\mathbf{n}-\mathbf{u})!} \mathbf{G}(\mathbf{s}_\ell|\theta)^{\ell-1} [\mathbf{G}(\mathbf{s}_\mathbf{u}|\theta) - \mathbf{G}(\mathbf{s}_\ell|\theta)]^{\mathbf{u}-\ell-1} \\ &\quad \times [\mathbf{1} - \mathbf{G}(\mathbf{s}_\mathbf{u}|\theta)]^{\mathbf{n}-\mathbf{u}} \mathbf{g}(\mathbf{s}_\ell|\theta) \mathbf{g}(\mathbf{s}_\mathbf{u}|\theta) \end{aligned}$$

  where $G(x|\theta) = \int g(z|\theta)dz$
  $\Rightarrow$ the (known) joint distribution of (ℓ, u)-th order statistics of $\{X_k\}$. ✓

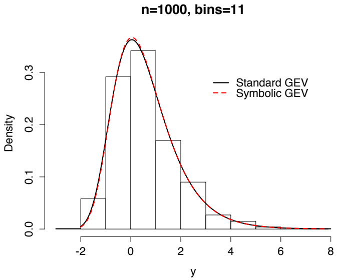
- When $S_\ell = \min_k X_k$ and $S_u = \max_k X_k$:

$$\mathbf{L}(\mathbf{s}_1, \mathbf{s}_\mathbf{n}|\theta) \propto \mathbf{n}(\mathbf{n}-\mathbf{1}) [\mathbf{G}(\mathbf{s}_\mathbf{n}|\theta) - \mathbf{G}(\mathbf{s}_1|\theta)]^{\mathbf{n}-2} \mathbf{g}(\mathbf{s}_1|\theta) \mathbf{g}(\mathbf{s}_\mathbf{n}|\theta), \quad \mathbf{s}_1 < \mathbf{s}_2$$

  $\Rightarrow$ the (known) joint distribution of min/max of $\{X_k\}$. ✓

- **Symbolic $\rightarrow$ Classical check:**
  If $S_\ell \rightarrow S_u = x$ (with $n = 1$) then $L(s_\ell, s_u|\theta) = g(x|\theta)$. ✓

Specific cases (3): Random histograms



- Underlying data $X_1, \ldots, X_n \in \mathbb{R}^p \sim h(x|\omega)$.
- Collected into histogram (random counts) with fixed bins as:

$$\mathbf{S} = (\mathbf{s_1}, \ldots, \mathbf{s_B})^\top$$
$$= (\#\mathbf{X_i} \in \mathbf{B_1}, \ldots, \#\mathbf{X_i} \in \mathbf{B_B})^\top$$

such that $\sum_b s_b = n$.

- The density of $X|S$ is

$$\mathbf{f(x|s)} = \prod_{\mathbf{b=1}}^{\mathbf{B}} \prod_{\ell=\mathbf{1}}^{\mathbf{s_b}} \mathbf{h^{(b)}(x_b^\ell|\omega) I(x_b^\ell \in B_b)}$$

where

- x_b^ℓ is the ℓ-th observation in bin B_b.
- $h^{(b)}(x|\omega) \propto h(x|\omega) I(x \in B_b)$.
- Enforces s_b observations in bin B_b.

Specific cases (3): Random histograms

- By construction the (prior) distribution of counts $S = (s_1, \ldots, s_B)^\top$ is

$$f(S|\omega) = \frac{n!}{s_1! \ldots s_B!} \prod P_b^h(\omega)^{s_b}$$

where

$$P_b^h(\omega) = \int_{B_b} h(x|\omega) dx$$

is the probability that any x will fall in bin B_b.

- Consequently

$$f(x, S|\omega) = \frac{n!}{s_1! \ldots s_n!} \prod_{i=1}^{n} h(x_i|\omega) \prod_{b=1}^{B} I\left(\sum_{i=1}^{n} I(x_i \in B_b) = s_b\right)$$

Specific cases (3): Random histograms

- As a result

$$g(S|x) = \frac{n!}{s_1! \dots s_B} \prod_{b=1}^{B} I\left(\sum_{i=1}^{n} I(x_i \in B_b) = s_b \right).$$

- Now if we want to fit the model $X_1, \dots, X_n \sim g(x|\theta)$, this gives us

$$\mathbf{L}(\mathbf{S}|\theta) \propto \int_{\mathbf{x}} \mathbf{g}(\mathbf{S}|\mathbf{x}) \prod_{\mathbf{k}=1}^{\mathbf{n}} \mathbf{g}(\mathbf{x_k}|\theta)\mathbf{dx}$$

$$\propto \frac{\mathbf{n}!}{\mathbf{s_1}! \dots \mathbf{s_n}!} \prod_{\mathbf{b}=1}^{\mathbf{B}} [\mathbf{P_b^g}(\theta)]^{\mathbf{s_b}}$$

where $P_b^g(\theta) = \int_{B_b} g(x|\theta)dx$

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

Specific cases (3): Random histograms

- **Limiting case:** recover classical likelihood as $B \to \infty$

$$\lim_{B\to\infty} L(S|\theta) \propto \lim_{B\to\infty} \frac{n!}{s_1!\dots s_B!} \prod_{b=1}^{B} \left[\int_{B_b} g(z|\theta)dz\right]^{s_b}$$
$$= L(X_1,\dots,X_n|\theta)$$

$\Rightarrow$ 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.
- **Computationally scalable:** Working with counts not computationally expensive latent data.
- **Some approximation of $L(S|\theta)$ to $L(x|\theta)$** depending on level of discretisation. Work needed to quantify this.
- More complicated if data are not *iid* but exchangeable (Zhang & Sisson, in preparation)

Outline

Motivation

- QUESTION: What is the expected maximum temperature across some region within the next 50 or 100 years?

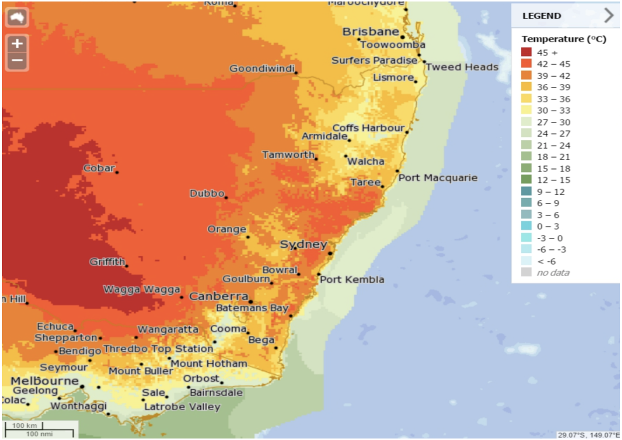


Figure: Heat wave in South East Australia (January 2017)

Motivation

- What do we know?
  - Environmental extremes are spatial $\Rightarrow$ SPATIAL EXTREMES
  - Max-stable processes are a convenient tool
- Drawbacks and challenges?
  - High dimensional distributions not always available, computationally costly $\Rightarrow$ Composite likelihood (Padoan et al. 2010)
  - Unfeasible for a large number of locations and temporal observations
- PROPOSAL: use Symbolic Data Analysis (SDA)

Max-stable processes

- **Definition:** Let $X_1, X_2, \ldots,$ be i.i.d replicates of $X(s), s \in \mathcal{S} \subset \mathbb{R}^d$.
  $Y(s)$ is a max-stable process if $\exists\, a_n(s) > 0$ and $b_n(s)$, continuous such that

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

- Spectral representation (de Haan, 1984; Schlather, 2002) $\Rightarrow$ Max-stable models

Gaussian extreme value model (Smith, 1990) defined by

$$Y(s) = \max_{1 \leq i} \{\zeta_i \phi_d(s; t_i, \Sigma)\}, s \in \mathbb{R}^d$$

where $(\zeta_i, t_i)_{1 \leq i}$ are the points of a point process on $(0, \infty) \times \mathbb{R}^d$,
For $d = 2$, the bivariate cdf of $(Y(s_1), Y(s_2))$, $s_1, s_2 \in \mathbb{R}^2$ is

$$P(Y(s_1) \leq y_1, Y(s_2) \leq y_2) = \exp\left(-\frac{1}{v_1}\Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{v_2}{v_1}\right) - \frac{1}{v_2}\Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{v_1}{v_2}\right)\right),$$

where $v_i = \left(1 - \xi_i \frac{y_i - \mu_i}{\sigma_i}\right)^{-\frac{1}{\xi}}, i = 1, 2$ and $a^2 = (z_1 - z_2)^T \Sigma^{-1} (z_1 - z_2)$

Composite Likelihood (1)

- Let $\mathbf{X} = (X_1, \ldots, X_N)$ denote a vector of N i.i.d. rv's taking values in $\mathbb{R}^K$ with realisation $\mathbf{x} = (x_1, \ldots, x_N) \in \mathbb{R}^{K \times N}$ and density function $g_{\mathbf{X}}(\cdot; \theta)$.
- Define a subset of $\{1, \ldots, K\}$ by $\mathbf{i} = (i_1, \ldots, i_j)$, where $i_1 < \cdots < i_j$ with $i_j \in \{1, \ldots, K\}$ for $j = 1, \ldots, K-1$.
- Then for $n = 1, \ldots, N$, $x_n^{\mathbf{i}} \in \mathbb{R}^j$ defines a subset of x_n and $\mathbf{x}^{\mathbf{i}} = (x_1^{\mathbf{i}}, \ldots, x_N^{\mathbf{i}}) \in \mathbb{R}^{j \times N}$, defines a subset of $\mathbf{x}$.

The **j-wise composite likelihood function, $\mathbf{CL}^{(j)}$** , is given by

$$L_{CL}^{(j)}(\mathbf{x}; \theta) = \prod_{\mathbf{i}} g_{\mathbf{X}^{\mathbf{i}}}(\mathbf{x}^{\mathbf{i}}; \theta),$$

where $g_{\mathbf{X}^{\mathbf{i}}}$ is a $j-$dimensional likelihood function.

Composite Likelihood (2)

- When $j = 2$, the *pairwise* composite log-likelihood function, $l_{CL}^{(2)}$ is given by

$$l_{CL}^{(2)}(\mathbf{x};\theta) = \sum_{i_1=1}^{K-1} \sum_{i_2=i_1+1}^{K} \log g_{\mathbf{X}^{\mathbf{i}}}(\mathbf{x}^{i_1}, \mathbf{x}^{i_2};\theta) \Rightarrow \frac{NK(K-1)}{2} \text{ terms}$$

- The resulting **maximum j-wise composite likelihood estimator** $\hat{\theta}_{CL}^{(j)}$ is asymptotically consistent and distributed as

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

where $G(\theta) = \mathbf{H}(\theta)\mathbf{J}(\theta)^{-1}\mathbf{H}(\theta)$, $\mathbf{J}(\theta) = \mathbb{V}(\nabla l_{CL}^{(j)} l(\theta))$ is a variability matrix and $\mathbf{H}(\theta) = -\mathbb{E}(\nabla l_{CL}^{2(j)}(\theta))$ is a sensitivity matrix.

Histogram-valued symbols (1)

- Consider we are only interested in a subset of size j of the K dimensions
- Let $\mathrm{b}^{\mathbf{i}}$ be the subset of b defining the coordinates of a $j-$**dimensional histogram bin** and let $\mathbf{B}^{\mathbf{i}} = (B^{i_1}, \ldots, B^{i_j})$ be the vector of the number of marginal bins.

The symbolic likelihood function associated with the vector of counts $\mathbf{s}_j^{\mathbf{i}} = (s_{\mathbf{1}^{\mathbf{i}}}^{\mathbf{i}}, \ldots, s_{\mathbf{B}^{\mathbf{i}}}^{\mathbf{i}})$ of length $B^{i_1} \times \cdots \times B^{i_j}$ is

$$L(\mathbf{s}_j^{\mathbf{i}}; \theta) = \frac{N!}{s_{\mathbf{1}^{\mathbf{i}}}^{\mathbf{i}}! \cdots s_{\mathbf{B}^{\mathbf{i}}}^{\mathbf{i}}!} \prod_{\mathrm{b}^{\mathbf{i}}=\mathbf{1}^{\mathbf{i}}}^{\mathbf{B}^{\mathbf{i}}} P_{\mathrm{b}^{\mathbf{i}}}(\theta)^{s_{\mathrm{b}^{\mathbf{i}}}^{\mathbf{i}}},$$

where $P_{\mathrm{b}^{\mathbf{i}}}(\theta) = \int_{\Upsilon_{b_{i_1}}^{i_1}} \ldots \int_{\Upsilon_{b_{i_j}}^{i_j}} g_X(x; \theta) dx$ and g_X is a $j-$dim density.

Histogram-valued symbols (2)

- $\mathbf{s}_j = \{\mathbf{s}_{jt}^{\mathbf{i}}; t = 1, \ldots, T, \mathbf{i} = (i_1, \ldots, i_j), i_1 < \ldots < i_j\}$ represents the set of $j-$dimensional observed histograms for the symbolic-valued random variable $\mathbf{S}_j$

- The symbolic $j-$wise composite likelihood function (SCL$^{(j)}$) is given by

$$L_{SCL}^{(j)}(\mathbf{s}_j; \theta) = \prod_{t=1}^{T} \prod_{\mathbf{i}} L(\mathbf{s}_{jt}^{\mathbf{i}}; \theta)$$

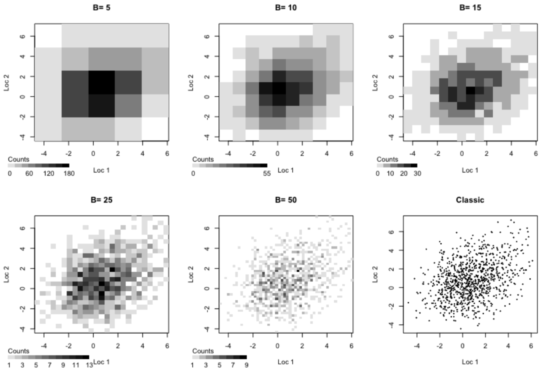
- Components of the **Godambe matrix** are given by

$$\hat{H}(\hat{\theta}_{SCL}^{(j)}) = -\frac{1}{N} \sum_{t=1}^{T} \sum_{\mathbf{i}} \nabla^2 l(\mathbf{s}_{jt}^{\mathbf{i}}; \hat{\theta}_{SCL}^{(j)})$$

$$\hat{J}(\hat{\theta}_{SCL}^{(j)}) = \frac{1}{N} \sum_{t=1}^{T} \left(\sum_{\mathbf{i}} \nabla l(\mathbf{s}_{jt}^{\mathbf{i}}; \hat{\theta}_{SCL}^{(j)}) \right) \left(\sum_{\mathbf{i}} \nabla l(\mathbf{s}_{jt}^{\mathbf{i}}; \hat{\theta}_{SCL}^{(j)}) \right)^{\top}$$

Simulation experiments: the set up

- K locations are generated uniformly on a $(0, 40) \times (0, 40)$ grid
- N realisations of the Smith model are generated for each location
- MLE's are obtained using $\mathrm{CL}^{(2)}$ and $\mathrm{SCL}^{(2)}$



Experiement 1 - Increasing the number of bins

- $N = 1000$, $K = 15$, $T = 1$, $\Sigma = \begin{bmatrix} 300 & 0 \\ 0 & 300 \end{bmatrix}$, Repetitions $= 1000$

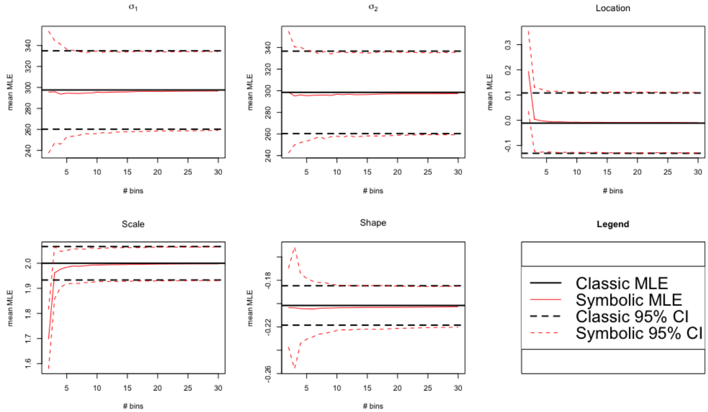


Figure: Mean of MLEs for $\theta = (\sigma_{11}, \sigma_{12}, \sigma_{22}, \mu, \sigma, \xi)$ using $\mathrm{CL}^{(2)}$ and $\mathrm{SCL}^{(2)}$, for increasing number of bins in bivariate histograms.

Experiement 2 - Computation time

- $B = 25,\ K = 10, 100,\ T = 1,$ Repetitions $= 10$

	$K = 10$			$K = 100$		
N	t_c	t_s	t_{hist}	t_c	t_s	t_{hist}
100	9.8	18.6	0.7	9758.6	1594.5	72.3
500	27.6	26.2	0.8	45040.1	2218.8	74.2
1000	71.9	22.5	0.8	-	2238.0	78.8
5000	291.8	19.0	0.8	-	2650.2	81.7
10000	591.7	23.8	0.9	-	2356.6	85.8
50000	2626.8	24.2	1.7	-	2300.6	131.6
100000	5610.7	25.4	2.4	-	2766.9	188.2
500000	31083.1	23.2	7.5	-	3111.5	627.1

Table: Mean computation times (sec) to optimise the regular and symbolic composite likelihood (t_c and t_s), and to aggregate the data into bivariate histograms (t_{hist})

Experiement 3 - Convergence of variances (1)

- $B = 25$, $N = 1000$, $K = 10$, Number of repetitions $= 1000$

T	σ_{11}	σ_{12}	σ_{22}	μ	σ	ξ
4	226.93	97.63	167.27	0.105	0.051	0.030
5	203.04	87.36	149.66	0.095	0.047	0.028
10	143.92	61.95	106.04	0.071	0.036	0.021
20	102.23	44.04	75.27	0.054	0.029	0.016
40	72.93	31.48	53.64	0.043	0.024	0.013
50	65.52	28.31	48.16	0.040	0.023	0.012
100	47.38	20.55	34.71	0.034	0.020	0.011
200	34.87	15.23	25.42	0.030	0.018	0.010
1000	21.12	13.08	13.11	0.025	0.016	0.010
Classic	16.65	10.53	10.69	0.020	0.014	0.009

Table: Mean variances calculated from $\text{CL}^{(2)}$ and $\text{SCL}^{(2)}$ for $\theta = (\sigma_{11}, \sigma_{12}, \sigma_{22}, \mu, \sigma, \xi)$ for increasing T.

Experiement 3 - Convergence of variances (2)

- $\hat{J}(\hat{\theta}^{(j)}_{SCL})$ requires $T \to N$ and $\mathbf{B} \to \infty$ for the convergence towards the classical Godambe matrices to occur.

- For T fixed, convergence still occurs as $\mathbf{B} \to \infty$ towards a different expression

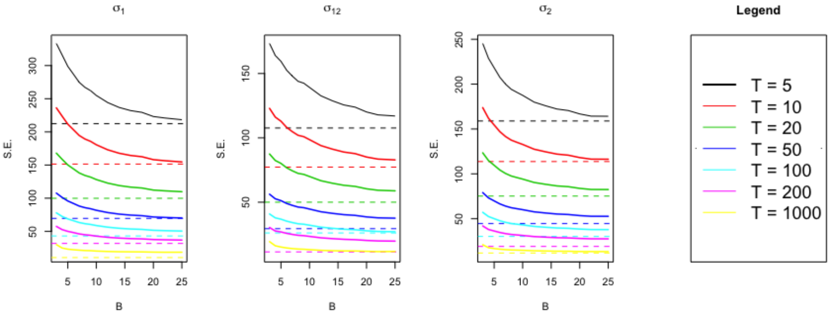


Figure: Mean variances calculated from SCL$^{(2)}$ for fixed T and increasing $\mathbf{B}$.

Real data analysis: an overview

- Maximum temperatures across Australia
- **Data:**
  - Focus on fortnighly maxima at $K = 105$ locations over summer months
  - 3 sets: historical ($N = 970$), RCP4.5 and RCP8.5 (both $N = 540$)
- Bivariate histograms are constructed for all pairs of locations for $B = 15, 20, 25, 30$.

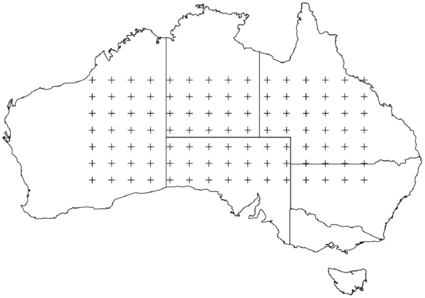


Figure: Study region

Model fitting

- Fit the Smith model with mean and variance parameters as linear functions of space

	σ_{11}	σ_{12}	σ_{22}	ξ
B	Historical Data			
15	176.4 (0.285)	-28.7 (0.032)	76.8 (0.329)	-0.266 (0.053)
20	164.2 (0.289)	-29.3 (0.030)	74.3 (0.469)	-0.264 (0.049)
25	162.4 (0.217)	-29.9 (0.033)	75.3 (0.284)	-0.264 (0.049)
30	161.6 (0.201)	-32.3 (0.029)	74.4 (0.234)	-0.264 (0.050)
B	RCP4.5 Data			
15	160.9 (0.942)	-34.1 (0.083)	79.0 (0.222)	-0.249 (0.074)
20	163.5 (0.595)	-41.1 (0.073)	77.6 (0.245)	-0.249 (0.076)
25	150.3 (0.349)	-33.1 (0.065)	70.7 (0.170)	-0.250 (0.073)
30	150.2 (0.150)	-31.6 (0.024)	70.7 (0.154)	-0.250 (0.069)
B	RCP8.5 Data			
15	128.7 (0.860)	-19.6 (0.092)	67.7 (0.392)	-0.232 (0.061)
20	128.0 (0.630)	-19.6 (0.129)	66.6 (0.332)	-0.231 (0.059)
25	136.0 (0.395)	-15.1 (0.093	59.4 (0.317)	-0.234 (0.060)
30	129.9 (0.401)	-13.6 (0.083)	56.4 (0.294)	-0.233 (0.055)

Figure: MLEs using the $SCL^{(2)}$ for various values of B.

Estimated location parameter

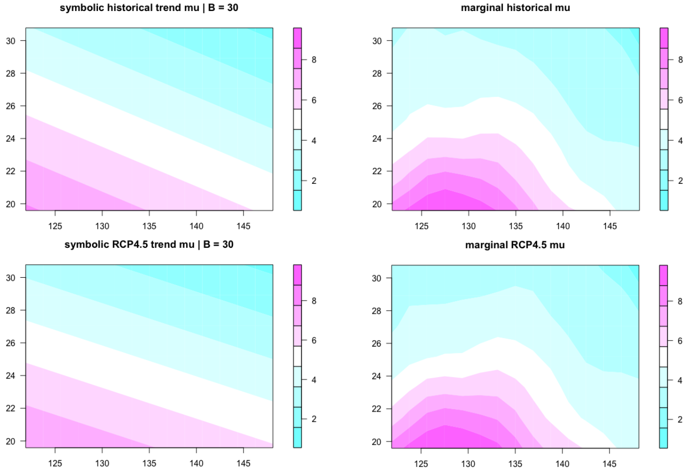


Figure: Estimated surfaces for the location parameter using the $l_{SCL}^{(2)}$ function (left) and marginal GEV estimations (right)

Examples of return level plots

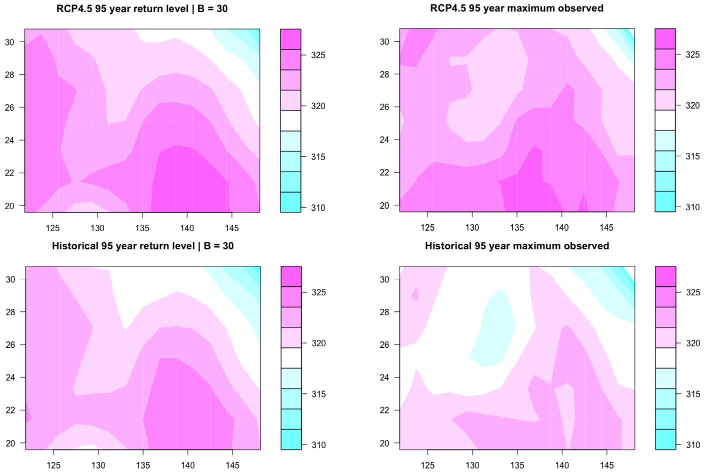


Figure: Estimated 95 year return levels using the $I^{(2)}_{SCL}$ function (left) and observed 95 year return levels (right)

Outline

Summary

- **Completely new approach to SDA:**
  - Based on fitting underlying (classical) model $\Rightarrow$ Much better!
  - View latent (classical) data through symbols
  - **Recovers existing models** for symbols but is more general
  - **Recovers classical model** as $S \to x$
  - Works for more general symbols than currently in use
  - Illustration of **practical use in extremes**
- **Working on:**
  - Characterise trade-off between accuracy and computation
  - Finalising procedure for **distribution valued symbols** (Gaussians, etc.)
  - **Design** symbols for best performance

THANK YOU!