Fitting models to underlying data using aggregates

Boris Beranger, Huan Lin, Scott Sisson, Tom Whitaker

SAMSI Program on Combinatorial Probability, 19 April 2021

What is Symbolic Data Analysis?

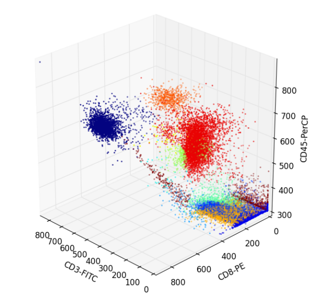
Existing and new SDA models

Statistical analyses using aggregates

- Fitting a GEV
- Meta-analyses
- Spatial Extremes
- Classification

Conclusion

Rise of non-standard data forms



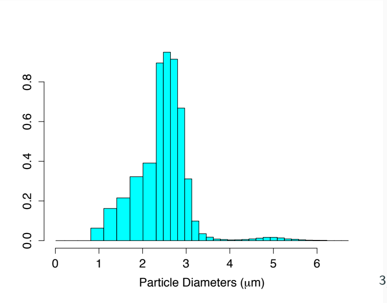
Standard statistical methods analyse classical datasets

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

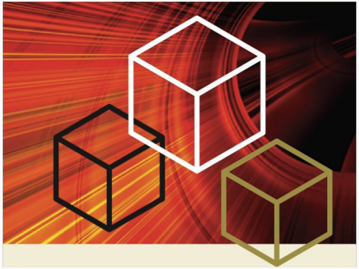
However: Increasingly see non-standard data forms for analysis.

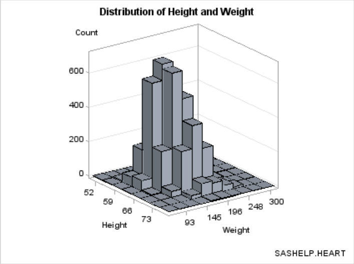
Simple non-standard forms:

- 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



Symbolic Data Analysis



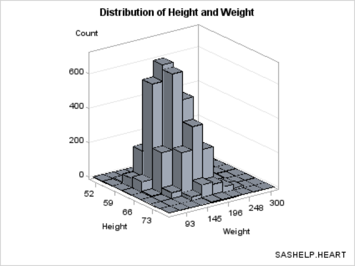


- Established by Diday & co-authors in 1990s.
- Basic unit of data is a distribution rather than usual datapoint.
  - interval (a, b)
  - p-dim hyper-rectangle
  - histogram
  - weighted list etc.
  - can be complicated by "rules"
- Classical data are special case of symbolic data:

  E.g. symbolic interval $s = (a, b)$ equivalent to classical data point x if $x = a = b$.
  Or histogram $\rightarrow \{x_i\}$ as $\#$ bins $\rightarrow \infty$.

  So 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 indentified.

- 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?

How to analyse symbolic data?

A good idea in principle, however:

- Poorly developed in terms of inferential methods.
- Current approaches:
  - Descriptive statistics (means, covariances)
    ⇒ Methods based on $1^{st}/2^{nd}$ moments: clustering, PCA etc.
  - Ad-hoc approaches (e.g. regression)
    ⇒ Can be plain wrong for inference/prediction.
  - Single technique for constructing likelihood functions
    ⇒ Limited model-based inferences
- Over-prevalence of models for intervals (a, b) & assuming uniformity
  ⇒ 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.

What is Symbolic Data Analysis?

Existing and new SDA models

Statistical analyses using aggregates

- Fitting a GEV
- Meta-analyses
- Spatial Extremes
- Classification

Conclusion

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

Symbol: $S = (S^1, \ldots, S^d)^\top$

For random intervals $[a_i, b_i]$, $i = 1, \ldots, n$: $S_i = (a_i, b_i)^\top$ or $S_i = (m_i, \log r_i)^\top$

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

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

Issues:

- Model unstable/collapses as $a_i \to b_i$ (classic data)
- How to fit equivalent models for classical data to symbols?
  - Fit to means? How to account for variation? etc.
- Symbols are summaries of classical data, $S = \pi(X_1, \ldots, X_N)$
  - Model can only predict symbols
- 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, Submitted)

Define $S = \pi(X_{1:N}) : [\mathcal{X}]^N \to \mathcal{S}$ such that $x_{1:N} \mapsto \pi(x_{1:N})$ then,

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

where

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

Gist

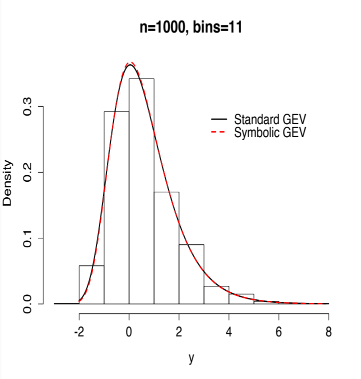
Fitting the standard classical model, when the data are viewed only through symbols S

Example: No generative model $L(x|\theta)$

- $g(S|x, \phi) = g(S|\phi) \Rightarrow L(S|\theta) = g(S|\phi)$
- Directly modelling symbol = existing likelihood approach (Le Rademacher & Billard, 2011) ✓

Modelling a histogram with random counts

Aggregation: $S = \pi(X_{1:N}) : \mathbb{R}^{d\times N} \to \mathcal{S} = \{0,\ldots,N\}^{B^1\times\cdots\times B^d}$ such that $x_{1:N} \mapsto \left(\sum_{i=1}^n \mathbb{I}\{x_i \in \mathcal{B}_1\},\ldots,\sum_{i=1}^n \mathbb{I}\{x_i \in \mathcal{B}_{\boldsymbol{B}}\}\right)$



- Assume some fixed bins $\mathcal{B}_1,\ldots,\mathcal{B}_{\boldsymbol{B}}$ and let $s = (s_1,\ldots,s_B)^\top, \sum_b s_b = n$
- If the X_i are *iid* then likelihood is multinomial:

$$L(s|\theta) \propto \frac{n!}{s_1!\ldots s_B!}\prod_{b=1}^{B} p_b(\theta)^{s_b}$$

where $p_b(\theta) \propto \int_{\mathcal{B}_b} f(z|\theta)dz$ under the model. ✓

- More complicated if data are not *iid* (Zhang, Beranger & Sisson, 2020)

Modelling a histogram with random counts

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

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

  So 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.
- Can consider histogram with random bins

Modelling a histogram with random bins

Aggregation: $S = \pi(X_{1:N}) : \mathbb{R}^{d \times N} \to \mathcal{S} = \{(a_1, \ldots, a_B) \in \mathbb{R}^B : a_1 \leq \cdots \leq a_B\} \times \mathbb{N}$ such that $x_{1:N} \mapsto (x_{(k_1)}, \ldots, x_{(k_B)}, N)$ then

$$L(s|\theta) = n! \prod_{b=1}^{B} f(s_b|\theta) \prod_{b=1}^{B+1} \frac{\left(f(s_b|\theta) - f(s_{b-1}|\theta)\right)^{k_b - k_{b-1} - 1}}{(k_b - k_{b-1} - 1)!}.$$

- Fixed $k_1, \ldots, k_B$
- When $B = 2$, $k_1 = l$ and $k_2 = u$ with $l, u = 1, \ldots, n; l \neq u$ $\Longrightarrow$ random intervals.
- Symbolic → Classical check: if $B = N \Longrightarrow L(s|\theta) = f(x|\theta)$. ✓

What is Symbolic Data Analysis?

Existing and new SDA models

Statistical analyses using aggregates

- Fitting a GEV
- Meta-analyses
- Spatial Extremes
- Classification

Conclusion

What is Symbolic Data Analysis?

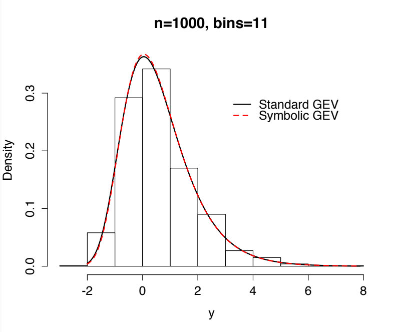
Existing and new SDA models

Statistical analyses using aggregates

- Fitting a GEV
- Meta-analyses
- Spatial Extremes
- Classification

Conclusion

Fitting a GEV



Mean MSE $\times 10^{-3}$ (1000 reps)

B	μ	σ	ξ
5	2.977	7.675	4.091
10	1.385	1.030	0.916
20	1.278	0.762	0.682
1000	1.277	0.809	0.662
Standard	1.268	0.725	0.547

- Use R's `hist` command to construct histograms, $n = 1,000$
- Use `fgev` command in `evd` package for standard approach
- Accuracy increases with more bins
- Accuracy close to using full dataset with only 20 bins
  (No real advantage to 1000 bins over 20)

Fitting a GEV

Time in seconds

n	100	1K	10K	100K	1M	10M	100M
Standard	0.018	0.047	0.431	2.860	$(*)$	$(*)$	$(*)$
Symbolic (total)	0.060	0.062	0.062	0.107	0.247	2.217	42.994
Symbolic (hist)	0.055	0.057	0.059	0.104	0.243	2.209	42.943
Symbolic (mle)	0.005	0.005	0.004	0.003	0.004	0.007	0.051

- Standard initially faster than symbolic for small datasets $\sim 1K$
- Symbolic scales much better $>$ 1K
- $* =$ `fgev` crashed on my laptop!
- However, most time for symbolic on histogram construction
- Actual symbolic optimisation super fast (obviously)
- Possible laptop caching problems around 100M
- Faster ways to construct histogram counts than `hist` for really large datasets (e.g. map-reduce using DeltaRho)

What is Symbolic Data Analysis?

Existing and new SDA models

Statistical analyses using aggregates

- Fitting a GEV
- Meta-analyses
- Spatial Extremes
- Classification

Conclusion

Estimating the sample mean from a 5-number summary

Individual studies report certain statistics: sample min (q_0), max (q_4) and quartiles (q_1, q_2, q_3)

The interest is to estimate the sample mean and variance.

- Sample mean estimator (Luo et al., 2018):

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

  where $w_1 = 2.2/(2.2 + n^{0.75})$ and $w_2 = 0.7 - 0.72/n^{0.55}$.

- Sample sd estimators (Wan et al., 2014; Shi et al., 2018):

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

  where $\zeta(n) = 2\Phi^{-1}\left(\frac{n-0.375}{n+0.25}\right)$, $\eta(n) = 2\Phi^{-1}\left(\frac{0.75n-0.125}{n+0.25}\right)$,
$\theta_1(n) = (2 + 0.14n^{0.6})\Phi^{-1}(\frac{n-0.375}{n+0.25})$ and $\theta_2(n) = (2 + \frac{2}{0.07n^{0.6}})\Phi^{-1}(\frac{0.75n-0.125}{n+0.25})$.

$\Longrightarrow$ **Each estimator assume that the data is Normally distributed!**

Estimating the sample mean from a 5-number summary

The vector $(q_0, q_1, q_2, q_3, q_4)$ corresponds to a **random bin histogram with B=4 bins**.

For simplicity let $n = 4Q + 1, Q \in \mathbb{N}$ such that $k = (1, Q + 1, 2Q + 1, 3Q + 1, n)$.

Assuming normality of the underlying data, the symbolic MLEs are
$\hat{\theta} = (\hat{\mu}, \hat{\sigma}) \approx (\bar{x}, \sqrt{(n-1)/n}s)$ which provide direct estimates $(\hat{\bar{x}}_*, \hat{s}_*) = (\hat{\mu}, \sqrt{n/(n-1)}\hat{\sigma})$.

Experiment:

- Generate data from Normal and Lognormal distributions
- Compare estimated and true sample values $(\hat{\bar{x}} - \bar{x}_0)$ and $(\hat{s} - s_0)$
- Average over $10,000$ replicates.
- Consider several n.

Normally distributed data

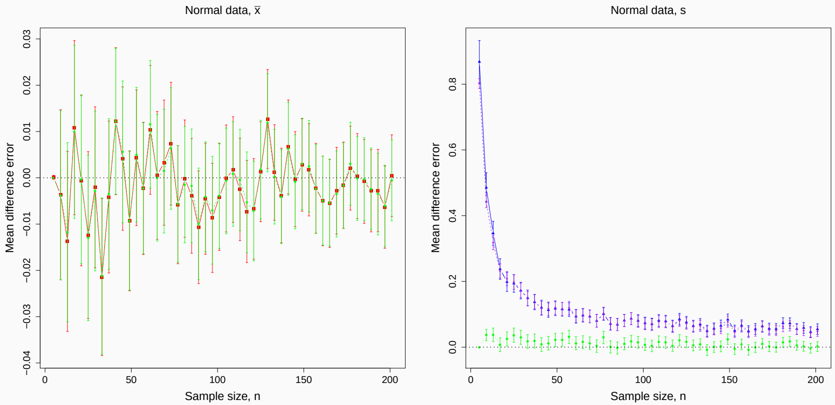


Figure 1: Mean difference errors for normally distributed data. Colouring indicates the SDA estimates (green), $\hat{\hat{x}}_L$ (red), $\hat{s}_W$ (blue) and $\hat{s}_S$ (purple). Confidence intervals indicate ±1.96 standard errors.

Lognormally distributed data

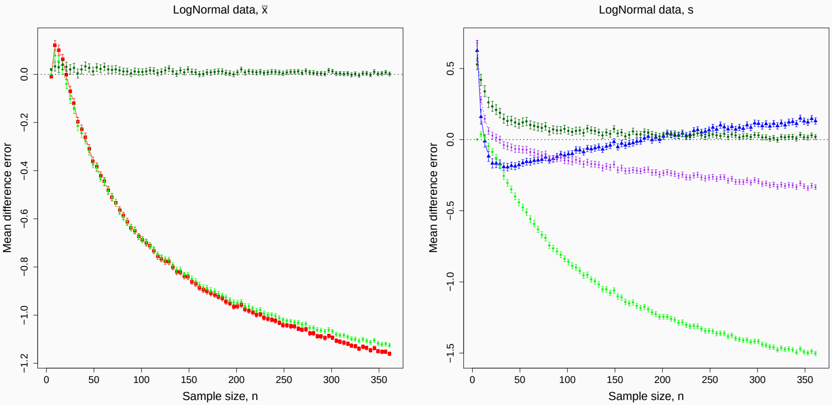


Figure 2: Mean difference errors for log-normally distributed data. Colouring indicates the SDA estimates (light and dark green), $\hat{\hat{x}}_L$ (red), $\hat{s}_W$ (blue) and $\hat{s}_S$ (purple). Confidence intervals indicate ±1.96 standard errors.

Some first steps into symbol design

The efficiency of the symbolic MLEs is clearly influenced by the form and specification of the symbol. How many bins to choose and where to put them?

Consider univariate random interval $S = (s_l, s_u, n)$ constructed using **symmetric upper and lower order statistics**, and 2-bin random histogram by including the sample median, q_2. For sample sizes $n = 4Q + 1, Q \in \mathbb{N}$, we have $l = i, u = n + 1 - i$ for the interval and $k = (i, 2Q + 1, n + 1 - i)$ for the histogram.

Experiment:

- Consider $i = 1, \ldots, 2Q$
- Draw $10,000$ datasets of size $n = 21, 81$ and 201 (i.e. $Q = 5, 20,$ 50) from a $N(\mu_0, \sigma_0)$
- Compute the rescaled symbolic MLEs $(\hat{\mu}_t, \tilde{\sigma}_t)$ where $\tilde{\sigma}_t = \sqrt{n/(n-1)}\hat{\sigma}_t$
- Calculate the relative mean square errors (RMSE) defined by

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

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

Some first steps into symbol design

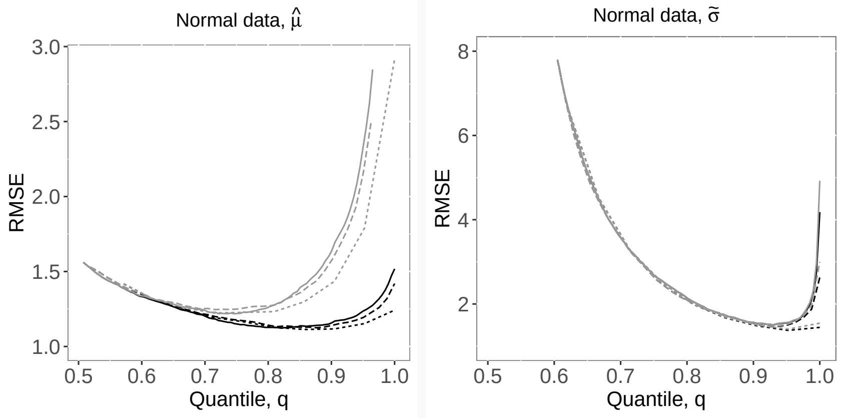


Figure 3: $\mathrm{RMSE}_{\hat{\mu}}$ (left) and $\mathrm{RMSE}_{\tilde{\sigma}}$ (right) as a function of quantile $q = (n + 1 - i)/n$ for $i = 1, \ldots, (n + 1)/2$. Grey and black lines respectively denote random intervals and histograms. Solid, long-dashed and short-dashed lines indicate samples of size $n = 21, 81$ and 201 respectively.

What is Symbolic Data Analysis?

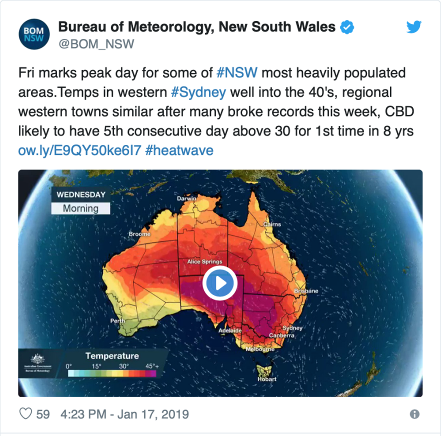
Existing and new SDA models

Statistical analyses using aggregates

- Fitting a GEV
- Meta-analyses
- Spatial Extremes
- Classification

Conclusion

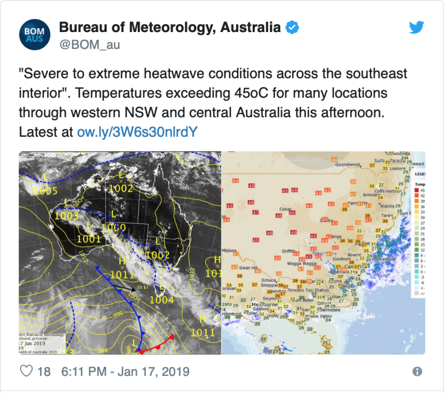
Spatial Extremes (Whitaker, Beranger & Sisson, 2020)



Bureau of Meteorology, New South Wales
@BOM_NSW

Fri marks peak day for some of #NSW most heavily populated areas.Temps in western #Sydney well into the 40's, regional western towns similar after many broke records this week, CBD likely to have 5th consecutive day above 30 for 1st time in 8 yrs ow.ly/E9QY50ke6I7 #heatwave

♡ 59 4:23 PM - Jan 17, 2019



Bureau of Meteorology, Australia
@BOM_au

"Severe to extreme heatwave conditions across the southeast interior". Temperatures exceeding 45oC for many locations through western NSW and central Australia this afternoon. Latest at ow.ly/3W6s30nlrdY

♡ 18 6:11 PM - Jan 17, 2019

Q: 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?

Spatial Extremes (Whitaker, Beranger & Sisson, 2020)

- Max-stable processes are a useful tool to analyse Spatial Extremes
- For e.g. the d.f. of the Gaussian max-stable process model

$$P(Y_1(t) \leq y_1, \ldots, Y_K(t) \leq y_K) = \exp\left\{ -\sum_{j=1}^{K} \frac{1}{y_j} \Phi_{K-1}\left(c^{(j)}(\mathrm{y}); \Sigma^{(j)} \right) \right\}$$

- The d.f. of such models becomes rapidly intractable with the number of spatial locations $\Longrightarrow$ Composite Likelihood methods (Padoan et al., 2010)
- Still unfeasible for a large number of locations and temporal observations!!
- K **dimensional histograms?!**

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

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

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**

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

when

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

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

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

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

Composite symbolic likelihoods

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

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

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

when

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

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

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

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 1: 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 2: 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

What is Symbolic Data Analysis?

Existing and new SDA models

Statistical analyses using aggregates

- Fitting a GEV
- Meta-analyses
- Spatial Extremes
- Classification

Conclusion

Classification

- $Y \in \Omega = \{1, \dots, K\}$ (response), $X \in \mathbb{R}^D$ (explanatory)

- Multinomial Logistic Regression: for realisations $\mathrm{x} \in \mathbb{R}^{D \times N}$, $y \in \Omega^N$, parameters $\beta \in \mathbb{R}^{(D+1) \times K}$, the likelihood is given by

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

  where

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

- Other model: One-vs-rest
- Prediction: $Y_n^{\mathrm{Pred}} = \arg\max_{k \in \Omega} P_{\mathrm{Model}}(Y = k | X = X_n), \forall n$
- Prediction accuracy: $PA^{\mathrm{Model}} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{Y_n^{\mathrm{Pred}} = Y_n\}$

Classification

- Let $\mathrm{X}^{(k)} = (X_n | Y_n = k, n = 1, \ldots, N) \in \mathbb{R}^{D \times N_k}$
- If $N_k = \sum_{n=1}^{N} \mathbb{1}\{Y_n = k\}$ is huge then $\mathrm{X}^{(k)}$ can be aggregated
- Histogram-valued symbol leads to likelihood

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

- Statistical improvement: mixture symbolic and classical contributions
- Computational improvements: Composite Likelihood (again!) **but** requires some adjustment.

Classification - Example

- Use a Supersymmetric (SUSY) benchmark dataset which consists of:
  - Binary response ($K = 2$): signal process (which produces supersymmetric particles) vs background process
  - $N = 5$ million observations
  - $D = 18$ features (8 kinematic properties, 10 functions)
- Comparison with optimal sub-sampling method (Wang et al., 2018 JASA)
- Training data: 4 500 000 obs.
- Test data: 500 000 obs.
- We consider the following:
  - One-vs-Rest model
  - Marginal composite likelihood
  - Histogram with random bins $L_{\mathrm{OO}}^{(1)}$
  - Histogram with random counts $L_{\mathrm{SO}}^{(1)}$

Classification - Example

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

Table 3: Prediction accuracies percentage (computing time in seconds) on the Supersymmetric dataset using histograms with B bins per margins.

- Wang et al. (2018) obtain a prediction accuracy of 78.2 with a computation time of 86.1 seconds.
- Simulation study: as good or better prediction accuracy, shorter computation time
- Sub-sampling will produce better MSE of the regression coefficients.

What is Symbolic Data Analysis?

Existing and new SDA models

Statistical analyses using aggregates

- Fitting a GEV
- Meta-analyses
- Spatial Extremes
- Classification

Conclusion

Summary

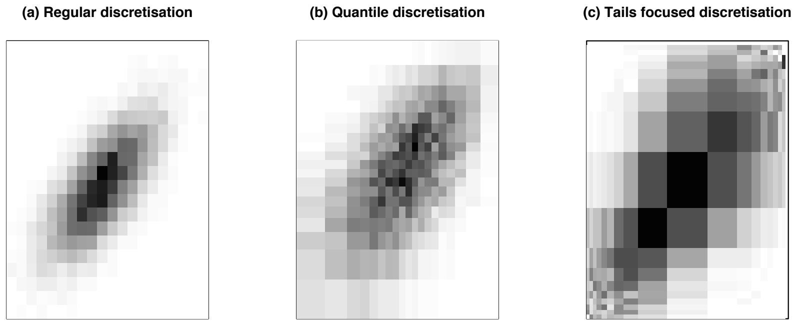
Completely new approach to SDA:

- Based on fitting underlying (classical) model
    - Radically different approach to existing SDA methods
- Views latent (classical) data through symbols
- Recovers known (some of the) existing models for symbols but is more general
- Works for more general symbols than currently in use
- Other works: internet traffic data (Rahman, Beranger & Sisson, 2020)

Still to do/Working on:

- Properties of symbolic based estimators (Prosha's PhD thesis)
- Implement more sophisticated statistical techniques using Symbols
- Characterise impact of using symbols on accuracy
    - Trade-off of accuracy vs computation
- Design of symbols for best performance
    - Histogram setting: How many bins? Bin locations?

How to design symbolic data?



How to design symbols to most efficiently represent dataset without (much) loss of critical information?

E. g. Linear regression with 10 million datapoints.

THANK YOU

Manuscripts:

- New models for symbolic data. Beranger, Lin & Sisson. https://arxiv.org/pdf/1805.03316.pdf.
- Composite likelihood methods for histogram-valued random variables. Whitaker, Beranger & Sisson (2020). *Stats & Computing*, **30**, pp.1459-1477.
- Logistic regression models using aggregated data. Whitaker, Beranger & Sisson. Whitaker, Beranger & Sisson (2021). *JCGS*, in press.

Contact:

B.Beranger@unsw.edu.au