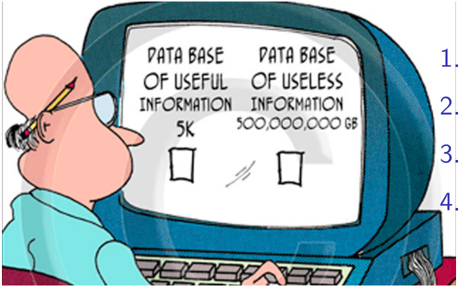
New Models for Symbolic Data

Boris Beranger, Jaslene Lin
Tom Whitaker, Scott A. Sisson,

UNSW & ACEMS

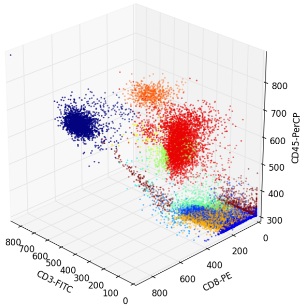
ANU, 28th February, 2019

Talk Outline



1. What is Symbolic Data Analysis?
2. Existing and new SDA models
3. Examples
4. Discussion

Rise of non-standard data forms



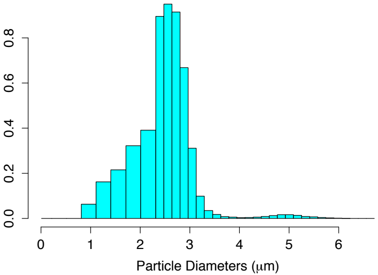
Standard statistical methods analyse classical datasets

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

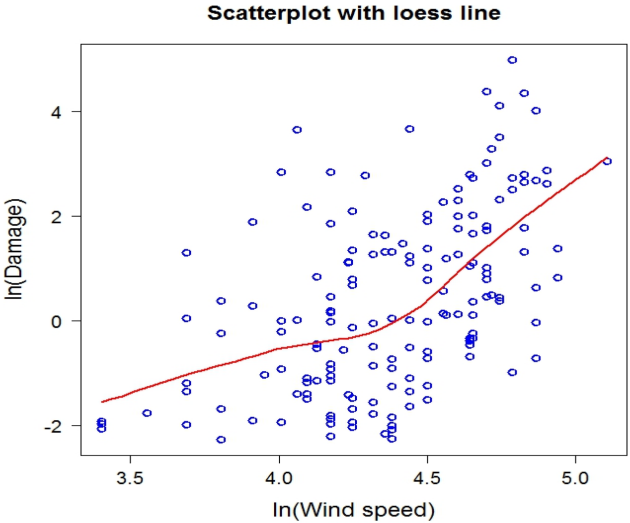
However: Increasingly see non-standard data forms for analysis.

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

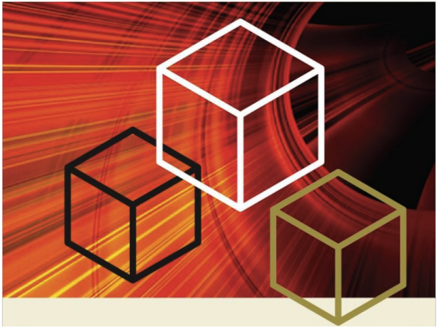


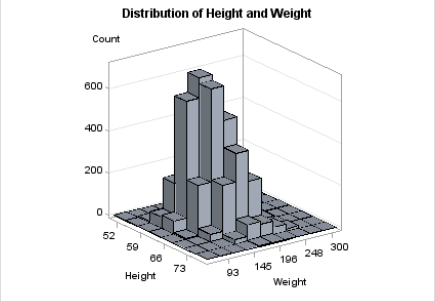
Example: Discretised data = histogram



- 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





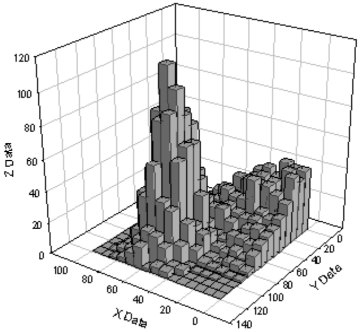
- Established by Diday & coauthors in 1990s.
- Basic unit of data is a distribution rather than usual datapoint.
  - interval (a, b)
  - p-dim hyper-rectangle
  - histogram
  - weighted list etc.
  - can be complicated by "rules"
- Classical data are special case of symbolic data:

  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!)

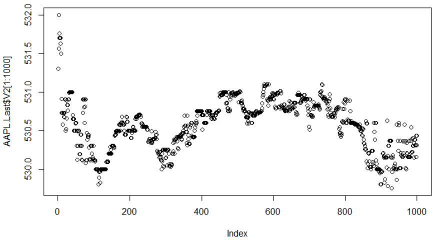
Possible use in data privacy?
Individual can't be indentified.

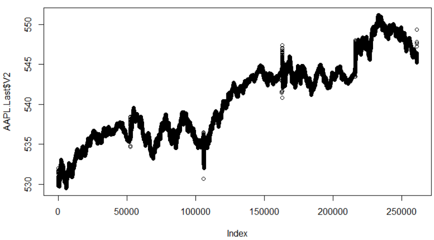
- Can arise naturally (measurement error):
  E.g. blood pressure, particulate histogram, truncation/rounding.
- 'Big Data' context:
  - Symbolic data points can summarise a complex & very large dataset in a compact manner.
  - Retaining maximal relevant information in original dataset.
  - Collapse over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

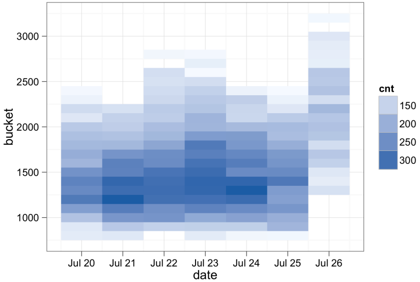
Statistical question:

How to do statistical analysis for this form of data?

Tick time series data







Too much data to analyse all ticks.

Collapse data to e.g.
one histogram per day.

Analysis of histograms now tractable.
(Though method perhaps unclear.)

In general: Reduction to symbols is question and data dependent.

How to analyse symbolic data?

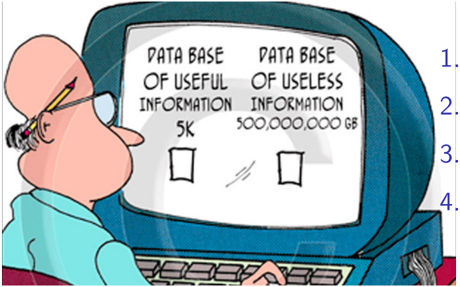
A good idea in principle, however:

- 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 (a, b) & 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.

Talk Outline



1. What is Symbolic Data Analysis?
2. Existing and new SDA models
3. Examples
4. Discussion

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

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

E.g. For random intervals $[a_i, b_i]$, $i = 1, \ldots, n$:

- $S_i = (a_i, b_i)^\top$
- $S_i = (m_i, \log r_i)^\top$

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

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

Problems:

- 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 as summaries

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 random interval

Aggregation: $S = \pi(X_{1:N}) : \mathbb{R}^N \to \mathcal{S} = \{(a_1, a_2) \in \mathbb{R}^2 : a_1 \leq a_2\} \times \mathbb{N}$
such that $x_{1:N} \mapsto (x_{(l)}, x_{(u)}, N)$.

Let $s = (s_l, s_u, n)$ with $s_l = X_{(\ell)}$, $s_u = x_{(u)}$ $\ell < u$ and $x_j \sim f(X|\theta)$:

$$\begin{aligned} L(s|\theta) &\propto \int_x g(s|x, \phi) L(x|\theta) dx \\ &= \int I(X_{(l)} = s_l \;\&\; X_{(u)} = s_u) \prod_j f(X_j|\theta) dX_{1:n} \\ &= \frac{n!}{(\ell-1)!(u-\ell-1)!(n-u)!} f(s_l|\theta) f(s_u|\theta) F(s_l|\theta)^{\ell-1} \\ &\quad \times [F(s_u|\theta) - F(s_l|\theta)]^{u-\ell-1} [1 - F(s_u|\theta)]^{n-u} \end{aligned}$$

$\Rightarrow$ the joint distribution of ℓ-th and u-th order statistics from $f(x|\theta)$. ✓

Symbolic $\rightarrow$ Classical check:
If $s_l \to s_u = x$ and $n = 1$ then $L(s|\theta) = f(x|\theta)$. ✓

Modelling a random rectangle

Aggregation 1: Marginal maxima and minima $S = \pi(X_{1:N}) : \mathbb{R}^{2\times N} \to \mathcal{S} = \{(a_1, a_2) \in \mathbb{R}^2 : a_1 \leq a_2\}^2 \times \{2, \dots, \min(4, n)\} \times \mathcal{T} \times \mathbb{N}$ such that $x_{1:N} \mapsto \left((x_{(1),i}, x_{(n),i})_{i=1,2}, p, I(p), N\right)$.

- p: number of points involved in constructing the rectangle
- $I(p)$: locations of the points (taking values in $\mathcal{T}$)

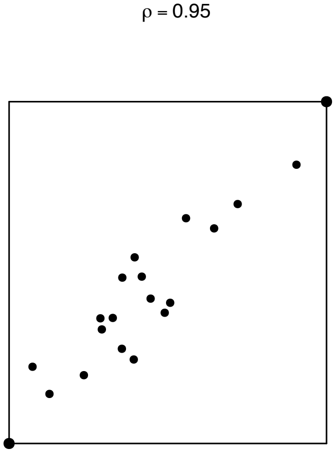
For $s = (s_{\min}, s_{\max}, s_p, s_{I_p}, n)$

$$L(s|\theta) = \frac{n!}{(n - s_p)!} \left[\int_{s_{\min}}^{s_{\max}} f(z|\theta)\mathrm{d}z\right]^{n-s_p} \times \ell_{s_p}.$$

- If $s_p = 2$ then $s_{I_p} = (s_{\min}, s_{\max})$ and $\ell_2 = f(s_{\min}|\theta)f(s_{\max}|\theta)$.

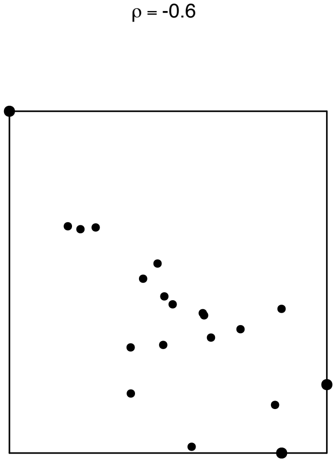
Modelling a random rectangle

Aggregation 1: Marginal maxima and minima



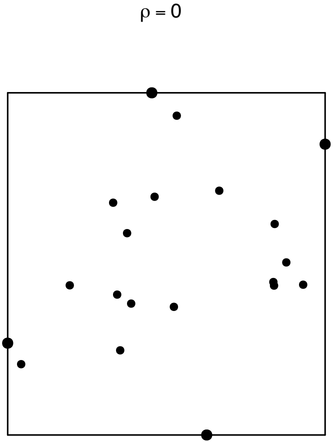
Modelling a random rectangle

Aggregation 1: Marginal maxima and minima



Modelling a random rectangle

Aggregation 1: Marginal maxima and minima



Modelling a random rectangle

Aggregation 2: Marginal order statistics
$S = \pi(X_{1:N}) : \mathbb{R}^{d\times N} \to \mathcal{S} = \{(a_1, a_2) \in \mathbb{R}^2 : a_1 \leq a_2\}^2 \times \mathbb{N}$ such that:

- [Sequential nesting]:
  $x_{1:N} \mapsto \left(\left(\left(x_{(l_i),i}, x_{(u_i),i}\right) | \{x_{(l_j),j} < x_j < x_{(u_j),j}; j < i\}\right)_{i=1,2}, N\right).$

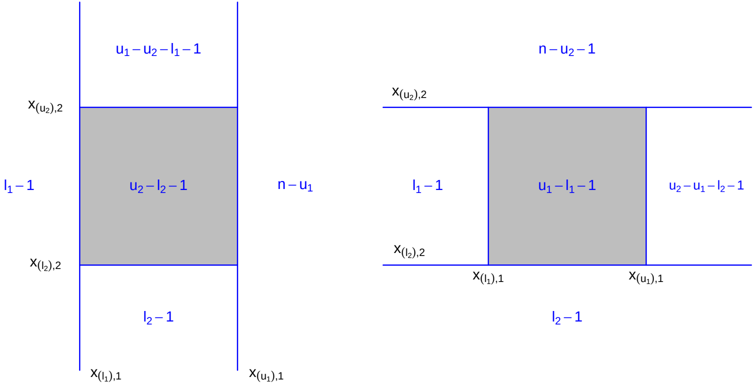
$$L(s|\theta) \propto \mathbb{P}(s_{\boldsymbol{l}} < X < s_{\boldsymbol{u}})^{u_2 - l_2 - 1} f_{X_1}(s_{l,1}) f_{X_1}(s_{u,1}) \prod_{i=1}^{2} p_i(s_{\boldsymbol{l}}) q_i(s_{\boldsymbol{u}}).$$

Modelling a random rectangle

Aggregation 2: Marginal order statistics

$S = \pi(X_{1:N}) : \mathbb{R}^{d \times N} \to \mathcal{S} = \{(a_1, a_2) \in \mathbb{R}^2 : a_1 \leq a_2\}^2 \times \mathbb{N}$ such that:

- [Sequential nesting]:



Modelling a random rectangle

Aggregation 2: Marginal order statistics

$S = \pi(X_{1:N}) : \mathbb{R}^{d\times N} \to \mathcal{S} = \{(a_1, a_2) \in \mathbb{R}^2 : a_1 \leq a_2\}^2 \times \mathbb{N}$ such that:

- [Sequential nesting]:

  $x_{1:N} \mapsto \left(\left(\left(x_{(l_i),i}, x_{(u_i),i}\right) | \{x_{(l_j),j} < x_j < x_{(u_j),j}; j < i\}\right)_{i=1,2}, N\right).$

  $$L(s|\theta) \propto \mathbb{P}(s_{\boldsymbol{l}} < X < s_{\boldsymbol{u}})^{u_2 - l_2 - 1} f_{X_1}(s_{l,1}) f_{X_1}(s_{u,1}) \prod_{i=1}^{2} p_i(s_{\boldsymbol{l}}) q_i(s_{\boldsymbol{u}}).$$

  where $p_1(s_{\boldsymbol{l}}) = F_{X_1}(s_{l,1})^{l_1 - 1}$, $q_1(s_{\boldsymbol{u}}) = (1 - F_{X_1}(s_{u,1}))^{n - u_1}$

- [Iterative segmentation]:

  $x_{1:N} \mapsto \left(\left(x_{(l_i),i} | \{x_j < x_{(l_j),j}; j < i\}, x_{(u_i),i} | \{x_j > x_{(u_j),j}; j < i\}\right)_{i=1,2}, N\right)$

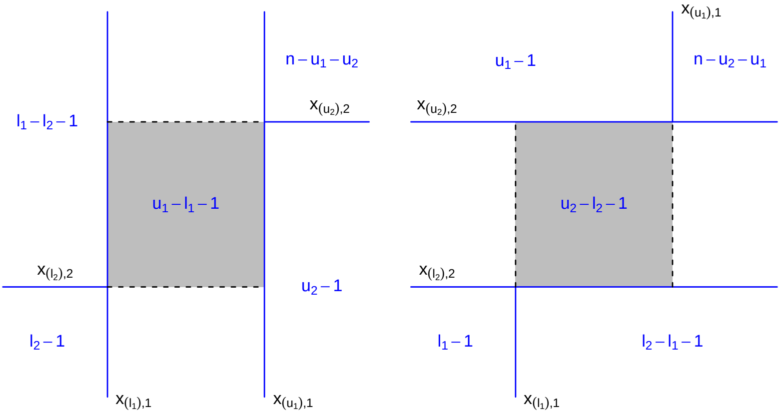
  $$L(s|\theta) \propto \mathbb{P}(s_{l,1} < X_1 < s_{u,1})^{u_1 - l_1 - 1} f_{X_1}(s_{l,1}) f_{X_1}(s_{u,1}) \prod_{i=2}^{3} p_i(s_{\boldsymbol{l}}) q_i(s_{\boldsymbol{u}}).$$

Modelling a random rectangle

Aggregation 2: Marginal order statistics

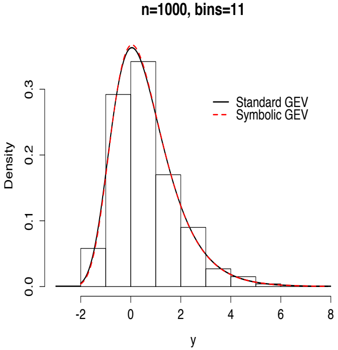
$S = \pi(X_{1:N}) : \mathbb{R}^{d\times N} \to \mathcal{S} = \{(a_1, a_2) \in \mathbb{R}^2 : a_1 \leq a_2\}^2 \times \mathbb{N}$ such that:

- [Iterative segmentation]:



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
$x_{1:N} \mapsto \left(\sum_{i=1}^{n} \mathbb{I}\{x_i \in \mathcal{B}_{\mathbf{1}}\},\ldots,\sum_{i=1}^{n} \mathbb{I}\{x_i \in \mathcal{B}_{\boldsymbol{B}}\}\right)$



- Assume some fixed bins $\mathcal{B}_{\mathbf{1}},\ldots,\mathcal{B}_{\boldsymbol{B}}$ 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, 2019)

Modelling a histogram with random counts

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

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

  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.

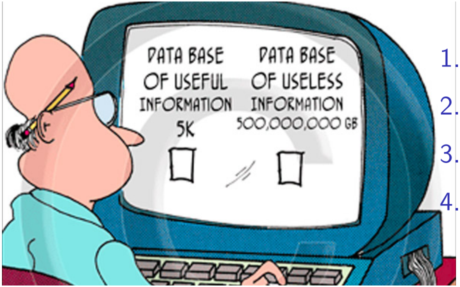
Modelling a histogram with random bins

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

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

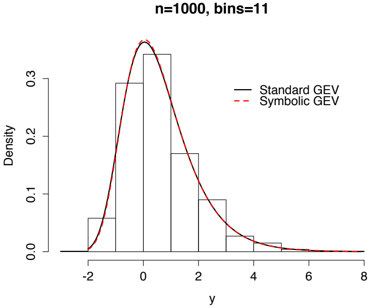
- Fixed $k_1,\ldots,k_B$
- When $B = 2$, $k_1 = l$ and $k_2 = u$ with $l, u = 1,\ldots,n; l \neq u$ $\Longrightarrow$ random intervals.
- Symbolic $\rightarrow$ Classical check: if $B = N \Longrightarrow L(s|\theta) = f(x|\theta)$. ✓

Talk Outline



1. What is Symbolic Data Analysis?
2. Some existing and new SDA models
3. Examples
4. Discussion

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)

Random rectangles

[USING MIN/MAX]

- SDA literature does not use as much information, best likelihood model:

$$L_{\emptyset}(s;\theta) = \sum_{t_p} \sum_{t_{l_p}} L_{\mathrm{full}}((s_{\min}, s_{\max}, t_p, t_{l_p}, n); \theta)\mathbb{P}(S_p = t_p, S_{l_p} = t_{l_p}; \theta),$$

- Other alternative: $L_{2d}(s;\theta)$, L_{full} with $S_p = 2d$.
- **Data:** $m = 50$ classes of $n_c = 5, 10, 50, 100$ obs from $N_2(\mu_0, \Sigma_0)$
  $\mu_0 = (2, 5)^\top$, $\mathrm{diag}(\Sigma_0) = (\sigma^2_{0,1}, \sigma^2_{0,2}) = (0.5, 0.5)$ and correlation $\rho_0 = 0, 0.5, 0.9$.
  $T = 1000$ replicates.

Random rectangles

[USING MIN/MAX]

n_c		5	10	100	1,000	100,000
$\rho_0 = 0.0$	L_4	−0.004	−0.003	0.001	0.000	0.000
		(0.056)	(0.032)	(0.015)	(0.008)	(0.004)
	$L_\emptyset$	−0.055	−0.018	−0.009	−0.005	—[a]
		(0.399)	(0.029)	(0.016)	(0.008)	—[a]
	L_{full}	−0.009	0.001	−0.001	0.011	0.000
		(0.087)	(0.082)	(0.100)	(0.108)	(0.004)
0.5	L_4	0.157	0.082	0.015	0.002	0.029
		(0.048)	(0.038)	(0.016)	(0.009)	(0.021)
	$L_\emptyset$	0.677	0.077	0.003	−0.001	—[a]
		(0.067)	(0.049)	(0.017)	(0.012)	—[a]
	L_{full}	0.508	0.503	0.494	0.488	0.327
		(0.058)	(0.055)	(0.083)	(0.076)	(0.259)
0.9	L_4	0.425	0.290	0.095	0.036	0.016
		(0.055)	(0.060)	(0.034)	(0.016)	(0.036)
	$L_\emptyset$	0.935	0.937	0.188	0.025	—[a]
		(0.010)	(0.010)	(0.355)	(0.020)	—[a]
	L_{full}	0.902	0.901	0.900	0.900	0.902
		(0.017)	(0.014)	(0.016)	(0.016)	(0.015)

Random rectangles

[USING ORDER STATISTICS]

- **Data:** $m = 20$ classes of $n_c = 60$ obs from $N_2(\mu_0, \Sigma_0)$
  $\mu_0 = (2, 5)^\top$, $\sigma^2_{0,1} = \sigma^2_{0,2} = 0.5$ and correlation $\rho_0 = 0.7$.

	Orders (l, u)	σ_1	ρ	σ_2
$L_{\mathrm{sn},x}$	$((6,5),(55,35))$	0.4992	0.6933	0.5050
		(0.0019)	(0.0255)	(0.0054)
	$((16,6),(45,24))$	0.4981	0.6402	0.5043
		(0.0021)	(0.0273)	(0.0107)
	$((20,5),(41,16))$	0.4991	0.6396	0.5054
		(0.0027)	(0.0256)	(0.0129)
$L_{\mathrm{is},x}$	$((6,3),(55,3))$	0.4993	0.7130	0.4900
		(0.0019)	(0.0067)	(0.0037)
	$((16,10),(45,2))$	0.4981	0.7037	0.4806
		(0.0021)	(0.0039)	(0.0064)
	$((20,7),(41,14))$	0.4993	0.7465	0.4871
		(0.0027)	(0.0128)	(0.0037)

- Smaller sd for first conditioned component

Random rectangles

[USING ORDER STATISTICS]

- **Data:** $m = 20$ classes of $n_c = 60$ obs from $N_2(\mu_0, \Sigma_0)$
  $\mu_0 = (2,5)^\top$, $\sigma^2_{0,1} = \sigma^2_{0,2} = 0.5$ and correlation $\rho_0 = 0.7$.

	Orders (l, u)	σ_1	ρ	σ_2
$L_{\mathrm{sn},x}$	$((6,5),(55,35))$	0.4992	0.6933	0.5050
		(0.0019)	(0.0255)	(0.0054)
	$((16,6),(45,24))$	0.4981	0.6402	0.5043
		(0.0021)	(0.0273)	(0.0107)
	$((20,5),(41,16))$	0.4991	0.6396	0.5054
		(0.0027)	(0.0256)	(0.0129)
$L_{\mathrm{is},x}$	$((6,3),(55,3))$	0.4993	0.7130	0.4900
		(0.0019)	(0.0067)	(0.0037)
	$((16,10),(45,2))$	0.4981	0.7037	0.4806
		(0.0021)	(0.0039)	(0.0064)
	$((20,7),(41,14))$	0.4993	0.7465	0.4871
		(0.0027)	(0.0128)	(0.0037)

- is provide more information about joint upper and lower values

Peer-to-peer loan data

- Data from the U.S. peer-to-peer lending company *LendingClub* available from the Kaggle platform (`https://www.kaggle.com/wendykan/lending-club-loan-data`)
- $887,373$ loans issued during 2007–2015
- Grade, from A1 (least risky) to G5 (most risky), based on risk and market conditions, which defines the interest rate

Goal: examine the link between the borrower's log annual income (in $US) and loan grade

- Analysis on full data, using reference SDA technique (LRB) and ours
- **Aggregation** of income data per risk group into a 5-bin histogram
- Models: $X_{ij} \sim N(\mu_i, \sigma_i^2)$ and $X_{ij} \sim SN(\mu_i, \sigma_i^2, \gamma_i)$
  - $\mu_i \sim T_3(c_0 + c_1 i + c_2 i^2, \tau^2)$
  - $\sigma_i^2 \sim IG(\alpha, \beta)$
  - $\gamma_i \sim N(\eta, \epsilon)$

Peer-to-peer loan data

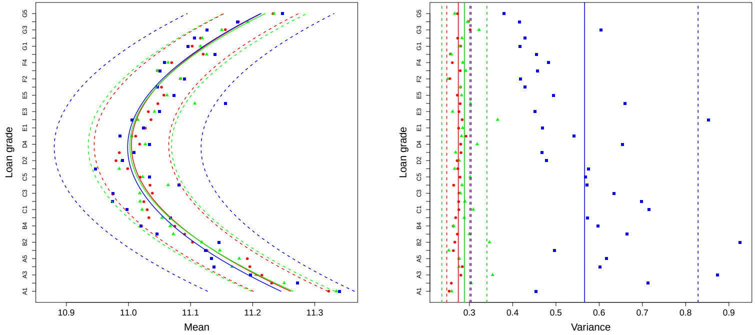


Figure: Fitted group means and variances (solid lines) when the underlying distribution is Normal, using the classical (red) and symbolic (green) likelihoods and the LRB model (blue). Dashed lines indicate pointwise 95% confidence intervals. Points denote $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ under the classical and symbolic models, and the sample mean and variance of each grade histogram for the LRB model.

Peer-to-peer loan data

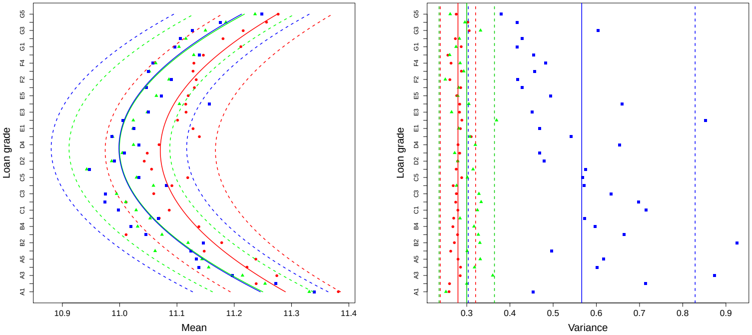


Figure: Fitted group means and variances (solid lines) when the underlying distribution is Skew-Normal, using the classical (red) and symbolic (green) likelihoods and the LRB model (blue). Dashed lines indicate pointwise 95% confidence intervals. Points denote $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ under the classical and symbolic models, and the sample mean and variance of each grade histogram for the LRB model.

Peer-to-peer loan data

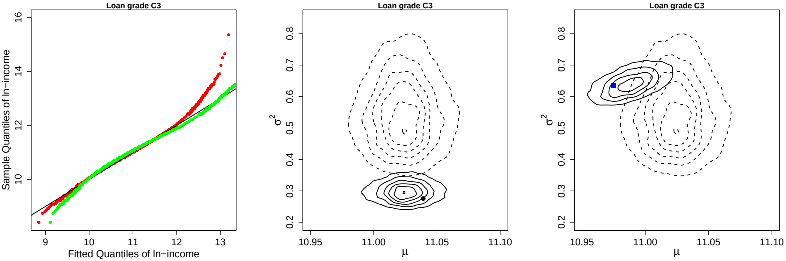
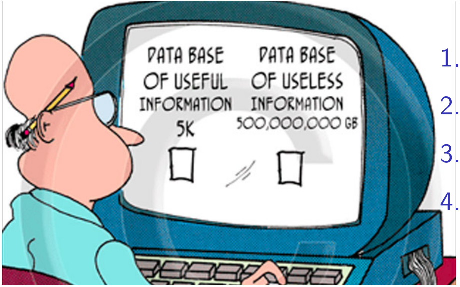


Figure: Predictive inference for loan grade C3 ($n_{C3} = 50,161$).

Table: Mean (s.e.) likelihood evaluation times (seconds $\times 10^{-3}$).

	Normal	Skew-Normal
Classical	3.886(0.478)	90.754(0.097)
New Symbolic	1.551(0.045)	12.721(0.034)
LRB	0.498(0.001)	0.476(0.001)

Talk Outline



1. What is Symbolic Data Analysis?
2. Some existing and new SDA models
3. Examples
4. Discussion

Summary

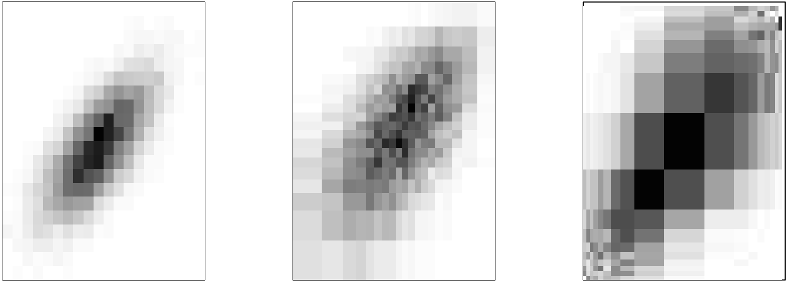
Completely new approach to SDA:

- Based on fitting underlying (classical) model
  - Radically different approach to existing SDA methods
  - Ours is much better!
- Views latent (classical) data through symbols
- Recovers known existing models for symbols but is more general
- Works for more general symbols than currently in use

Still to do/Working on:

- Implement more sophisticated statistical techniques using Symbols (Tom's PhD)
- 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?



(a) Regular discretisation (b) Quantile discretisation (c) Tails focused discretisation

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.
- A composite likelihood based approach for max-stable processes using histogram-valued variables. Whitaker, Beranger & Sisson. In prep.

Contact:

B.Beranger@unsw.edu.au