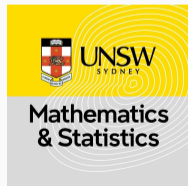


Using symbolic data to understand underlying data behaviour

Boris Beranger, Huan Lin, Scott Sisson, Tom Whitaker

One World YoungStats webinar, 08 November 2021



Motivation

[YoungStats blog post](#) on Advancement in Symbolic Data Analysis:

“With the development of digital systems, very large datasets have become routine. However, standard statistical approaches do not have the power or flexibility to analyse these efficiently, and extract the required knowledge.”

Big data → small (symbolic) data

Statistical questions:

- How to **summarise a complex & very large dataset** in a compact manner while retaining maximal relevant information in original dataset?
- How to do **statistical analysis** using symbolic data?

Useful for: Data storage, computational efficiency, data privatisation, data with non-standard form

A novel approach

Statistical analyses using aggregates

- Meta-analyses

- Classification

Conclusion

One possible approach (Beranger, Lin & Sisson, Submitted)

Define $S = \pi(X_{1:N}) : [\mathcal{X}]^N \rightarrow \mathcal{S}$ 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

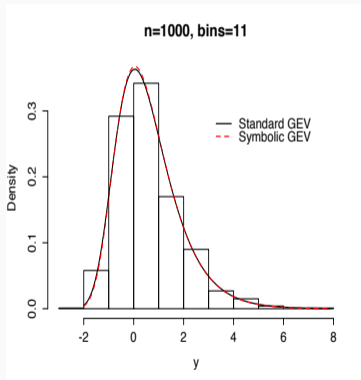
Fitting the standard classical model, when the data are viewed only through symbols S

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: $S = \pi(X_{1:N}) : \mathbb{R}^{d \times N} \rightarrow \mathcal{S} = \{0, \dots, N\}^{B^1 \times \dots \times B^d}$ such that
 $x_{1:N} \mapsto (\sum_{i=1}^n \mathbb{I}\{x_i \in \mathcal{B}_1\}, \dots, \sum_{i=1}^n \mathbb{I}\{x_i \in \mathcal{B}_B\})$



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

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

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)

Example - a histogram with random counts

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

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

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**

A novel approach

Statistical analyses using aggregates

Meta-analyses

Classification

Conclusion

A novel approach

Statistical analyses using aggregates

Meta-analyses

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{x}_L
- Sample sd estimators (Wan et al., 2014; Shi et al., 2018): \hat{s}_W and \hat{s}_S .

⇒ **Each estimator assume that the data is Normally distributed!**

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

Estimating the sample mean from a 5-number summary

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

Normally distributed data

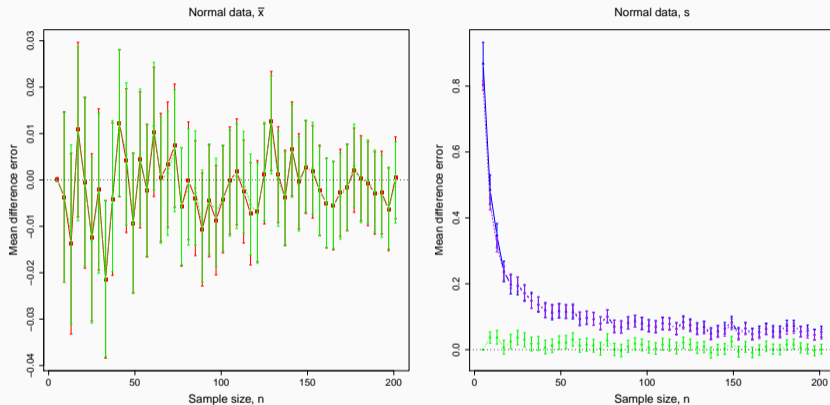


Figure 1: Mean difference errors for normally distributed data. Colouring indicates the SDA estimates (green), \hat{x}_L (red), \hat{s}_W (blue) and \hat{s}_S (purple). Confidence intervals indicate ± 1.96 standard errors.

Lognormally distributed data

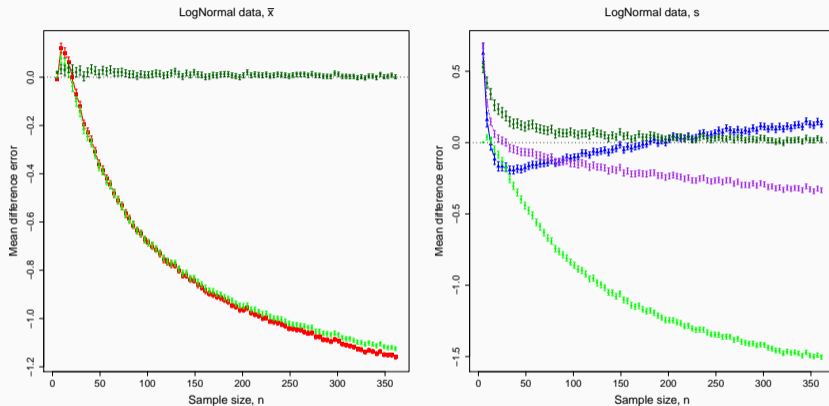


Figure 2: Mean difference errors for log-normally distributed data. Colouring indicates the SDA estimates (light and dark green), \hat{x}_L (red), \hat{s}_W (blue) and \hat{s}_S (purple). Confidence intervals indicate ± 1.96 standard errors.

Some first steps into [symbol design](#) in (Beranger, Lin & Sisson, Submitted).

A novel approach

Statistical analyses using aggregates

Meta-analyses

Classification

Conclusion

Classification

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

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

where

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

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

Classification

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

$$L_{\text{SM}}(\mathbf{s}; \beta) \propto \prod_{k \in \Omega} \prod_{b_k=1_k}^{B_k} \left(\int_{\mathbf{r}_{b_k}} P_{\text{M}}(Y = k | X = x) dx \right)^{s_{b_k}}$$

- **Statistical improvement:** mixture symbolic and classical contributions
- **Computational improvements:** Composite Likelihood (based on Whitaker, Beranger & Sisson, 2020) **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:
 - Marginal composite likelihood
 - Histogram with random counts $L_{SO}^{(1)}$

Classification - Example

Likelihood	Bins						
	6	8	10	12	15	20	25
$L_{SO}^{(1)}$	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 1: 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.

A novel approach

Statistical analyses using aggregates

- Meta-analyses

- Classification

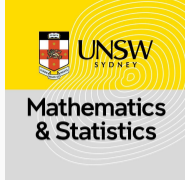
Conclusion

Completely new approach to SDA:

- Based on fitting underlying (classical) model
- 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





Still to do/Working on:

- Properties of symbolic based estimators (Prosha Rahman'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?



THANK YOU

Manuscripts:

-  New models for symbolic data. Beranger, Lin & Sisson.
-  Logistic regression models using aggregated data. Whitaker, Beranger & Sisson (2021). *JCGS*, in press.
-  Composite likelihood methods for histogram-valued random variables. Whitaker, Beranger & Sisson (2020). *Stats & Computing*, **30**, pp.1459-1477.
-  Likelihood-based inference for modelling packet transit from thinned flow summaries. Rahman, Beranger, Roughan & Sisson.

Contact:

 @borisberanger

 B.Beranger@unsw.edu.au

 www.borisberanger.com