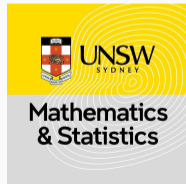


Logistic Regression Models for Aggregated Data

Boris Beranger, Tom Whitaker, Scott Sisson,

RSFAS Summer Research Camp, 2 December 2022



Big data \longrightarrow small (symbolic) data

General 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

In this talk

- Large datasets are aggregated into histograms.
- Use these summaries in order to fit a logistic regression at the underlying data level.

A possible approach to modelling aggregated data

Logistic regression using aggregates

Discussion

One possible approach to modelling aggregated data (Beranger, Lin & Sisson, 2022)

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

$$L(S|\theta) \propto \int_{\mathcal{X}} g(S|x, \phi) L(x|\theta) dx$$

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

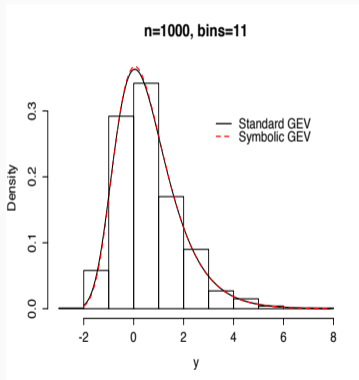
Fitting the standard classical model, when the data are viewed only through *symbols* S

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) ✓

Random count histogram

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)

Random count histogram

- 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.

Random bin histogram

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

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

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

A possible approach to modelling aggregated data

Logistic regression using aggregates

Discussion

Classification - classical data

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

Multinomial Logistic Regression

Consider realisations $\mathbf{x} \in \mathbb{R}^{D \times N}$, $y \in \Omega^N$, parameters $\beta \in \mathbb{R}^{(D+1) \times K}$.
The **standard classical likelihood** is given by

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

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 logistic regression.

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 - aggregated data

- 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

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

- **Statistical improvement:** mixture symbolic and classical contributions
- **Computational improvements:** Can the above integral be easily computed? \implies **Composite Likelihood** (based on Whitaker, Beranger & Sisson, 2020) **but** requires some adjustment.

Classification - aggregated data

Interesting result on existence of MLEs

Standard logistic regression:

- 1) $\hat{\beta} = \arg \max L(\mathbf{Y}, \mathbf{X}|\beta)$ exists and is unique if there is neither complete nor quasi-complete separation of the data (Albert and Anderson, 1984).

Histogram-based logistic regression:

- 2) $\hat{\beta}_S = \arg \max L_S(\mathbf{S}|\beta)$ exists and is unique if the set of histograms $(\mathbf{S}^1, \dots, \mathbf{S}^K)$ does not exhibit complete nor quasi-complete separation¹ of the data (Whitaker, Beranger & Sisson, 2021).

- 2) is stronger than 1), so $2) \rightarrow 1)$
- So if $\nexists \hat{\beta}_S \Rightarrow \nexists \hat{\beta}$ (i.e. if $\exists \hat{\beta} \Rightarrow \exists \hat{\beta}_S$)
- However $\hat{\beta}_S$ can exist where $\hat{\beta}$ does not
- So can do something in SDA that you can't with classical data (Is this useful?)

¹For modified definitions of separation compared to Albert and Anderson (1984)

Classification - aggregated data

Likelihood evaluation requires to compute

$$\int_{\mathbf{r}_{b_k}} P_M(Y = k | X = x) dx$$

Our options:

1. Need to do d -dimensional computational integration
2. ... **abuse** ideas from composite symbolic likelihoods
 - We can integrate $L_S(\mathcal{S}|\beta)$ for univariate predictor;
 - **Construct composite likelihood over all univariate predictor likelihoods;**
 - Or over all 2-dimensional predictor likelihoods, etc.

The Good:

- Gets around high-dimensional integration (1-d is particularly good)

Classification - aggregated data

The Bad:

- Each marginal event is **not** an unbiased estimating equation
- So this is not a “true” composite likelihood
- The estimates of β will be biased
- All parameters depressed $\beta \downarrow 0$ (known result)

The (partial) Fix:

- Can reduce the bias using some modifications to this composite likelihood following ideas in a related context by Cramer (2007)
- Does **not** eliminate it
- However prediction can still be good if reduction in $\beta \downarrow 0$ is similar for all parameters
- This is what we found to happen in practice

Composite symbolic likelihood

- Assume the interested is in a subset of size j of the K dimensions.
- Let \mathbf{b}^i be the subset of \mathbf{b} defining the coordinates of a j —**dimensional histogram bin** and let $\mathbf{B}^i = (B^{i_1}, \dots, B^{i_j})$ be the vector of the number of marginal bins.
- The symbolic likelihood function associated with the vector of counts $\mathbf{s}_j^i = (s_{1^i}^i, \dots, s_{B^i}^i)$ of length $B^{i_1} \times \dots \times B^{i_j}$ is

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

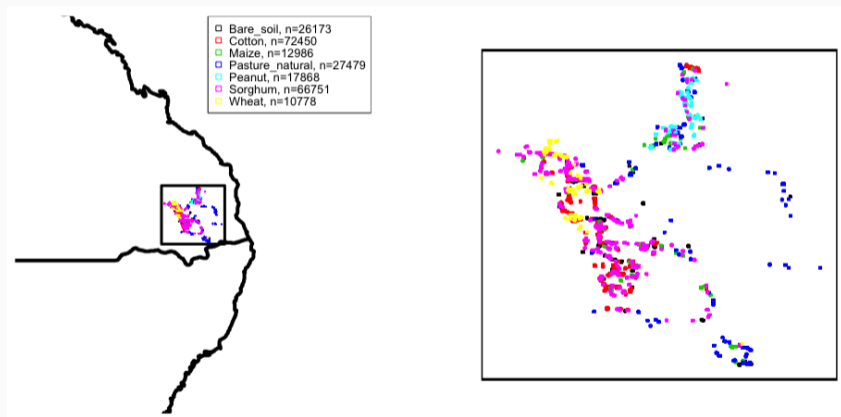
where $P_{b^i}(\theta) = \int_{\gamma_{b_{i_1}}^{i_1}} \dots \int_{\gamma_{b_{i_j}}^{i_j}} g_X(x; \theta) dx$ and g_X is a j —dim density.

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

$$L_S^{(j)}(\mathbf{s}_j; \theta) = \prod_{t=1}^T \prod_i L(\mathbf{s}_{j_t}^i; \theta)$$

Classification - Example 1

Prediction of crop types from satellite images



- $\sim 250K$ pixels with 7-dimensional predictor variable x
- 7 response categories with known ground truth
- Use (modified) symbolic composite likelihood over all 1-D predictors

Classification - Example 1

Crop type	N_k	Bins						$L_M(\mathbf{X}, \mathbf{Y} \beta)$
		6	8	10	12	15	20	
Cotton	72 450	90.5	90.6	92.8	93.6	94.0	94.1	92.2
Sorghum	66 751	74.6	74.8	75.7	76.4	76.2	76.3	80.3
Pasture Natural	27 479	75.7	75.4	76.0	76.8	77.0	77.1	77.6
Bare Soil	26 173	88.0	89.6	89.2	90.0	89.5	90.1	91.0
Peanut	17 868	81.2	81.3	81.5	81.5	81.9	81.6	82.9
Maize	12 986	9.7	9.9	10.2	10.4	10.3	10.4	14.2
Wheat	10 778	3.4	4.0	4.8	5.0	5.2	5.7	10.3
Overall	234 485	74.6	75.5	76.4	77.1	77.2	77.2	78.1
Time (secs)		(164)	(162)	(221)	(229)	(276)	(508)	(6071)

Table 1: Crop specific and overall prediction accuracies (%) using univariate marginal histograms with B bins. The likelihood optimisation times (in seconds) are reported in the last row.

- Pretty good results with 10 bins
 - Overall accuracy 76.4% (histogram) versus 78.1% (classical)
 - Poorer performance for less numerous crops (wheat, maize)
- ...and 27× faster

Classification - Example 2

- 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 $L_{\text{SO}}^{(1)}$

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 2: 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 possible approach to modelling aggregated data

Logistic regression using aggregates

Discussion

Summary

Based on a new approach to SDA:

- Aims at fitting underlying (classical) model
- Views latent (classical) data through symbols
- Logistic regression for large datasets as accurate as sub-sampling method but faster

Current & Future work:

- Properties of symbolic based estimators (Prosha Rahman's PhD thesis)

Properties of symbolic based estimators

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

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

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

when

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

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

$$\sqrt{N} (\hat{\theta} - \theta) \rightarrow \mathcal{N}(??(\theta, Bins), ??(\theta, Bins)^{-1}).$$

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

Properties of symbolic based estimators

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

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

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

when

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

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

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

- Similarly work in progress.

Summary

Based on a new approach to SDA:




- Aims at fitting underlying (classical) model
- Views latent (classical) data through symbols
- Logistic regression for large datasets as accurate as sub-sampling method but faster

Current & Future work:

- Properties of symbolic based estimators (Prosha Rahman's PhD thesis)
- Design of symbols for best performance (Hakiim Jamaluddin's PhD thesis)
 - Histogram setting: How many bins? Bin locations?
- More general symbols
- Characterise impact of using symbols on accuracy
 - Trade-off of accuracy vs computation

THANK YOU

Manuscripts:

-  New models for symbolic data. Beranger, Lin & Sisson (2022). *ADAC*, to appear.
-  Logistic regression models using aggregated data. Whitaker, Beranger & Sisson (2021). *JCGS*, **30**(4), pp.1049-1067
-  Composite likelihood methods for histogram-valued random variables. Whitaker, Beranger & Sisson (2020). *Stats & Computing*, **30**, pp.1459-1477.