Logistic Regression Models for Aggregated Data

Boris Beranger, Tom Whitaker, Scott Sisson,

IFCS ,23 July 2022

Motivation

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

Conclusion

One possible approach to modelling aggregated data (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

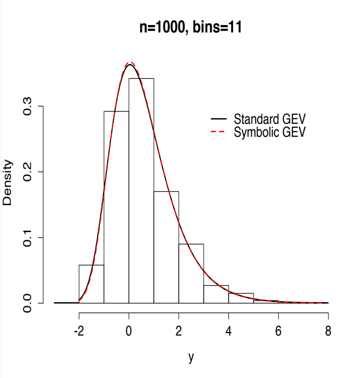
Fitting the standard classical model, when the data are viewed only through *symbols* S

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



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

Classification - classical data

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

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

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

  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 - aggregated data

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

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

- Statistical improvement: mixture symbolic and classical contributions

- Computational improvements: Composite Likelihood (based on Whitaker, Beranger & Sisson, 2020) **but** requires some adjustment.

Composite symbolic likelihood

- Assume the interested is 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.

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

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

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 $L_{\mathrm{SO}}^{(1)}$

Classification - Example

Likelihood	Bins						
	6	8	10	12	15	20	25
$L_{\mathrm{SO}}^{(1)}$	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 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 possible approach to modelling aggregated data

Logistic regression using aggregates

Conclusion

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

Future work:

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

UNSW SYDNEY Mathematics & Statistics

ACEMS

THANK YOU

Manuscripts:

- New models for symbolic data. Beranger, Lin & Sisson.
- 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.

Contact:

@borisberanger B.Beranger@unsw.edu.au www.borisberanger.com