Advances in data analysis using aggregated data

Boris Beranger, Hakiim Jamaluddin, Prosha Rahman, Scott Sisson,

CFE-CMStatistics 2024, 16 December 2024

UNSW Data Science Hub
uDASH
UNSW Sydney Mathematics & Statistics

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? What **properties** do the estimators have?

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

In this talk

1. Present a general framework for data analysis through summaries
2. Asymptotic results (Prosha's work)
3. Design of histogram aggregation functions (Hakiim's work)

A possible approach to modelling aggregated data

Asymptotic results

Design of aggregation functions

One possible approach to modelling aggregated data

(Beranger, Lin & Sisson, 2023)

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 $\Rightarrow$ (Le Rademacher & Billard, 2011)

Random bin histogram

Assume some fixed $k_1, \dots, k_B$

Aggregation:

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

Likelihood

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

Key points:

- When $B = 2$, $k_1 = l$ and $k_2 = u$ with $l, u = 1, \dots, n; l \neq u$ $\implies$ random intervals.
- Can recover classical likelihood if $B = n \implies L(s|\theta) = f(x|\theta)$.

A possible approach to modelling aggregated data

Asymptotic results

Design of aggregation functions

Convergence of summaries

Setting:

Take random intervals, i.e., random bin histogram with $B = 2$, $k_1 = l$, $k_2 = u$, and aggregation function $\pi(X_{1:n}) = (X_{(l)}, X_{(u)})$.

Things to consider:

Conditions on the sequences $1 \leq l_n \leq u_n \leq n$ are needed to ensure asymptotically nondegenerate intervals: $l_n/n \to l_0$ and $u_n/n \to u_0$.

Approach:

Order statistics can be obtained from quantiles of the empirical distribution function (van der Vaart, 1998)

Convergence of summaries

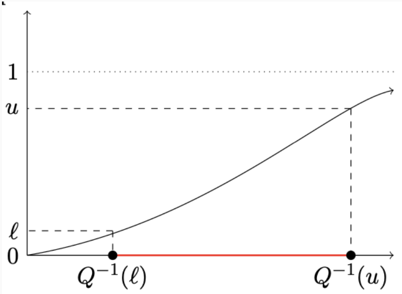
Let $Q \in \mathcal{P}$ be a continuous distribution with empirical measure μ_n

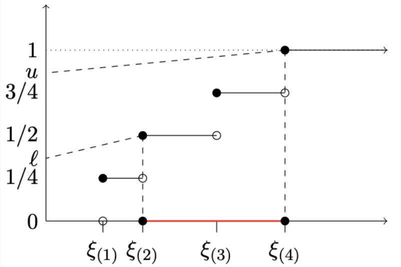
Interval-valued aggregation

Let $\boldsymbol{P} = \{(l, u) \in \mathbb{R}^2 : 0 < l \leq u < 1\}$ and $\mathbb{R}^2_{\preceq} = \{(a, b) \in \mathbb{R}^2 : a \leq b\}$

$$\begin{aligned} r : &\mathcal{P} \times \boldsymbol{P} \to \mathbb{R}^2_{\preceq} \\ &(Q, (l, u)) \mapsto \left(Q^{-1}(l), Q^{-1}(u)\right) \end{aligned}$$

Accordingly, the quantiles of μ_n are $r(\mu_n, (l, u)) = \left(X_{\lceil nl \rceil}, X_{\lceil nu \rceil}\right)$.





Convergence of summaries

Convergence

The random interval $R_n(l, u) := r(\mu_n, (l, u))$ converges uniformly in probability to $R_\infty(l, u) := r(Q, (l, u))$.

Extensions

- Random rectangle (interval-valued aggregation to $\mathbb{R}^d$):
  R_n^d converges weakly to R_∞^d
- Random histograms:
  H_n^b converges uniformly almost surely to H_∞^b.
- Two distribution-valued aggregations with similar convergence properties

Convergence of the likelihood

Denote $S_n = \pi_n(X_1, \ldots, X_n) = \pi(\mu_n)$, such that in the interval example $S_n(\omega) = R_n(l, u)(\omega)$

1 aggregate $\Rightarrow$ the limit of the likelihood $\mathcal{L}_n$ is determined by the limit of the sequence of densities f_{S_n}.

Suppose we fit the model P_θ, therefore

- $\star$ $\mu_n \to P_\theta$ weakly
- $\star$ $S_n \to \pi(P_\theta)$ in probablity

Limit likelihood

For some true $\theta_0 \in \Theta$, we then get:

$$\begin{aligned}\mathcal{L}_\infty(\theta, \omega) &= \lim_{n\to\infty} f_{S_n}(S_n(\omega)) \\ &= \lim_{n\to\infty} f_{\pi_n(X_1,\ldots,X_n)}(S_n(\omega)) \\ &= \delta(\pi(P_{\theta_0}) - \pi(P_\theta))\end{aligned}$$

Convergence of the likelihood

Convergence

1. The summary likelihood $\mathcal{L}_n \to \mathcal{L}_\infty$ uniformly in Θ in probability.
2. The MLE $\hat{\theta}_n \to \theta_0$ in probability and is a consistent estimator.

Extension

Convergence can be established for multiple data summaries (under some assumptions)

Summary

- As we get more distributions, and data per distribution, the likelihood will consistently estimate θ_0.
- Interest is now in the rate this happens (so we can design distributions with the most efficient rate).

A possible approach to modelling aggregated data

Asymptotic results

Design of aggregation functions

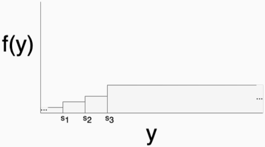
Illustrative example

Generate $1,001$ samples from $\mathcal{N}(10,1)$.

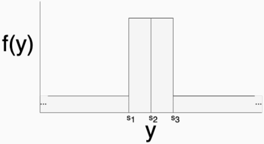
Aggregation into 4 bin histograms with bins based on order statistics

Fit the true model. Repeat $1,000$ times.

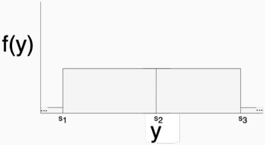
$\mathbf{k} = (100, 200, 300)$
$\mu : 9.15(1.18)$
$\sigma : 0.14(1.36)$



$\mathbf{k} = (400, 500, 601)$
$\mu : 10.00(1.81)$
$\sigma : 0.41(3.09)$



$\mathbf{k} = (100, 500, 601)$
$\mu : 10.00(0.81)$
$\sigma : 0.09(0.94)$



Statistical decision theory

Let $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ be some unknown parameter of interest and $\mathbf{d} \in \mathcal{D}$ be some decision.

A loss function $L(\boldsymbol{\theta}, \mathbf{d})$ measures the consequence of each decision $\mathbf{d}$, e.g., quadratic loss:

$$L(\boldsymbol{\theta}, \mathbf{d}) = (\boldsymbol{\theta} - \mathbf{d})^\top \mathbf{Q} (\boldsymbol{\theta} - \mathbf{d})$$

This is not available since $\boldsymbol{\theta}$ is unknown so we refer to the expected loss

In the Bayesian framework, take some prior $p(\boldsymbol{\theta})$, the best belief about the distribution of $\boldsymbol{\theta}$ is the posterior $p(\boldsymbol{\theta}|\boldsymbol{s})$.

Posterior expected loss

$$\rho(p(\theta|\mathbf{s}), \mathbf{d}) \equiv \mathbb{E}_{\theta|\mathbf{s}}\left[L(\theta, \mathbf{d})\right] = \int_{\Theta} L(\theta, \mathbf{d}) p(\theta|\mathbf{s}) \mathrm{d}\theta,$$

Statistical decision theory

Take $\mathbf{Q} = \mathbb{I}_p$, then $\mathbf{d}^* = \arg\min_{\mathbf{d} \in \mathcal{D}} \rho(p(\theta|\mathrm{y}), \mathbf{d}) = \mathbb{E}_{\theta|\mathbf{s}}[\theta]$ and

$$\rho(\pi(\theta|\mathbf{s}), \mathbf{d}^*) = \sum_{i=1}^{p} \mathbb{E}_{\theta_i|\mathbf{s}}\left[\left(\theta_i - \mathbb{E}_{\theta_i|\mathbf{s}}[\theta_i]\right)^2\right] = \sum_{i=1}^{p} \mathbb{V}_{\theta_i|\mathbf{s}}(\theta_i).$$

Optimal design

An optimal symbolic data design minimises the minimised posterior expected loss (MPEL) function, $\mathbf{s}^* = \arg\min_{\mathbf{s}} \rho(p(\theta|\mathbf{s}), \mathbf{d}^*)$

Experiment: where to put the bins of a histogram?

Setup

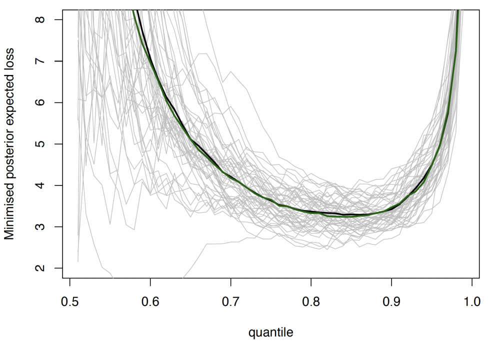
True model: $Y \sim \mathcal{N}(\mu = 50, \sigma = 17)$.

$n = 201$ observations;

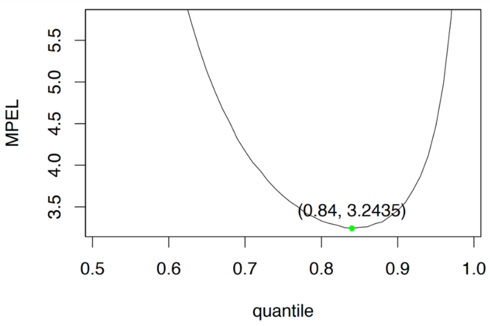
$B = 2$ (3 bins) with **symmetric** quantiles;

Compute the posterior Loss for varying quantiles;

Repeat 50 times (expensive!)



Experiment: where to put the bins of a histogram?



For a normal distribution, the suggests to use the 16 and 84% quantiles.

More (non-symmetric) quantiles:

$B = 2$	$B = 3$	$B = 4$	$B = 5$
(0.14,0.85)	(0.09,0.52,0.91)	(0.07,0.32,0.74,0.95)	(0.05,0.22,0.52,0.79,0.96)

Summary(!)

- Likelihood-based framework to fit model through summaries;
- Limit results ensuring the convergence of the summaries and the likelihood; Estimators have good properties: consistent!
- Bayesian framework for summary design.

THANK YOU

B.Beranger@unsw.edu.au

UNSW Data Science Hub uDASH

UNSW SYDNEY Mathematics & Statistics