Estimation Efficiency under a Two-Sample Density Ratio Model

Archer Gong Zhang

Department of Statistics
University of British Columbia

SSC 2022 Annual Meeting

Acknowledgement

This talk is based on the joint work with my Ph.D. supervisor:

Figure: Dr. Jiahua Chen.

Outline

Outline

Motivation

In many disciplines, data are collected as multiple samples from similar and connected populations:

$$
\begin{gathered}
x_{0,1}, x_{0,2}, \ldots, x_{0,n_0} \overset{i.i.d.}{\sim} G_0(x) \\
x_{1,1}, x_{1,2}, \ldots, x_{1,n_1} \overset{i.i.d.}{\sim} G_1(x) \\
\vdots \\
x_{m,1}, x_{m,2}, \ldots, x_{m,n_m} \overset{i.i.d.}{\sim} G_m(x),
\end{gathered}
$$

where $G_0, G_1, \ldots, G_m$ share some common features.
For example,

- in economics, scientists collect survey datasets of individual and household incomes from year to year;
- in network studies, people's activities on social networks in different periods of time are collected as multiple samples.

Example: How to analyze data look like these?

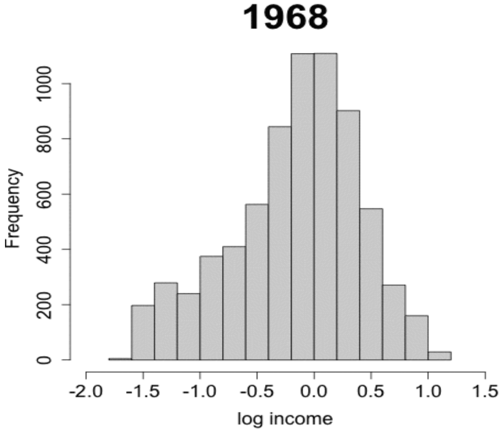

Figure: Histograms of log household relative income from 1968 to 1988. Data source: UK Family Expenditure Survey.

Different approaches to statistical analysis

Parametric approaches
Choose a suitable parametric model (e.g., normal) for each of the multiple populations
Pros: good statistical efficiency
Cons: consequence of model misspecification may be serious
No ☹️

Different approaches to statistical analysis

Parametric approaches	Nonparametric approaches
Choose a suitable parametric model (e.g., normal) for each of the multiple populations	Do not place distributional assumptions on the populations
Pros: good statistical efficiency	Pros: free from the risk of model misspecification
Cons: consequence of model misspecification may be serious	Cons: low statistical efficiency
No ☹️	No 🙁

Different approaches to statistical analysis

Parametric approaches	Nonparametric approaches	A Semiparametric approach
Choose a suitable parametric model (e.g., normal) for each of the multiple populations	Do not place distributional assumptions on the populations	Do not place parametric assumptions on each population
Pros: good statistical efficiency	Pros: free from the risk of model misspecification	**Model the connection between the multiple population distributions**
Cons: consequence of model misspecification may be serious	Cons: low statistical efficiency	A flexible & efficient compromise between parametric and nonparametric approaches
No ☹️	No 🙁	Yes! 😃

Outline

Density ratio model (DRM) [Anderson, 1979]

- $g_k(x)$: density of the kth population distribution G_k.
- DRM assumes that: for $k = 1, \ldots, m$,

$$\frac{g_k(x)}{g_0(x)} = \exp\left\{\alpha_k + \boldsymbol{\theta}_k^\top \mathbf{q}(x)\right\}$$

unknown parameters to be estimated

vector-valued function: basis function

- We call G_0 the base distribution; any G_k may serve as the base distribution.

Why DRM?

- DRM is flexible: G_0 is unspecified, allowing it to cover many distribution families.

Distribution family	q(x)
Normal	$(x,\ x^2)$
Gamma	$(x,\ \log x)$
Exponential family	Sufficient stats
...	...

- With an appropriate $\mathbf{q}(x)$, DRM allows us to use the pooled data to estimate G_k rather than use data only from G_k.

$$\Downarrow$$

gain in statistical efficiency!

Inference method on the base distribution G_0

- The base distribution G_0 is left unspecified in DRM.
- Assign a parametric distribution to $G_0 \implies$ DRM being fully parametric.
- We use a nonparametric method: the empirical likelihood (EL) [Owen, 1988].
- Owen [2001]: "EL keeps the effectiveness of likelihood methods and does not impose a known family distribution on the data".

Figure: Art B. Owen: "Yes, I said it."

Outline

Efficiency of the DRM-based estimators

- Many studies have showed that some DRM-based estimators are more efficient than the nonparametric estimators.
- Motivated by these results, we are interested in how far we can push the efficiency of the DRM-based estimators.
- A "gold standard" is the parametric estimator: estimator under a parametric model (e.g., a normal model).
- When the parametric model is correctly specified, parametric estimators (such as MLE) are usually the most efficient.
- Is it likely that the DRM-based estimators can be as efficient as the parametric estimators? Or When?

A two-sample scenario

If there are two samples from populations G_0 and G_1, and $n_0 \gg n_1$:

- The larger sample is expected to characterize the whole population G_0 with high accuracy: G_0 can be roughly seen as “known”.
- The DRM can then be regarded as a fully parametric model for G_1:

$$g_1(x) = g_0(x)\exp\{\alpha + \boldsymbol{\theta}^\top \mathbf{q}(x)\}.$$

- We therefore expect the DRM estimators for G_1 to achieve parametric efficiency.
- We study the efficiency of some estimators for G_1 when:

$$n_0/n_1 \to \infty \qquad \text{as } n_0, n_1 \to \infty.$$

A parametric model

- We consider an exponential family model for the two samples:

$$x_{0,1}, \ldots, x_{0,n_0} \overset{i.i.d.}{\sim} g_0(x) = B(x) \exp\{\boldsymbol{\eta}_0^\top \mathbf{q}(x) + A(\boldsymbol{\eta}_0)\},$$
$$x_{1,1}, \ldots, x_{1,n_1} \overset{i.i.d.}{\sim} g_1(x) = B(x) \exp\{\boldsymbol{\eta}_1^\top \mathbf{q}(x) + A(\boldsymbol{\eta}_1)\}.$$

- Recall the two-sample DRM with the same $\mathbf{q}(x)$:

$$g_1(x)/g_0(x) = \exp\{\alpha + \boldsymbol{\theta}^\top \mathbf{q}(x)\}.$$

- The DRM contains this exponential family model:

$$\begin{pmatrix} \alpha \\ \boldsymbol{\theta} \end{pmatrix} = \begin{pmatrix} A(\boldsymbol{\eta}_1) - A(\boldsymbol{\eta}_0) \\ \boldsymbol{\eta}_1 - \boldsymbol{\eta}_0 \end{pmatrix}.$$

- The MLEs under this exponential family model are the parametric estimators ("gold standard").

Theoretical results

We prove that under the two-sample scenario, the DRM-based estimators of the following parameters achieve the same asymptotic efficiency as the parametric estimators:

- Model parameters $(\alpha, \boldsymbol{\theta})$ under the DRM;
- Population distribution $G_1(x)$;
- Quantiles of G_1.

Efficiency of DRM quantile estimators: an ideal case

- We illustrate the efficiency of the DRM quantile estimator under an ideal situation:

$$G_0(x) = G_1(x).$$

- Focus on ξ_p: the pth quantile for G_1.
- Let $k = n_0/n_1$. Assuming k does not evolve with n_0, n_1, we use the result by Chen and Liu [2013] to show that in this case:

$$\underbrace{n_1 \mathrm{Var}(\hat{\xi}_p)}_{\text{DRM Var}} = \frac{1}{k+1} \underbrace{\left[\frac{p(1-p)}{g_1^2(\xi_p)}\right]}_{\text{Nonparametric Var}} + \frac{k}{k+1} \underbrace{\left[n_1 \mathrm{Var}(\tilde{\xi}_p)\right]}_{\text{Parametric Var}}.$$

Simulation with data from normal distributions

- Focus on the pth quantile for G_1.
- Both samples are generated from $N(0, 1)$.
- The DRM-based quantile estimate is obtained assuming only the knowledge of the most appropriate $\mathbf{q}(x) = (x, x^2)^\top$.
- Two competitors that only use sample from G_1 (with a smaller size):
  - MLE of quantile derived under the normal model;
  - Nonparametric empirical quantile.

Simulation results (numbers are $\times n_1$, based on 1000 repetitions)

Levels p	DRM-based		MLE		Nonparametric	
	Bias	Var	Bias	Var	Bias	Var
	$n_0 = k \times n_1,\ n_1 = 1000,\ k = 10$					
0.01	−0.02	4.91	0.03	3.81	0.51	13.53
0.05	−0.01	2.61	0.02	2.42	0.09	4.53
0.10	0.00	1.98	0.02	1.87	0.03	3.29
0.50	0.00	1.10	0.01	1.03	0.05	1.58
	$n_0 = k \times n_1,\ n_1 = 1000,\ k = 100$					
0.01	−0.06	3.94	−0.05	3.83	0.58	13.66
0.05	−0.05	2.46	−0.05	2.41	0.11	4.45
0.10	−0.05	1.86	−0.05	1.85	0.01	2.88
0.50	−0.05	0.97	−0.04	0.96	−0.04	1.52

- As k increases, the variances of the DRM estimators approach those of the MLEs.
- Our “weighted average” result is supported.

Summary

- We prove that in the two-sample scenario where $n_0/n_1 \to \infty$, some DRM estimators for G_1 achieve parametric efficiency.
- Our contribution is new and particularly useful in applications where we have one large historical sample and one small sample to make inference on.
- Simulation results on quantile estimation support our theoretical findings.

References I

J. Anderson. Multivariate logistic compounds. *Biometrika*, 66(1):17–26, 1979.

J. Chen and Y. Liu. Quantile and quantile-function estimations under density ratio model. *The Annals of Statistics*, 41 (3):1669–1692, 2013.

A. B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.

A. B. Owen. *Empirical Likelihood*. Chapman & Hall/CRC, New York, 2001.

Thank you!

We hope someday you may find DRM useful in your research! :-)