Density ratio model with data-adaptive basis function

Archer Gong Zhang

Department of Statistics
University of British Columbia

Joint work with Dr. Jiahua Chen

2021 CANSSI Showcase

Motivation

- In many applications, data are collected as multiple samples from <u>similar and connected</u> populations:

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

  where $G_0, G_1, \ldots, G_m$ share some common features.
- E.g., in economics, multiple survey datasets of individual and household incomes are collected annually.

Example: How to analyze data look like these?

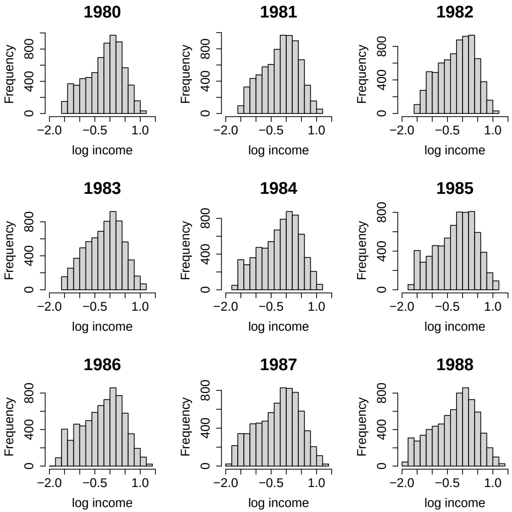

Figure: Histograms of log household relative income from 1980 to 1988. Data source: UK Family Expenditure Survey.

How to get these similar populations connected?

A semi-parametric density ratio model [Anderson, 1979]:

- no distributional assumption on each population
- model the relationship between the multiple population distributions

Density ratio model (DRM)

- $g_k(x)$: density of the k-th population distribution G_k.
- DRM assumes that:

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

  for $k = 1, \ldots, m$.
- $\mathbf{q}(x)$: a vector-valued function, called the *basis function*.
- $(\alpha_k, \boldsymbol{\theta}_k)$: unknown parameters to be estimated.

Why DRM?

- DRM covers 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.

An open problem in the use of DRM

- DRM assumes the knowledge of the basis function $\mathbf{q}(x)$.
- Complete knowledge about $\mathbf{q}(x)$ is impossible in applications.
- How to choose $\mathbf{q}(x)$ based on data remains an open problem.
- We propose a data-adaptive approach to the choice of $\mathbf{q}(x)$, which helps alleviate the risk of model misspecification.

A closer look at the basis function

- Re-write the DRM assumption as:

$$Q_k(x) = \log \frac{g_k(x)}{g_0(x)} = \alpha_k + \boldsymbol{\theta}_k^\top \mathbf{q}(x),$$

  for $k = 0, 1, \ldots, m$,
- $Q_0(x), \ldots, Q_m(x)$ are all linear combinations of $\mathbf{q}(x)$.
- Intuitively, it is appropriate to form $\mathbf{q}(x)$ by the dominant modes of variation of $Q_0(x), \ldots, Q_m(x)$.

Functional principal component analysis (FPCA)

- FPCA is a dimension reduction technique on functional data, in our case, $\{Q_0(x), \ldots, Q_m(x)\}$.
- Via FPCA, $Q_0(x), \ldots, Q_m(x)$ can be represented by some functional principal components (FPCs):

$$Q_k(x) - \frac{1}{m+1}\sum_{r=0}^{m} Q_r(x) = \sum_{j=1}^{d} \beta_j^k \psi_j(x).$$

- FPCs $\psi_1(x), \ldots, \psi_d(x)$ are the dominant modes of variation among $Q_0(x), \ldots, Q_m(x)$.

Estimation of the FPCs

Estimate of density: $\hat{g}_k(x)$

- Via kernel density estimation

Estimate of $Q_k(x)$: $\hat{Q}_k(x)$

- Via $\hat{Q}_k(x) = \log \dfrac{\hat{g}_k(x)}{\hat{g}_0(x)}$

Estimates of FPCs: $\{\hat{\psi}_1(x), \dots, \hat{\psi}_d(x)\}$

- Via linear algebra

Data-adaptive basis function

- Use these estimated FPCs to form the data-adaptive $\mathbf{q}(x)$:

$$\hat{\mathbf{q}}(x) = (\hat{\psi}_1(x), \ldots, \hat{\psi}_d(x)).$$

- In applications, we do not know what d is most appropriate.
- In our paper [Zhang and Chen, 2021], we suggest some adaptive ways to choose d.

UK household income data

- We consider a survey data from the Family Expenditure Survey in UK, from 1968 to 1988. (accessible on `https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=200016`)
- The data contain yearly samples on the incomes and expenditures of > 7000 households (HHs) each year.
- Variable of interest: log-transformed HH relative income.

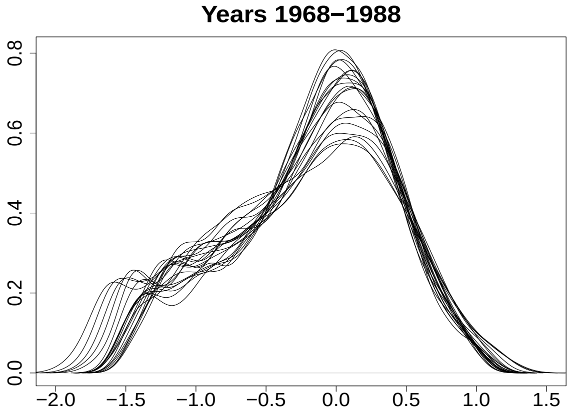

Figure: Kernel density estimators based on HH relative income data.

Apparently, there is some connection between these distributions.

Real-data based simulation procedure

We study the empirical likelihood [Owen, 2001] based quantile estimation under the DRM [Chen and Liu, 2013].

Data from 1968–1981: training data	**Data from 1982–1988: test data**
obtain the adaptive q(x)	create multiple samples by sampling with sizes 1000
	fit the DRM to these multiple samples with the adaptive q(x)
	obtain the DRM-based estimates
	repeat for 1000 times

Performance of some quantile estimators

Simulated mean squared errors (MSEs) of the 10th, 30th, 50th, 70th, and 90th percentiles, averaged across the years 1982–1988.

Method	Average MSE ($\times$1000) of quantile estimators					
	10%	30%	50%	70%	90%	avg.
FPC-1	1.86	0.62	0.37	0.16	0.31	0.66
FPC-2	1.43	0.68	0.44	0.22	0.40	0.63
Adaptive	1.43	0.69	0.44	0.22	0.40	0.64
NP	1.78	1.41	0.84	0.57	0.67	1.05

- The proposed "Adaptive" estimators perform well, with a $\sim$ 39% gain in efficiency compared to the "NP" estimators.
- Our suggested adaptive approach usually selects $d = 2$ FPCs, which is also the best-performing d (FPC-2 in the Table).

Conclusions

- DRM with the proposed data-adaptive $\mathbf{q}(x)$ leads to efficiency gain.
- Our contribution gives users confidence in the validity and the effectiveness of data analysis via DRM.
- Other DRM-based inferences using the adaptive $\mathbf{q}(x)$ can be similarly developed.

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*. Chapman & Hall/CRC, New York, 2001.

A. G. Zhang and J. Chen. Density ratio model with data-adaptive basis function. *arXiv preprint arXiv:2103.03445*, 2021.

Thank you!

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