Learning the basis function in a semi-parametric density ratio model

Archer (Gong) Zhang

Department of Statistics
University of British Columbia

Joint work with Dr. Jiahua Chen

JSM 2020 Virtual Conference

Outline

Motivation

A semiparametric model: Density Ratio Model

A non-parametric inference method: Empirical Likelihood

How does the EL-DRM framework work in quantile estimation?

Our contribution

Real Data

Outline

Motivation

A semiparametric model: Density Ratio Model

A non-parametric inference method: Empirical Likelihood

How does the EL-DRM framework work in quantile estimation?

Our contribution

Real Data

Motivation

In many research disciplines, data are collected as multiple samples from similar populations.
For example,

- in the field of family studies, sociologists collect survey data on the social and economic characteristics of families (e.g., family incomes, marriage status) from years to years;
- in educational studies, students' performances are monitored from time to time, in the form of multiple samples;
- in social media-related studies, people's activities on Facebook or Twitter in different periods of time are naturally collected as multiple samples;
- etc...

Example: How to analyze data looks like this?

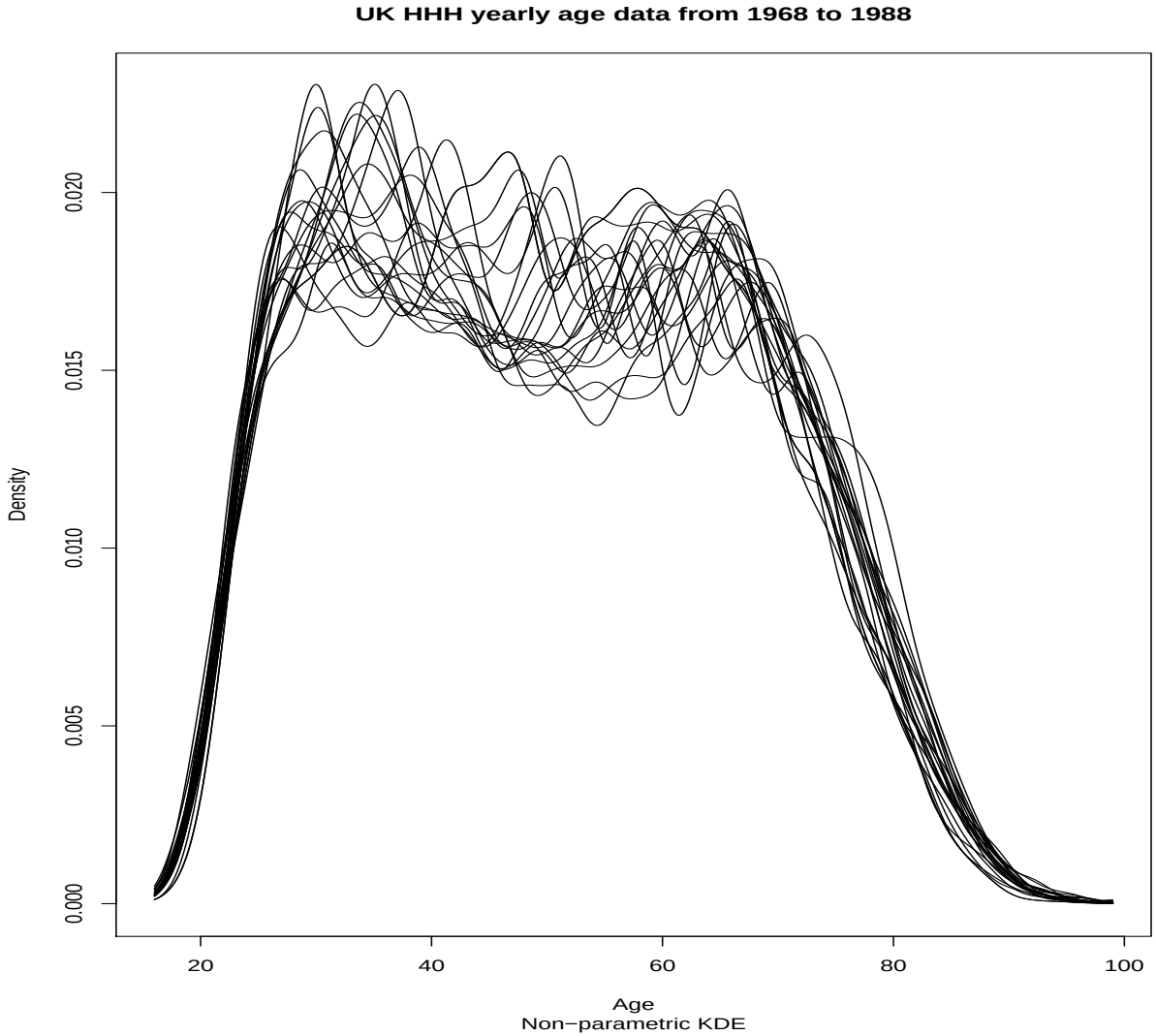


Figure: Data source: UK Family Expenditure Survey data on age of the head of the household (HHH),from the years 1968 to 1988.

Data as multiple samples

- In these applications, data are in the form of multiple samples:

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

- The population distributions $G_0, G_1, \ldots, G_m$ share some common features.
- Making use of this information will lead to better efficiency when conducting statistical inference on various aspects of the multiple distributions.

How to get these similar populations connected?

A fully parametric approach:

- choose a suitable parametric model (e.g., Normal distribution) for each of the multiple populations
- need to take the risk of model misspecification: a mild violation on model assumptions may lead to unreliable statistical conclusions
- ☹

How to get these similar populations connected?

A fully non-parametric approach:

- do not place distributional assumptions on the populations
- can avoid the risk of model misspecification
- usually leads to low statistical efficiency
- 

How to get these similar populations connected?

Middle grounds can be found! $\Longrightarrow$ a semi-parametric approach: the density ratio model [Anderson, 1979]:

- do not place distributional assumptions directly on each population
- model the connection between the multiple population distributions
- a flexible but efficient compromise between the parametric and non-parametric approaches
- ☺

Outline

Motivation

A semiparametric model: Density Ratio Model

A non-parametric inference method: Empirical Likelihood

How does the EL-DRM framework work in quantile estimation?

Our contribution

Real Data

Recall: Data as multiple samples

- Data are in the form of multiple samples (e.g., collected over years):

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

- $G_0, G_1, \ldots, G_m$ are the distributions of the multiple populations.
- Samples drawn from different populations (e.g., different years) are assumed to be independent of each other.

The Density Ratio Model (DRM)

- Let $g_k(x)$ denote the density function of the k-th population distribution G_k, for $k = 0, 1, \ldots, m$.
- The DRM models the relationship between $G_0, \ldots, G_m$ by assuming that the ratio of any two density functions is of a certain form.
- For $k = 0, 1, \ldots, m$,

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

- We call G_0 the base distribution.
- $\mathbf{q}(x)$ is some given vector-valued function, called the *basis function*.
- $(\alpha_k, \boldsymbol{\theta}_k)$ is some unknown vector-valued parameters to be estimated.

DRM is a flexible model

Many distribution families fall into the DRM category:

- Normal distributions satisfy the DRM with basis function $\mathbf{q}(x) = (x, x^2)$;
- Gamma distributions satisfy the DRM with basis function $\mathbf{q}(x) = (x, \log x)$;
- any exponential family.

On the other hand, to use DRM, we do not need to assume that the distributions $G_0, \ldots, G_m$ are Normal or Gamma.

$$\Downarrow$$

DRM is very flexible.

One subtle thing in DRM

- The base distribution G_0 is still left unspecified in DRM.
- If we assign a parametric distribution to the base distribution G_0 (for example, let G_0 be the Normal distribution), the DRM would reduce to a usual parametric model.

- Data analysis based on a parametric model must take the risk of model misspecification. ☹

A inference method: Empirical Likelihood

- Motivated by this observation, we use a non-parametric inference method: the empirical likelihood (EL) [Owen, 1988].
- Owen [2001]: EL is a nonparametric method of statistical inference. “It keeps the effectiveness of likelihood methods and does not impose a known family distribution on the data”.
- There have been many works on the EL approach under the DRM [e.g., Qin, 1993; Qin and Zhang, 1997; Fokianos et al., 2001; Qin, 1998; Chen and Liu, 2013; Cai et al., 2017].

Outline

Motivation

A semiparametric model: Density Ratio Model

A non-parametric inference method: Empirical Likelihood

How does the EL-DRM framework work in quantile estimation?

Our contribution

Real Data

EL for a single sample

- In principle, the likelihood of a distribution given a random sample is proportional to the probability of observing the sample under this distribution.
- The EL of a distribution is defined to be the likelihood as if the distribution is discrete.
- Suppose we have a sample of i.i.d. observations $x_1, \ldots, x_n$ drawn from a common distribution F.
- Let $p_i := P_F(X_i = x_i)$.
- The EL of the distribution F is defined as the probability of observing this random sample:

$$L_n(F) = \prod_{i=1}^{n} P_F(X_i = x_i) = \prod_{i=1}^{n} p_i = p_1 \times p_2 \times \cdots p_n.$$

- Following the work by Owen [1988], we require that

$$\sum_{i=1}^{n} p_i = p_1 + p_2 + \cdots + p_n = 1.$$

EL under DRM

- Let x_{kj} be the j-th observation from the k-th population, and let

$$p_{kj} = P_{G_0}(X = x_{kj}).$$

- The principle of EL leads to the EL under the DRM:

$$L_n(G_0, \ldots, G_m) = \prod_{k,j} P_{G_k}(X = x_{kj}) = \Big\{ \prod_{k,j} p_{kj} \Big\} \times \exp\Big\{ \sum_{k,j} \theta_k^{\mathsf{T}} \mathbf{q}(x_{kj}) \Big\}.$$

- The logarithm of EL, called log-EL, is a function of θ_k, p_{kj}:

$$\ell_n(\theta_k, p_{kj}) = \log L_n(G_0, \ldots, G_m) = \sum_{k,j} \log(p_{kj}) + \sum_{k,j} \theta_k^{\mathsf{T}} \mathbf{q}(x_{kj}).$$

What can we do with this log-EL function?

- Recall that in the classical likelihood theory, what can we do with the log likelihood function?
  - give point estimates on the parameters: maximum likelihood estimator
  - conduct hypothesis tests on parameters: likelihood ratio test
  - draw confidence intervals on the parameters
  - etc...
- We can do similar things with the log-EL function!
- We can regard the log-EL function as the usual log likelihood function in the classical likelihood theory.

Outline

Motivation

A semiparametric model: Density Ratio Model

A non-parametric inference method: Empirical Likelihood

How does the EL-DRM framework work in quantile estimation?

Our contribution

Real Data

Quantile estimation

- Data: $m+1$ independent random samples,

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

- Let ξ_k be the τ_k-th quantile of the k-th population, for $k = 0, 1, \ldots, m$.
  e.g., let ξ_1 be the 10% quantile of G_1.
- Target: efficiently estimate the quantiles $(\xi_0, \xi_1, \ldots, \xi_m)$ at the same time, by using the DRM to have these distributions connected.

The EL-DRM quantile estimator

- Following Owen [1988]'s convention, we require that for $r = 0, 1, \ldots, m$,

$$\sum_{k,j} p_{kj} \exp\left\{\boldsymbol{\theta}_r^\top \mathbf{q}(x_{kj})\right\} = 1. \tag{1}$$

- As in the classical likelihood theory, we obtain the maximizer $\hat{\boldsymbol{\theta}}_k, \hat{p}_{kj}$ of the log-EL function, subject to the constraints (1).
- The distributions $G_0, \ldots, G_m$ are fully characterized by $p_{kj}, \boldsymbol{\theta}_k$ in the framework of EL-DRM.
- Once we obtain the maximizer $\hat{\boldsymbol{\theta}}_k, \hat{p}_{kj}$, we can easily obtain the estimated distribution function $G_k(x)$, denoted as $\hat{G}_k(x)$.
- The EL-DRM quantile estimator for the k-th distribution is naturally defined as the value at which the estimated distribution $\hat{G}_k(x)$ first exceeds τ_k (e.g., 10%).

Outline

Motivation

A semiparametric model: Density Ratio Model

A non-parametric inference method: Empirical Likelihood

How does the EL-DRM framework work in quantile estimation?

Our contribution

Real Data

One problem in DRM remains unsolved

- In the current literature of DRM, researchers assume the knowledge of the basis function $\mathbf{q}(x)$.
- In real-world applications, complete knowledge about the most suitable basis function is impossible.
- How to specify a suitable basis function remains an unsolved problem.
- We propose an approach to specifying a basis function based on data.

Formulate a "suitable" basis function

- Recall the DRM assumption: for $k = 0, 1, \ldots, m$,

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

- Re-write the DRM assumption as

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

- The *log density ratios* $f_0(x), f_1(x), \ldots, f_m(x)$ are all linear combinations of the elements in the basis function $\mathbf{q}(x)$.
- Intuitively, if we regard $f_0, f_1, \ldots, f_m$ as functional data, the elements of basis function should represent the dominant modes of variation of such functional data.

Functional Principal Component Analysis (FPCA)

FPCA is very similar in idea to Principal Component Analysis, except that it deals with functional data.

- the "data" now are in the form of functions: $f_0, \ldots, f_m$;
- FPCA is a dimension reduction technique on functional data;
- Use only a small number of functions to well approximate $f_0, \ldots, f_m$:

$$f_k(x) \approx \sum_{i=1}^{l} \beta_i^k \varphi_i(x).$$

- $\varphi_1(x), \varphi_2(x), \ldots, \varphi_l(x)$, usually called the functional principal components (FPC), give l orthogonal dominant modes of variation among the functional data $f_0, \ldots, f_m$.
- We propose to use these FPC's as the basis function in DRM:

$$\mathbf{q}(x) = (\varphi_1(x), \varphi_2(x), \ldots, \varphi_l(x)).$$

Outline

Motivation

A semiparametric model: Density Ratio Model

A non-parametric inference method: Empirical Likelihood

How does the EL-DRM framework work in quantile estimation?

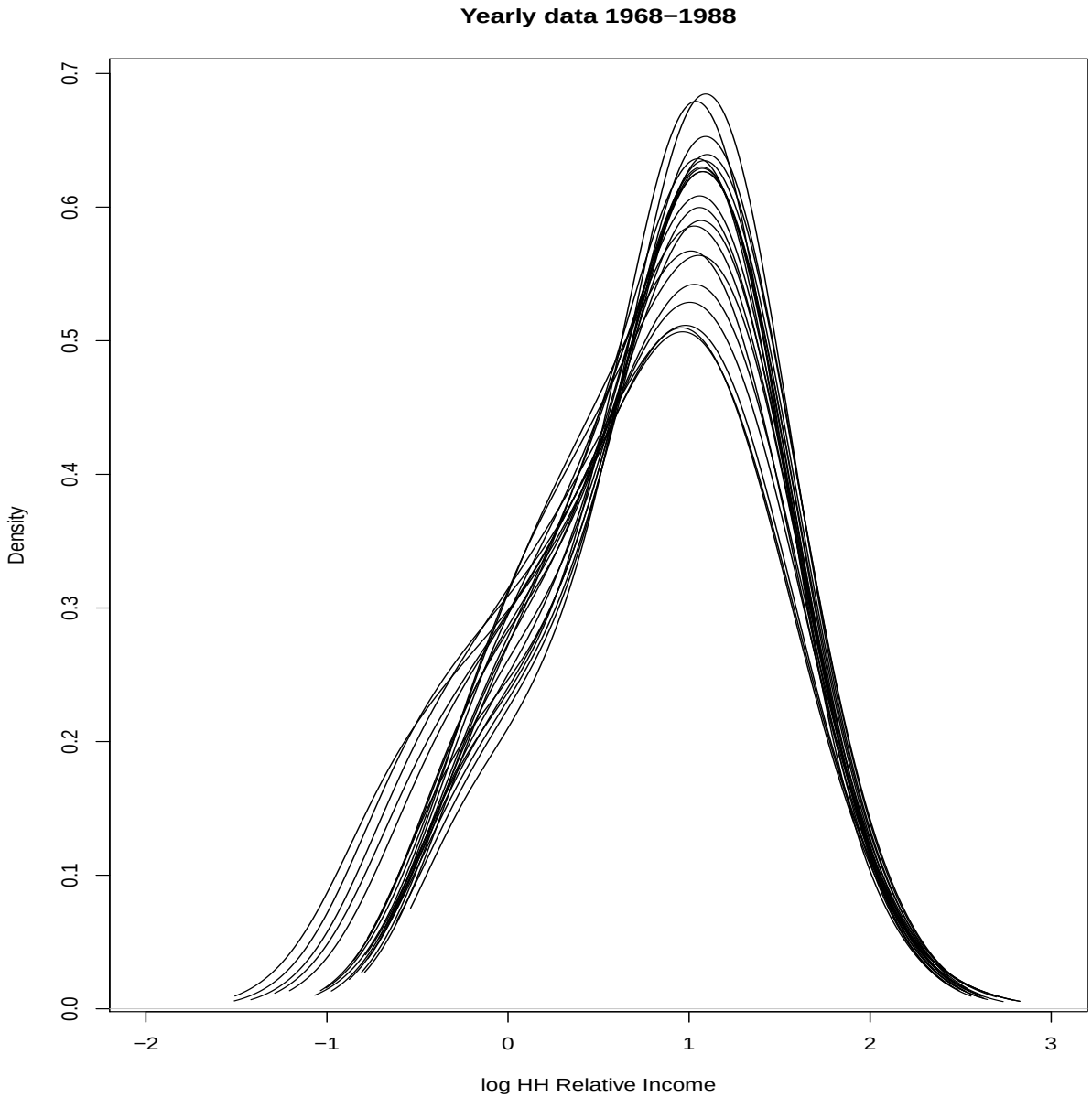
Our contribution

Real Data

UK household net income data

- Survey data from the Family Expenditure Survey in UK, from the years 1968 to 1988.
- The data consists of 21 yearly cross-sectional samples that include information about the incomes of more than 7000 households (HH's) each year.
- The variable of interest is the HH relative income: the ratio of the net HH income against the mean income of the population in the same year.
- The net HH income (aka disposable income) is the total income of all the HH members, after the income taxes have been accounted for.
- Here we focus on the log HH relative income.

Plots of the distributions for the log relative income from 1968 to 1988



Obervations

- The 21 curves represent the distributions of 21 populations (yearly, 1968-1988).
- It looks like there is some connection between these 21 distributions.
- It is reasonable to fit DRM to this data.
- However, the question is: which basis function should we use?
- We look at how the quantile estimation works under DRM when we use different basis functions.
- We focus on estimating the 21 10%-quantiles of the 21 populations.

Procedure on estimating 10% quantiles

- We regard the 21 (yearly, from 1968 to 1988) surveyed log HH relative income datasets as 21 populations.
- Each of the 21 populations is of size over 7000.
- The 10% sample quantiles of the 21 surveyed datasets are regarded as the 10% population quantiles.
- We randomly sample 100 observations (called sub-samples) from each of the 21 surveyed datasets.
- Fit the DRM to these 21 sub-samples and obtain the 10% EL-DRM quantile estimators for the 21 populations.
- Repeat the previous 2 steps for many times (for example, 500 times).
- Measure the mean squared error (MSE) of these 500 sets of 10% EL-DRM quantile estimators based on the sub-samples.

Performance of the 10% EL-DRM quantile estimators

Performance of the 10% EL-DRM quantile estimators for the first 6 years 1968-1973:

Table: MSE of the 10% EL-DRM quantile estimators for the years 1968-1973 using different basis functions (smaller is better).

DRM with basis function $\mathbf{q}(x)$	MSE of the 10% EL-DRM quantile estimators					
	Year 1968	Year 1969	Year 1970	Year 1971	Year 1972	Year 1973
(x, x^2)	0.0114	0.0113	0.0135	0.0127	0.0134	0.0133
$(\sqrt{\lvert x\rvert}, x, x^2, \log(1+\lvert x\rvert))$	0.0142	0.0148	0.0160	0.0146	0.0145	0.0150
1 FPC	0.0082	0.0063	0.0086	0.0074	0.0078	0.0096
2 FPC's	0.0089	0.0084	0.0101	0.0098	0.0096	0.0104
	Baseline					
sample quantile of sub-samples	0.0180	0.0180	0.0195	0.0177	0.0189	0.0197

Conclusions

- The DRM can efficiently take the connection between multiple populations into consideration.
- Using our proposed FPC's as the basis function in DRM leads to efficient estimation of the population quantiles.
- Besides estimation of quantiles, the EL-DRM framework also has successes in many other areas.
- We would love to see the DRM being applied and connected to sociological research studies! :)

References I

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

S. Cai, J. Chen, and J. V. Zidek. Hypothesis testing in the presence of multiple samples under density ratio models. *Statistica Sinica*, 27:716–783, 2017.

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

K. Fokianos, B. Kedem, J. Qin, and D. A. Short. A semiparametric approach to the one-way layout. *Technometrics*, 43(1):56–65, 2001.

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

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

J. Qin. Empirical likelihood in biased sample problems. *The Annals of Statistics*, pages 1182–1196, 1993.

J. Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

J. Qin and B. Zhang. A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 84(3):609–618, 1997.

Thank you!