Density Ratio Model with Data-Adaptive Basis Function

2022 DoSS Postdoctoral Day

Archer Gong Zhang
Department of Statistical Sciences
University of Toronto

Acknowledgement

This presentation is based on the joint work with my PhD supervisor at UBC.

Dr. Jiahua Chen

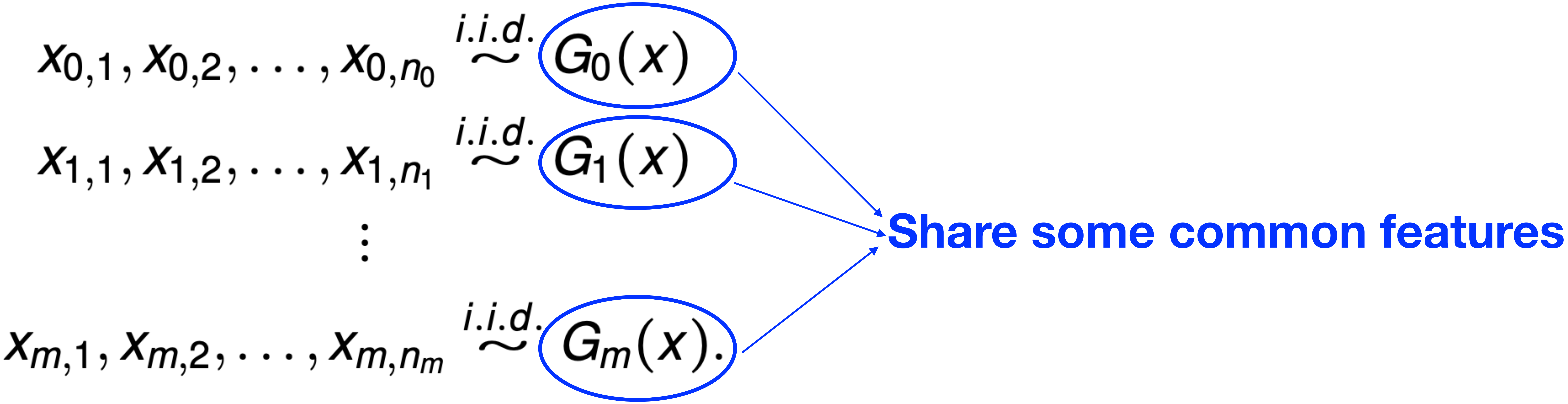
Outline

- Motivation
- A Semiparametric Model: Density Ratio Model
- Data-Adaptive Basis Function in the Density Ratio Model

Motivation

Motivation

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



$$x_{0,1}, x_{0,2}, \dots, x_{0,n_0} \overset{i.i.d.}{\sim} G_0(x)$$

$$x_{1,1}, x_{1,2}, \dots, x_{1,n_1} \overset{i.i.d.}{\sim} G_1(x)$$

$$\vdots$$

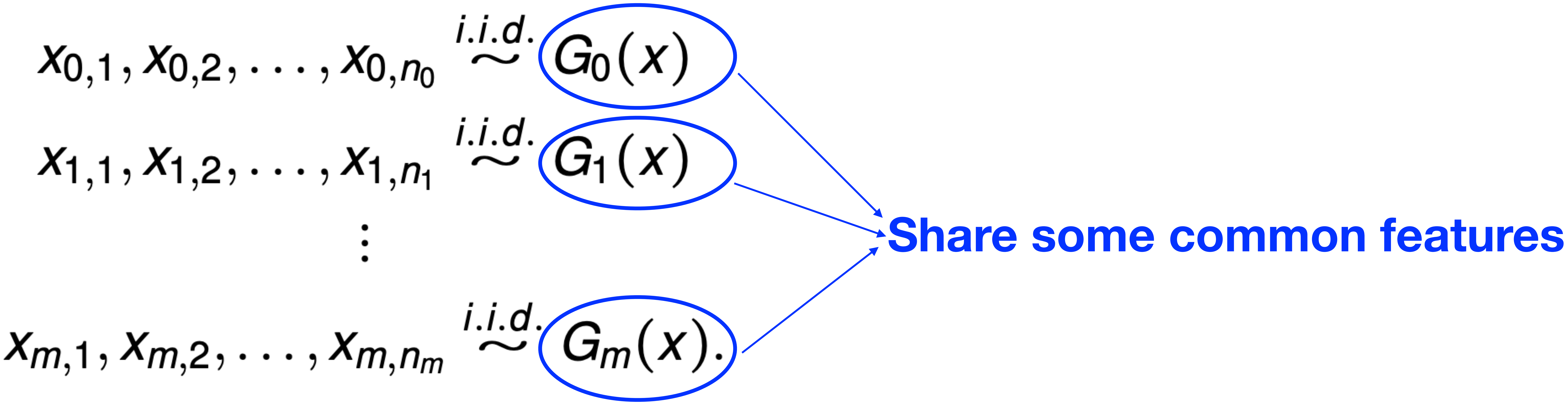
$$x_{m,1}, x_{m,2}, \dots, x_{m,n_m} \overset{i.i.d.}{\sim} G_m(x).$$

Share some common features

- E.g., to study the evolution of the economic status of a country, survey data sets of household income data are collected over multiple years: G_k is the population distribution for each year.

Motivation

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



$$x_{0,1}, x_{0,2}, \ldots, x_{0,n_0} \overset{i.i.d.}{\sim} \boxed{G_0(x)}$$

$$x_{1,1}, x_{1,2}, \ldots, x_{1,n_1} \overset{i.i.d.}{\sim} \boxed{G_1(x)}$$

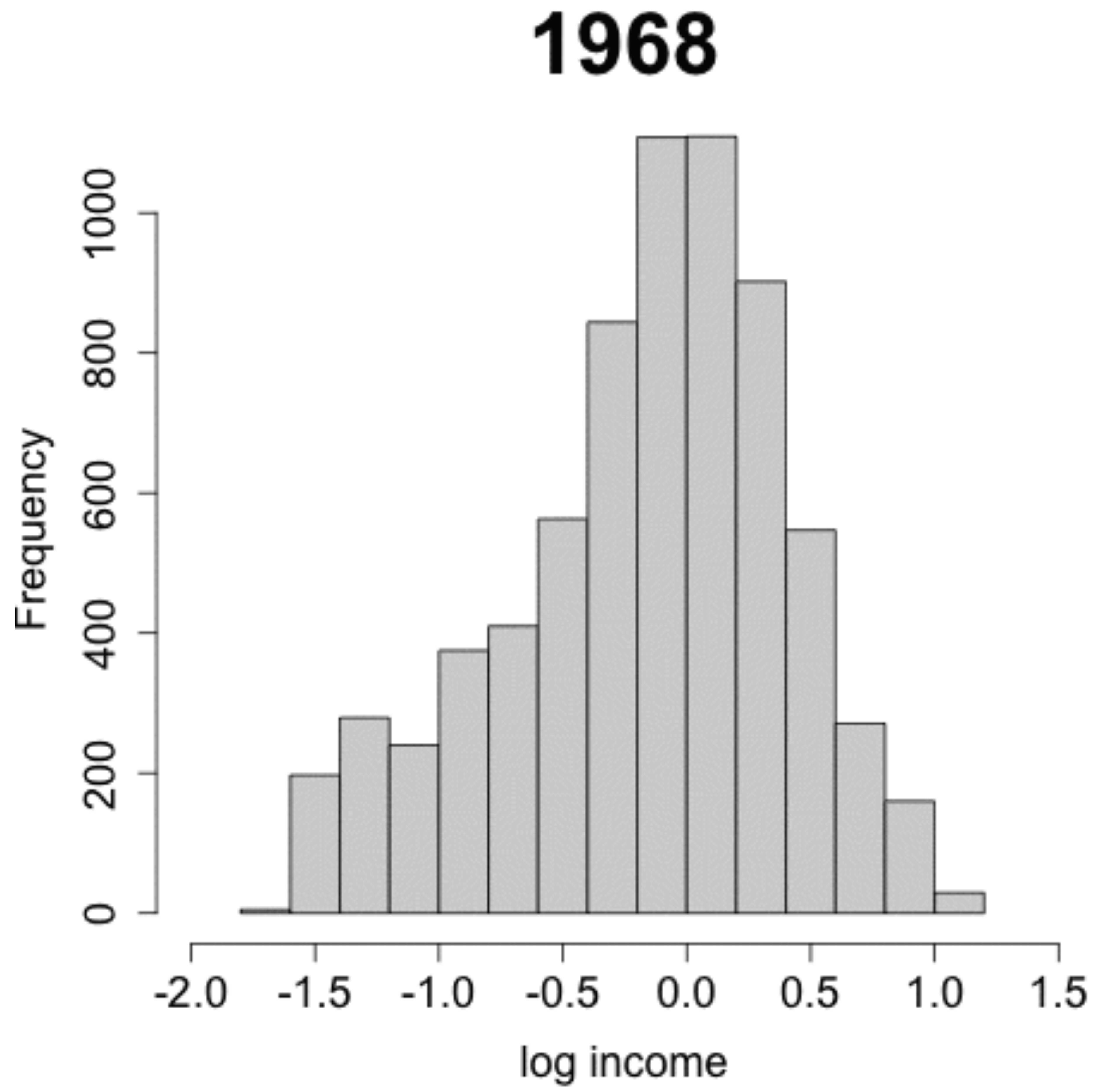
$$\vdots$$

$$x_{m,1}, x_{m,2}, \ldots, x_{m,n_m} \overset{i.i.d.}{\sim} \boxed{G_m(x).}$$

Share some common features

- E.g., in network studies, people’s activities on social networks collected in different time periods/locations naturally form multiple samples: G_k is the population distribution for each time period/location.

Example: How to analyze data like these?



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

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! 😃

A Semiparametric Model: Density Ratio Model

Density ratio model (DRM) (Anderson, 1979)

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

$$\frac{g_k(x)}{g_0(x)} = \exp\{\alpha_k + \theta_k^\top \boldsymbol{q}(x)\}$$

**unknown parameters
to be estimated**

**vector-valued function:
basis function**

- We call G_0 the base distribution; any G_k may serve the same purpose.
- Sample from G_k forms a biased sample from G_0 characterized by the exponential tilting!

Why DRM?

DRM: $g_k(x)/g_0(x) = \exp\{\alpha_k + \boldsymbol{\theta}_k^\top \boldsymbol{q}(x)\}$.

- DRM is flexible: G_0 is unspecified and users can choose a $\boldsymbol{q}(x)$ they wish, which allow it to cover many distribution families.

Distribution family	**Basis function $\boldsymbol{q}(x)$**
Normal	(x, x^2)
Gamma	$(x, \log x)$
Exponential family	Sufficient statistics
…	…

Why DRM?

DRM: $g_k(x)/g_0(x) = \exp\{\alpha_k + \boldsymbol{\theta}_k^\top \boldsymbol{q}(x)\}$.

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

↓

Gain in statistical efficiency!

- Can be useful for integrating data from heterogeneous sources: underlying data distributions may not be identical but probably connected.

Inference for the unspecified G_0

- If assigning a parametric form to G_0, DRM would reduce to a fully parametric model.
- Use a nonparametric inference method: empirical likelihood (EL; Owen, 1988).

Art B. Owen

Owen (2001): "EL keeps the effectiveness of likelihood methods and does not impose a known family distribution on the data."

(Pause for questions.)

Data-Adaptive Basis Function in the DRM

An open problem in DRM

- The benefit of DRM largely relies on the correct specification of the basis function $\boldsymbol{q}(x)$.
- Complete knowledge of $\boldsymbol{q}(x)$ is impossible in applications.
- Some remedies in the current literature:
    - choose a $\boldsymbol{q}(x)$ based on some exploratory data analysis;
    - Chen and Liu (2013): use a rich $\boldsymbol{q}(x)$ for "safety", e.g., $\boldsymbol{q}(x) = (|x|^{1/2}, x, x^2, \log(1+|x|))^\top$;
    - Fokianos (2007): select a $\boldsymbol{q}(x)$ among a number of candidates based on some model selection criterion.
- How to choose $\boldsymbol{q}(x)$ based on data remains an open problem.
- We propose a data-adaptive approach to the choice of $\boldsymbol{q}(x)$.

Our contribution helps further alleviate the risk of model misspecification!

A closer look at $\boldsymbol{q}(x)$

- Recall the DRM assumption: $g_k(x)/g_0(x) = \exp\{\alpha_k + \boldsymbol{\theta}_k^\top \boldsymbol{q}(x)\}$.

- The DRM is always satisfied if

$$\boldsymbol{q}(x) = (\log\{g_1(x)/g_0(x)\}, \ldots, \log\{g_m(x)/g_0(x)\}) .$$

- Therefore, the DRM is meaningful when all centred $\log\{g_k(x)/g_0(x)\}$ can be written as $\boldsymbol{\theta}_k^\top \boldsymbol{q}(x)$ for some lower-than-m dimensional $\boldsymbol{q}(x)$!

- Assume such a low-dim $\boldsymbol{q}(x)$ exists and

$$\mathbb{E}_{\bar{G}}[\boldsymbol{q}(X)] = \mathbf{0},$$

under $\bar{G} = \sum_{k=0}^{m} \rho_k G_k(x)$, where $\rho_k = \lim n_k / N_{\text{total}}$.

Formulate an appropriate $q(x)$

- Under these assumptions on $\boldsymbol{q}(x)$, define

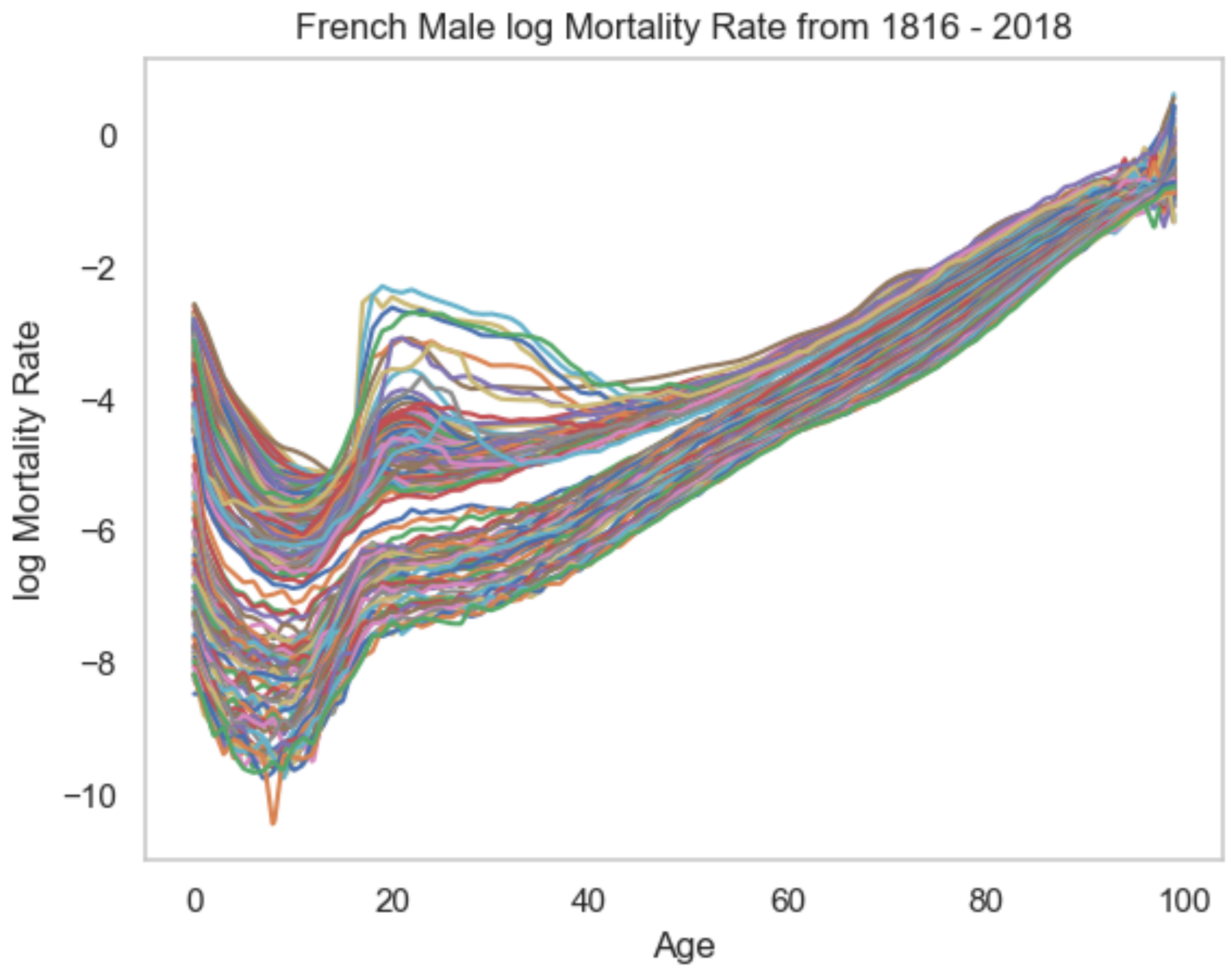
$$Q_k(x) := \log \frac{g_k(x)}{g_0(x)} - \boxed{\mathbb{E}_{\bar{G}}\left[\log \frac{g_k(X)}{g_0(X)}\right]} = \boldsymbol{\theta}_k^\top \boldsymbol{q}(x)\,.$$

α_k

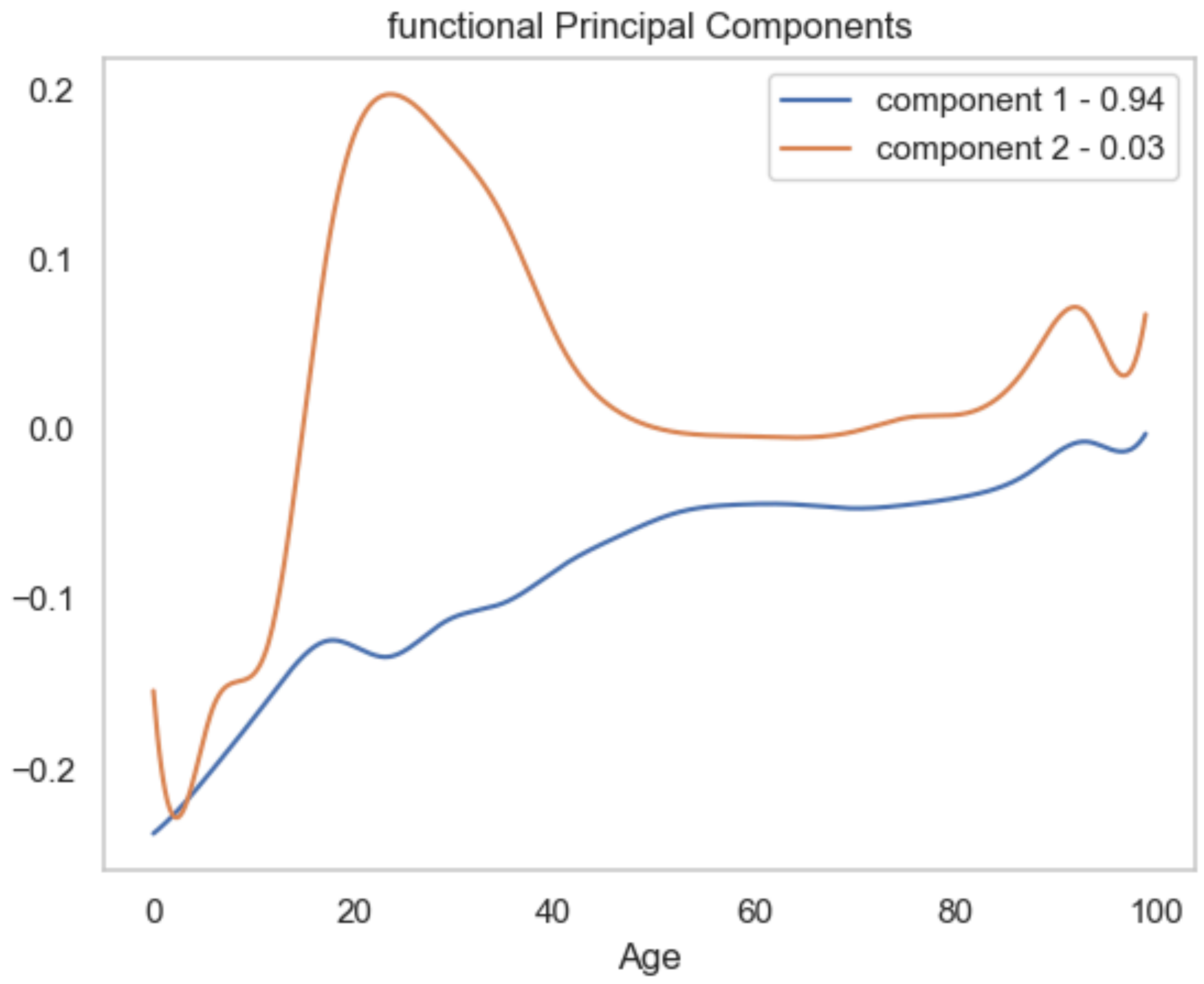
- $\{Q_0(x), \ldots, Q_m(x)\}$ forms a linear space.
- $\boldsymbol{q}(x)$ is made of elements in the basis of such a linear space.
- Idea: form $\boldsymbol{q}(x)$ by the dominant modes of variation of $\{Q_0(x), \ldots, Q_m(x)\}$.

Functional principal component analysis (FPCA)

FPCA is a dimension reduction technique on functional data (in our case: $\{Q_0(x), \ldots, Q_m(x)\}$) that aims to find their dominant modes of variation.


French Male log Mortality Rate from 1816 - 2018

Functional data: curves


functional Principal Components

Dominant modes: functional directions

Figure source: https://towardsdatascience.com/functional-principal-component-analysis-and-functional-data-91d21261ab7f

FPCA (cont'd)

- Via FPCA, $Q_0(x), \ldots, Q_m(x)$ can be represented by $d < m$ functional principal components (FPCs):

$$Q_k(x) - \underbrace{\boxed{\frac{1}{m+1}\sum_{r=0}^{m} Q_r(x)}}_{\text{centralization}} = \sum_{j=1}^{d} \beta_j^k \underbrace{\psi_j(x)}_{j\text{th FPC}}$$

- FPCs $\psi_1(x), \ldots, \psi_d(x)$ are the dominant modes of variation of $\{Q_0(x), \ldots, Q_m(x)\}$.
- They are "optimal": explain the most variability among $\{Q_0(x), \ldots, Q_m(x)\}$.

Recovery of the FPCs

Given complete knowledge of $Q_0(x), \ldots, Q_m(x)$, we can obtain the FPCs via linear algebra.

- Let $\boldsymbol{M}$ be an $(m+1) \times (m+1)$ matrix with the $(i \,.\, j)$th element

$$\boldsymbol{M}(i,j) = \mathbb{E}_{\bar{G}}[Q_i(X)Q_j(X)], \quad i,j = 0,\ldots,m\,.$$

- Let $\{\boldsymbol{v}_j\}_{j=1}^{d}$ be the set of eigenvectors of $\boldsymbol{M}$ corresponding to the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$.

Theorem

For $j = 1,\ldots,d$, the jth FPC $\psi_j(x)$ is given by

$$\psi_j(x) = \lambda_j^{-1/2}\boldsymbol{v}_j^\top \begin{pmatrix} Q_0(x) \\ \vdots \\ Q_m(x) \end{pmatrix}.$$

Estimation of the FPCs

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

- Via kernel density estimation: **(Silverman, 1986)**

$$\hat{g}_k(x) = \frac{1}{n_k h_k} \sum_{j=1}^{n_k} K(\frac{x - x_{k,j}}{h_k})$$

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

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

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

- Via the recovery of FPCs

We successfully prove that these estimated FPCs are consistent under some conditions.

Data-adaptive basis function $\boldsymbol{q}(x)$

- Use the top d of the estimated FPCs to form the data-adaptive $\boldsymbol{q}(x)$:

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

- In Zhang and Chen (2022), we proposed some ways to choose d adaptively:
  - proportion of explained variation in FPCA
  - model selection, e.g., BIC
- Given the adaptive $\hat{\boldsymbol{q}}(x)$, we re-use the data for model fitting and inference.

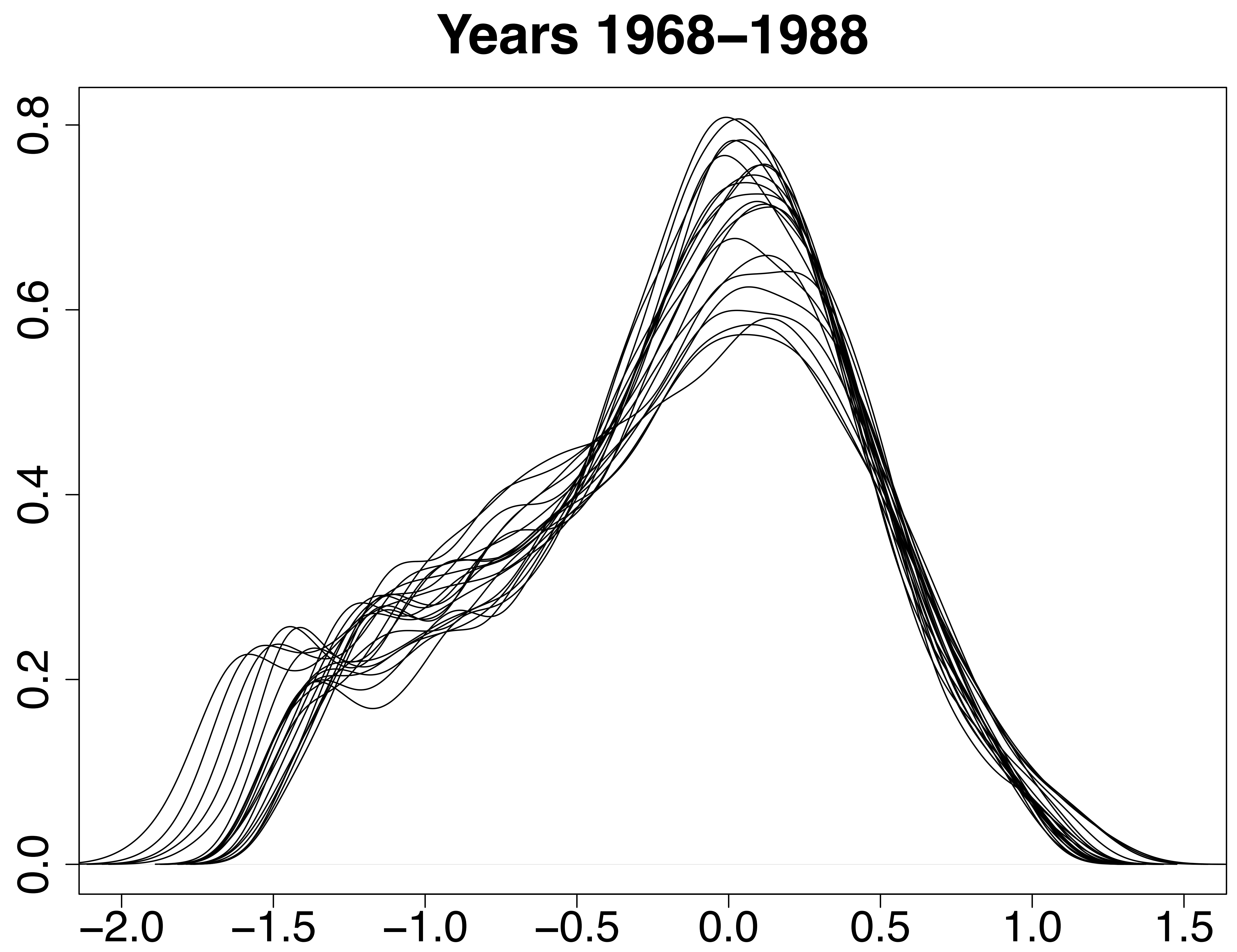
(Pause for questions.)

Real-data analysis

UK household income data

- We consider a survey data from the Family Expenditure Survey in UK, from 1968 to 1988.
- The data contain yearly samples on the incomes and expenditures of $> 7{,}000$ households (HHs) each year.
- Variable of interest: log-transformed HH relative income.

Exploratory analysis



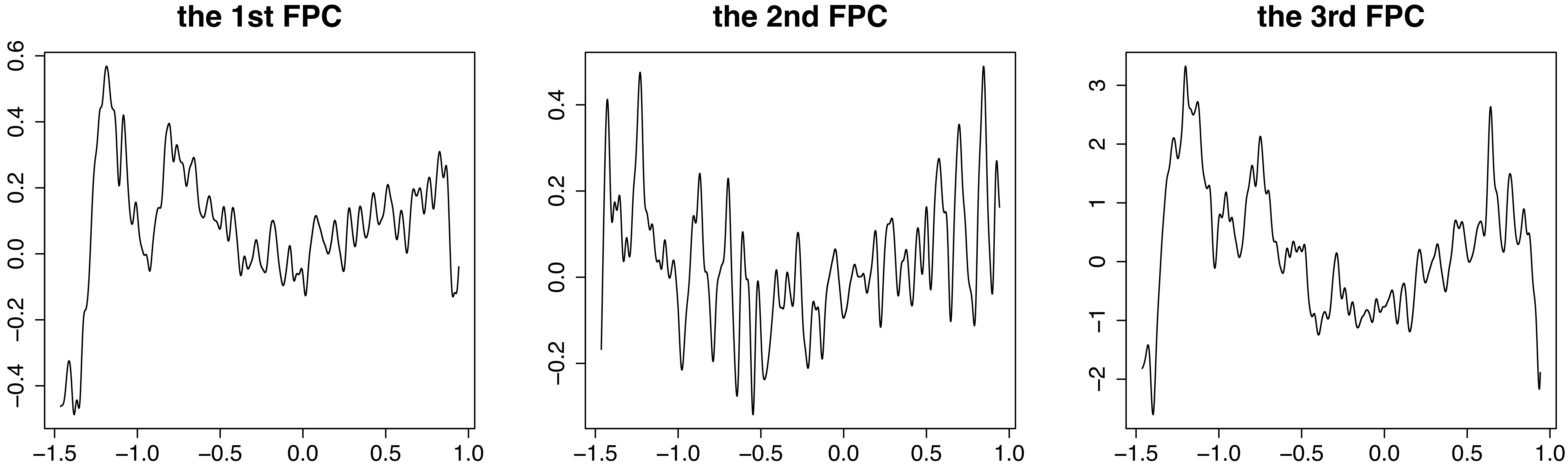
Kernel density estimators based on HH relative income data.
Apparently, there is some connection between these distributions.

Real-data based simulation procedure

We study the EL-based quantile estimation under the DRM.

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

Estimated FPCs based on real-data



First three FPCs obtained using the training data from 1968–1981. Note that there are some overall trends in these FPCs, suggesting the existence of some latent structures in the multiple populations.

Performance of quantile estimators (lower is better)

Method	Average MSE (×1000) of quantile estimators					
	10%	30%	50%	70%	90%	avg.
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
FPC-1	1.86	0.62	0.37	0.16	0.31	0.66
NP	1.78	1.41	0.84	0.57	0.67	1.05
Rich	1.88	1.39	0.91	0.56	0.54	1.06

1. The proposed "Adaptive" estimators perform well, with a $\approx 39\,\%$ gain in efficiency compared to the "NP" estimators.

Data pooling via DRM is helpful!

Performance of quantile estimators (lower is better)

Method	Average MSE (×1000) of quantile estimators					
	10%	30%	50%	70%	90%	avg.
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
FPC-1	1.86	0.62	0.37	0.16	0.31	0.66
NP	1.78	1.41	0.84	0.57	0.67	1.05
Rich	1.88	1.39	0.91	0.56	0.54	1.06

2. Our suggested adaptive approach usually selects $d = 2$, the best-performing d ("FPC-2").

Latent structure often exists in real data.

Performance of quantile estimators (lower is better)

Method	Average MSE (×1000) of quantile estimators					
	10%	30%	50%	70%	90%	avg.
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
FPC-1	1.86	0.62	0.37	0.16	0.31	0.66
NP	1.78	1.41	0.84	0.57	0.67	1.05
Rich	1.88	1.39	0.91	0.56	0.54	1.06

3. The safe choice “Rich” is not satisfactory here: barely ≈ “NP”.

Adaptive basis function is helpful under DRM!

Summary

- DRM with the proposed data-adaptive $\boldsymbol{q}(x)$ leads to efficiency gain in quantile estimation.

- Our contribution gives users confidence in the validity and the effectiveness of data analysis via DRM.

- Other DRM-based inferences using the adaptive $\boldsymbol{q}(x)$ can be similarly developed.

Some thoughts…

- Under the DRM, every G_k can be seen as a distributional shift version of G_0.
  - $g_k(x) = g_0(x)\ \exp\{\alpha_k + \boldsymbol{\theta}_k^\top \boldsymbol{q}(x)\}$.
- DRM offers an **<u>interpretable and efficient</u>** platform for the distributional shift with data from multiple connected sources/domains/environments/modalities, and could be particularly useful in:
  - out-of-distribution (OOD) generalization
  - transfer learning/domain adaptation
  - etc…

References

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.

K. Fokianos. Density ratio model selection. *Journal of Statistical Computation and Simulation*, 77(9):805–819, 2007.

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.

B. W. Silverman. Density Estimation for Statistics and Data Analysis, volume 26. CRC Press, 1986.

A. G. Zhang and J. Chen. Density ratio model with data-adaptive basis function. *Journal of Multivariate Analysis*, page 105043, 2022.

Thank you!

Questions & discussions are welcome! :-)