Estimation Efficiency under A Semiparametric Density Ratio Model

Archer Gong Zhang
Department of Statistical Sciences
University of Toronto

Jiahua Chen
Department of Statistics
University of British Columbia

Outline

- Motivation
- A Semiparametric Model: Density Ratio Model
- Estimation Efficiency under the Density Ratio Model
  - Quantile Estimation

Motivation

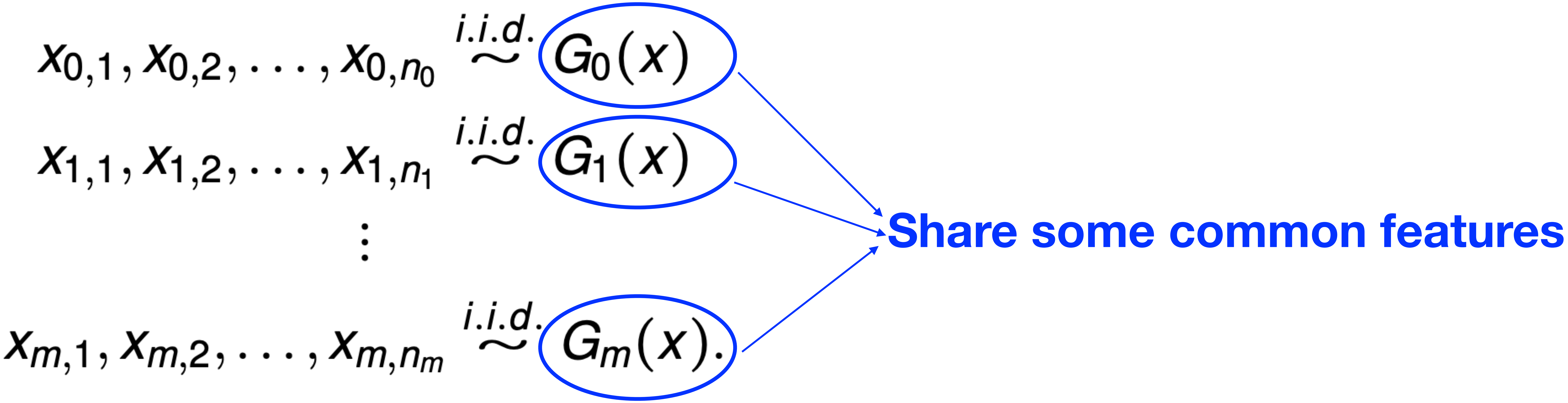
Motivation

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

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

Motivation

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



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

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

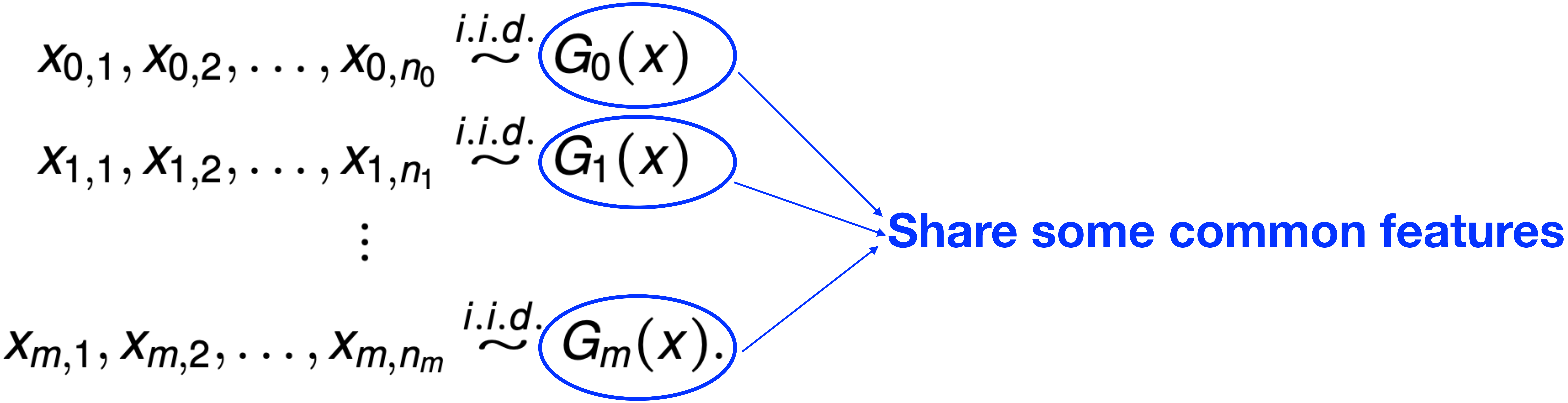
$$\vdots$$

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

Share some common features

Motivation

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



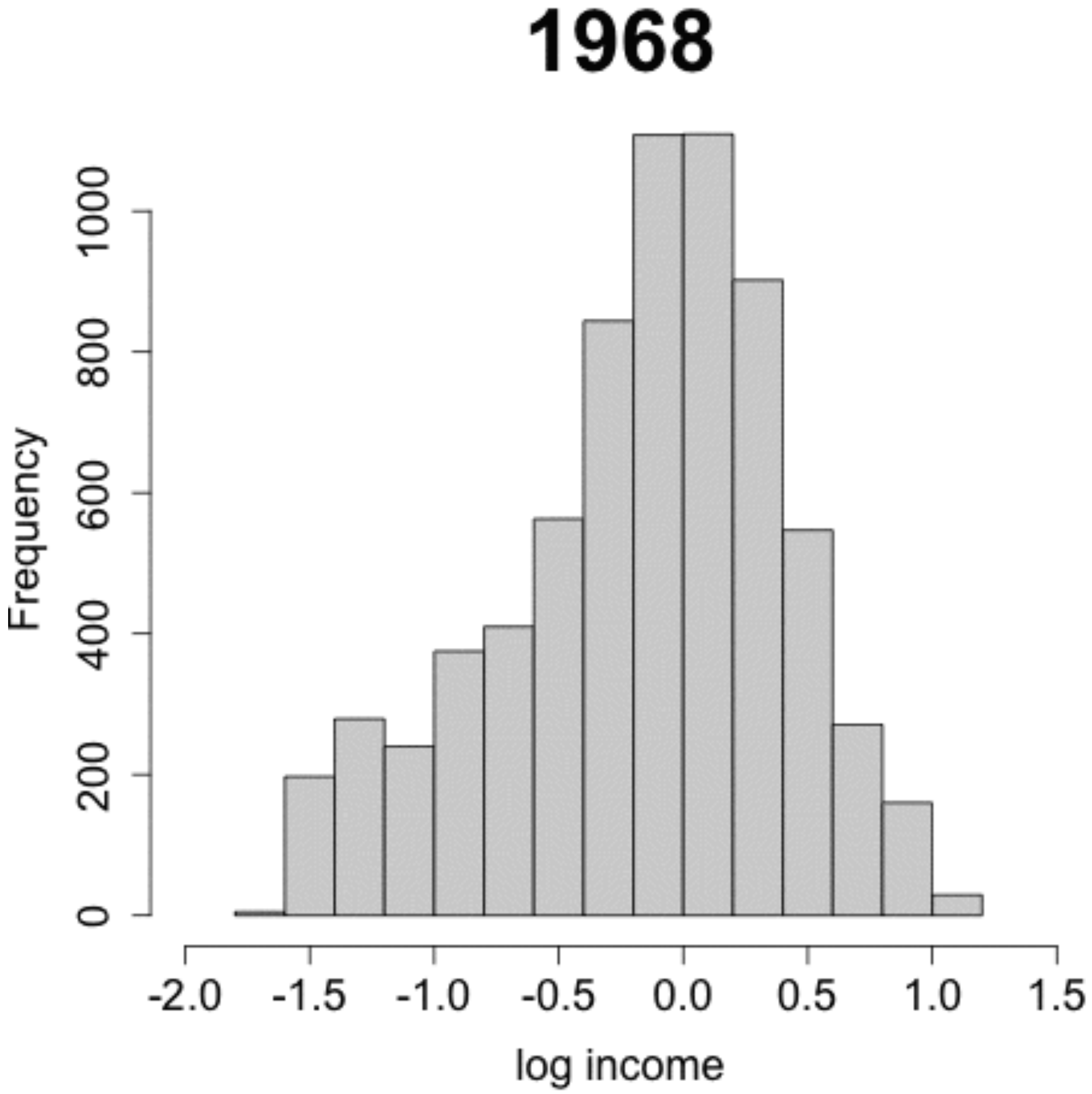
$$
\begin{aligned}
x_{0,1}, x_{0,2}, \ldots, x_{0,n_0} &\overset{i.i.d.}{\sim} G_0(x) \\
x_{1,1}, x_{1,2}, \ldots, x_{1,n_1} &\overset{i.i.d.}{\sim} G_1(x) \\
&\vdots \\
x_{m,1}, x_{m,2}, \ldots, x_{m,n_m} &\overset{i.i.d.}{\sim} G_m(x).
\end{aligned}
$$

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.

How to analyze data like these?

Income data from UK Family Expenditure Survey



Histograms of log household relative income from 1968 to 1988.

Data source: https://archiveshub.jisc.ac.uk/search/archives/412e6ebd-8de7-3e6e-b060-34d35cffaf15.

Different approaches to statistical analysis

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! 😃

A Semiparametric Model: Density Ratio Model

Density ratio model (DRM) (Anderson, 1979)

Density ratio model (DRM) (Anderson, 1979)

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

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

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

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

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.

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

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

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!

Why DRM?

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

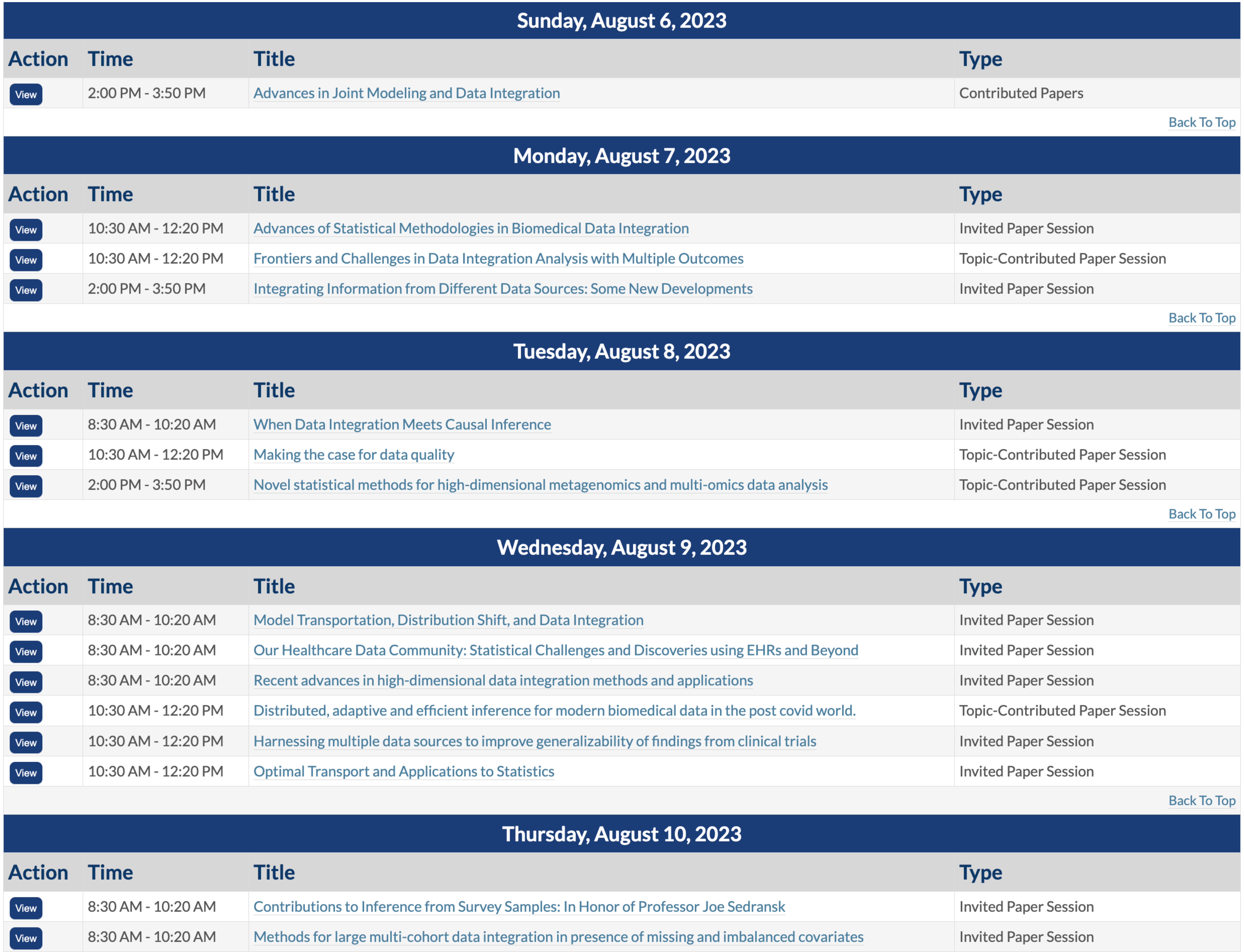
- With an appropriate basis function $\boldsymbol{q}(x)$, DRM allows us to use the pooled data to estimate G_k, rather than use data only from G_k.

↓

Gain in statistical efficiency!

- Every G_k can be seen as a distributional shift version of G_0!
  - Can be useful for integrating data from **connected** sources/domains.

Search “data integration” in JSM2023 program…popular recently!



Sunday, August 6, 2023

Action	Time	Title	Type
View	2:00 PM - 3:50 PM	Advances in Joint Modeling and Data Integration	Contributed Papers

Back To Top

Monday, August 7, 2023

Action	Time	Title	Type
View	10:30 AM - 12:20 PM	Advances of Statistical Methodologies in Biomedical Data Integration	Invited Paper Session
View	10:30 AM - 12:20 PM	Frontiers and Challenges in Data Integration Analysis with Multiple Outcomes	Topic-Contributed Paper Session
View	2:00 PM - 3:50 PM	Integrating Information from Different Data Sources: Some New Developments	Invited Paper Session

Back To Top

Tuesday, August 8, 2023

Action	Time	Title	Type
View	8:30 AM - 10:20 AM	When Data Integration Meets Causal Inference	Invited Paper Session
View	10:30 AM - 12:20 PM	Making the case for data quality	Topic-Contributed Paper Session
View	2:00 PM - 3:50 PM	Novel statistical methods for high-dimensional metagenomics and multi-omics data analysis	Topic-Contributed Paper Session

Back To Top

Wednesday, August 9, 2023

Action	Time	Title	Type
View	8:30 AM - 10:20 AM	Model Transportation, Distribution Shift, and Data Integration	Invited Paper Session
View	8:30 AM - 10:20 AM	Our Healthcare Data Community: Statistical Challenges and Discoveries using EHRs and Beyond	Invited Paper Session
View	8:30 AM - 10:20 AM	Recent advances in high-dimensional data integration methods and applications	Invited Paper Session
View	10:30 AM - 12:20 PM	Distributed, adaptive and efficient inference for modern biomedical data in the post covid world.	Topic-Contributed Paper Session
View	10:30 AM - 12:20 PM	Harnessing multiple data sources to improve generalizability of findings from clinical trials	Invited Paper Session
View	10:30 AM - 12:20 PM	Optimal Transport and Applications to Statistics	Invited Paper Session

Back To Top

Thursday, August 10, 2023

Action	Time	Title	Type
View	8:30 AM - 10:20 AM	Contributions to Inference from Survey Samples: In Honor of Professor Joe Sedransk	Invited Paper Session
View	8:30 AM - 10:20 AM	Methods for large multi-cohort data integration in presence of missing and imbalanced covariates	Invited Paper Session

Inference for the unspecified G_0

Inference for the unspecified G_0

- If assigning a parametric form to G_0, DRM would reduce to a fully parametric model.

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

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."

Estimation Efficiency under the Density Ratio Model

Efficiency of the DRM-based estimators

Efficiency of the DRM-based estimators

- Many studies have showed that some DRM-based estimators are more efficient than the nonparametric estimators.

Efficiency of the DRM-based estimators

- Many studies have showed that some DRM-based estimators are more efficient than the nonparametric estimators.

- An interested question: how far we can push their efficiency?

Efficiency of the DRM-based estimators

- Many studies have showed that some DRM-based estimators are more efficient than the nonparametric estimators.

- An interested question: how far we can push their efficiency?

- “Gold standard”: the parametric estimator derived under a parametric model (e.g., a normal model).

Efficiency of the DRM-based estimators

- Many studies have showed that some DRM-based estimators are more efficient than the nonparametric estimators.

- An interested question: how far we can push their efficiency?

- “Gold standard”: the parametric estimator derived under a parametric model (e.g., a normal model).

    - usually the most efficient (e.g., MLE).

Efficiency of the DRM-based estimators

- Many studies have showed that some DRM-based estimators are more efficient than the nonparametric estimators.

- An interested question: how far we can push their efficiency?

- "Gold standard": the parametric estimator derived under a parametric model (e.g., a normal model).

    - usually the most efficient (e.g., MLE).

- Is it likely that the DRM-based estimators can be as efficient as the parametric estimators?

Efficiency of the DRM-based estimators

- Many studies have showed that some DRM-based estimators are more efficient than the nonparametric estimators.

- An interested question: how far we can push their efficiency?

- “Gold standard”: the parametric estimator derived under a parametric model (e.g., a normal model).

    - usually the most efficient (e.g., MLE).

- Is it likely that the DRM-based estimators can be as efficient as the parametric estimators?

- Or When?

A two-sample scenario

A two-sample scenario

Consider two samples of sizes n_0, n_1 from G_0, G_1, with $n_0 \gg n_1$.

DRM that connects G_0, G_1:

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

A two-sample scenario

Consider two samples of sizes n_0, n_1 from G_0, G_1, with $n_0 \gg n_1$.

DRM that connects G_0, G_1:

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

can roughly be seen as "known": the larger sample is expected to estimate G_0 with high accuracy

A two-sample scenario

Consider two samples of sizes n_0, n_1 from G_0, G_1, with $n_0 \gg n_1$.

DRM that connects G_0, G_1:

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

DRM is then a fully parametric model for G_1

can roughly be seen as “known”: the larger sample is expected to estimate G_0 with high accuracy

A two-sample scenario

Consider two samples of sizes n_0, n_1 from G_0, G_1, with $n_0 \gg n_1$.

DRM that connects G_0, G_1:

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

DRM is then a fully parametric model for G_1

can roughly be seen as "known": the larger sample is expected to estimate G_0 with high accuracy

We therefore expect the DRM estimators for G_1 to achieve the "gold-standard"parametric efficiency!

A parametric submodel

A parametric submodel

- We consider an exponential family model for G_1:

$$x_{1,1}, \ldots, x_{1,n_1} \overset{\text{i.i.d.}}{\sim} g_1(x) = g_0(x)\ \exp\{\alpha + \boldsymbol{\theta}^\top \boldsymbol{q}(x)\}\,.$$

A parametric submodel

Known!

- We consider an exponential family model for G_1:

$$x_{1,1}, \ldots, x_{1,n_1} \overset{\text{i.i.d.}}{\sim} g_1(x) = \boxed{g_0(x)} \exp\{\alpha + \boldsymbol{\theta}^\top \boldsymbol{q}(x)\}\,.$$

A parametric submodel

Known!

- We consider an exponential family model for G_1:

$$x_{1,1}, \ldots, x_{1,n_1} \overset{\text{i.i.d.}}{\sim} g_1(x) = \boxed{g_0(x)} \exp\{\alpha + \boldsymbol{\theta}^\top \boldsymbol{q}(x)\}\,.$$

- It is a parametric submodel for the two-sample DRM with the same $\boldsymbol{q}(x)$:

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

A parametric submodel

- We consider an exponential family model for G_1:

Known!

$$x_{1,1}, \ldots, x_{1,n_1} \overset{\text{i.i.d.}}{\sim} g_1(x) = \boxed{g_0(x)} \exp\{\alpha + \boldsymbol{\theta}^\top \boldsymbol{q}(x)\}\,.$$

- It is a parametric submodel for the two-sample DRM with the same $\boldsymbol{q}(x)$:

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

- MLEs under this exponential family model are the “gold-standard” parametric estimators.

Our contribution

We theoretically prove that under the two-sample scenario, the following DRM-based estimators for G_1 achieve parametric efficiency asymptotically when $n_0/n_1 \to \infty$ as $n_0, n_1 \to \infty$:

- DRM model parameters $(\alpha, \boldsymbol{\theta})$;
- Distribution function $G_1(x)$;
- Quantiles of G_1.

Our contribution

We theoretically prove that under the two-sample scenario, the following DRM-based estimators for G_1 achieve parametric efficiency asymptotically when $n_0/n_1 \to \infty$ as $n_0, n_1 \to \infty$:

- DRM model parameters $(\alpha, \boldsymbol{\theta})$;
- Distribution function $G_1(x)$;
- Quantiles of G_1.

Our contribution is applicable and particularly useful in applications where one wishes to make efficient inference with a small sample, aided by another large historical sample.

Quantile Estimation

Efficiency of DRM quantile estimators

A special case when $G_1 = G_0$

Efficiency of DRM quantile estimators

A special case when $G_1 = G_0$

- Focus on ξ_p: the pth quantile for G_1.

Efficiency of DRM quantile estimators

A special case when $G_1 = G_0$

- 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:

Efficiency of DRM quantile estimators

A special case when $G_1 = G_0$

- 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:

$$\boxed{\mathrm{Var}(\hat{\xi}_p)} = \frac{1}{k+1}\boxed{\frac{p(1-p)}{n_1 g_1^2(\xi_p)}} + \frac{k}{k+1}\boxed{\mathrm{Var}(\tilde{\xi}_p)}.$$

Variance of DRM quantile

Variance of Nonpara. quantile

Variance of parametric quantile

Simulation with data from normal distributions

Parameter of interest: ξ_p — the pth quantile for G_1

- Generate two samples both from $N(0,1)$.
- Obtain the DRM quantile estimator only assuming the knowledge of the most appropriate $\boldsymbol{q}(x) = (x, x^2)^\top$.
- Two competitors that only use sample from G_1:
    - MLE of quantile derived under the normal model
    - Nonparametric empirical quantile

Performance of quantile estimators

Biases are $\times\sqrt{n_1}$; Variances are $\times n_1$; Based on 1000 repetitions

Levels p	DRM-based		MLE		Nonparametric	
	Bias	Var	Bias	Var	Bias	Var
	$n_0 = k \times n_1, \quad n_1 = 1000, \quad k = 10$					
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
	$n_0 = k \times n_1, \quad n_1 = 1000, \quad k = 100$					
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

Performance of quantile estimators

Biases are $\times\sqrt{n_1}$; Variances are $\times n_1$; Based on 1000 repetitions

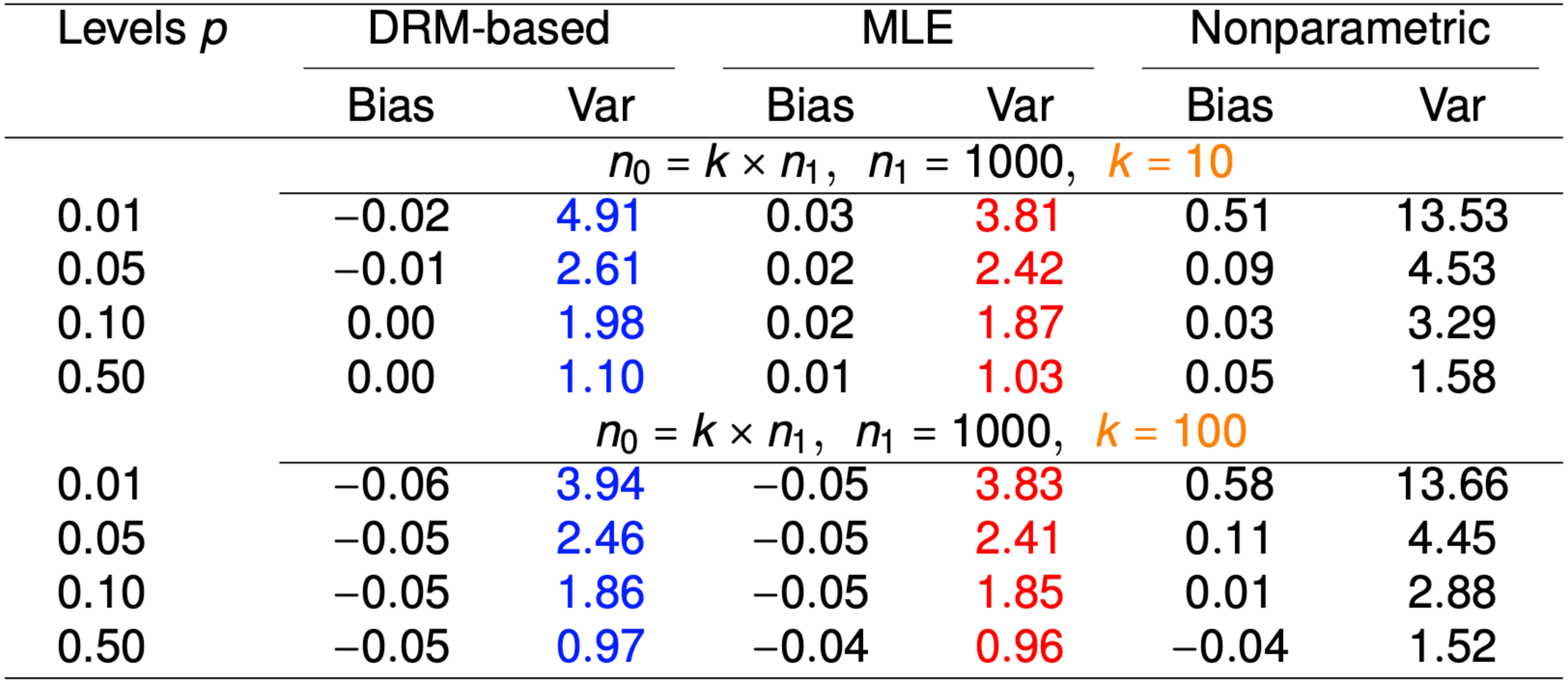
Levels p	DRM-based		MLE		Nonparametric	
	Bias	Var	Bias	Var	Bias	Var
			$n_0 = k \times n_1, \quad n_1 = 1000, \quad \boxed{k = 10}$			
0.01	−0.02	4.91	0.03	3.81	0.51	13.53
0.05	−0.01	2.61	0.02	2.42	0.09	4.53
0.10	0.00	1.98	0.02	1.87	0.03	3.29
0.50	0.00	1.10	0.01	1.03	0.05	1.58
			$n_0 = k \times n_1, \quad n_1 = 1000, \quad \boxed{k = 100}$			
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

1. As $\boxed{k}$ ↑, variances of the DRM estimators approach those of the MLEs.

Matches our theoretical result!

Performance of quantile estimators

Biases are $\times\sqrt{n_1}$; Variances 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

2. Our “weighted average” result is also well 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

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! :-)

Q & A