

## A selective review of statistical methods using calibration information from similar studies

Jing Qin, Yukun Liu & Pengfei Li

To cite this article: Jing Qin, Yukun Liu & Pengfei Li (2022): A selective review of statistical methods using calibration information from similar studies, Statistical Theory and Related Fields, DOI: [10.1080/24754269.2022.2037201](https://doi.org/10.1080/24754269.2022.2037201)

To link to this article: <https://doi.org/10.1080/24754269.2022.2037201>



© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 17 Feb 2022.



Submit your article to this journal [↗](#)



Article views: 227



View related articles [↗](#)



View Crossmark data [↗](#)

# A selective review of statistical methods using calibration information from similar studies

Jing Qin<sup>a</sup>, Yukun Liu<sup>b</sup> and Pengfei Li<sup>c</sup>

<sup>a</sup>National Institute of Allergy and Infectious Diseases, National Institutes of Health, Frederick, MD, USA; <sup>b</sup>KLATASDS – MOE, School of Statistics, East China Normal University, Shanghai, People's Republic of China; <sup>c</sup>Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada

## ABSTRACT

In the era of big data, divide-and-conquer, parallel, and distributed inference methods have become increasingly popular. How to effectively use the calibration information from each machine in parallel computation has become a challenging task for statisticians and computer scientists. Many newly developed methods have roots in traditional statistical approaches that make use of calibration information. In this paper, we first review some classical statistical methods for using calibration information, including simple meta-analysis methods, parametric likelihood, empirical likelihood, and the generalized method of moments. We further investigate how these methods incorporate summarized or auxiliary information from previous studies, related studies, or populations. We find that the methods based on summarized data usually have little or nearly no efficiency loss compared with the corresponding methods based on all-individual data. Finally, we review some recently developed big data analysis methods including communication-efficient distributed approaches, renewal estimation, and incremental inference as examples of the latest developments in methods using calibration information.

## ARTICLE HISTORY

Received 4 January 2021  
Revised 14 December 2021  
Accepted 10 January 2022

## KEYWORDS

Calibration information;  
empirical likelihood;  
estimating equations;  
generalized method of  
moments; meta-analysis



## 1. Introduction

Statistical inference with big data can be extremely challenging owing to the high volume and large variety of observed quantities. Currently, one of the most popular approaches to this problem in statistics and computer science is the divide-and-conquer paradigm. The basic idea of this method is to break down a problem recursively into two or more sub-problems of the same or related type, such that each sub-problem becomes simple enough to be solved easily. The solution to the original problem is the optimal combination of the solutions to the sub-problems. A closely related statistical method is called parallel and distributed inference. In essence, large amounts of observed data are stored in different machines in a distributed manner. The computation is often relatively inexpensive in each machine. Then, communication is essential to enable assembly of the available results from all machines. Many related references can be found in, for example, Jordan et al. (2019). Although many new statistical methods have been developed for big data analysis, most of them have roots in traditional statistical methods of combining auxiliary information.

Combining information from similar studies has been and will continue to be an extremely important

strategy in statistical inference. The most popular example of such methods is meta-analysis, in which the published results of multiple similar scientific studies are pooled to produce an enhanced estimate without using the raw individual data from each study. We refer to Borenstein et al. (2009) for a comprehensive introduction to meta-analysis. For various reasons such as privacy or capacity of computer storage, in massive data inference, only summarized data rather than the original individual data may be available. This poses a very challenging problem: how to conduct efficient updated inference by making full use of the summarized data? In recent years, many methods of combining information have been developed in economic studies, machine learning, and distributed statistical inference. The goal of this paper is to selectively review a few popular methods that are able to integrate information in different disciplines.

Utilizing external summary data or auxiliary information to obtain more accurate inference is an old and effective method in survey sampling. Owing to restrictions such as cost effectiveness or convenience, the variable of interest  $Y$  may be available for only a small portion of individuals. However, the explanatory variable  $X$  associated with  $Y$  may readily be

**CONTACT** Yukun Liu  ykliu@sfs.ecnu.edu.cn  Room 105, Building of Law and Bussiness, Sorth Branch, No 500 Dongchuan Road, East China Normal University (Minhang Campus), Minhang District, Shanghai 200241, China

This article has been republished with minor changes. These changes do not impact the academic content of the article.

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

available for all individuals. Cochran (1977) presented a comprehensive discussion on regression-type estimators making use of the summarized information from X. Chen and Qin (1993), Chen et al. (2002), and Wu and Sitter (2001) used empirical likelihood (EL; Owen, 1988) to incorporate such information in finite populations.

With advances in technology, many summarized statistical results have become available in public domains. For example, many aggregated demographic and socioeconomic status data are provided in the US census reports. The Surveillance, Epidemiology, and End Results (SEER) programme of the National Cancer Institute provides population-based cancer survival statistics such as covariate-specific survival probabilities. Imbens and Lancaster (1994) combined micro and macro data in economic studies through the generalized method of moments (GMM). Chaudhuri et al. (2008) showed that inclusion of population-level information could reduce bias and increase the efficiency of the parameter estimates in a generalized linear model setup. Wu and Thompson (2020) published an excellent monograph on combining auxiliary information in survey sampling.

In this paper, we consider two situations. In the first, the summarized information from different studies was derived using the same statistical model. Second, the summarized information was derived using statistical models that were similar but not exactly the same. In general, combining information in the former case is easier. The latter case is more complex, as one has to take into consideration the heterogeneity among different studies.

The rest of this paper is organized as follows. In Section 2, we briefly review two simple and popular meta-analysis methods for combining similar results. In Section 3, we review Owen's (1988) EL method and Qin and Lawless's (1994) over-identified parameter problem as examples of general tools for synthesizing information from summarized data. In particular, we present a new way of deriving the lower information bound for the over-identified parameter problem. Section 4 discusses enhanced inference by utilizing auxiliary information. Section 5 presents results on more flexible meta-analyses where information on different covariates are available in similar studies. Calibration of information from previous studies is described in Section 6. We discuss methods of using disease prevalence information for more efficient estimation in case-control studies in Section 7. The popular communication-efficient distributed statistical inference method used in machine learning is discussed in Section 8. Renewal estimation and incremental inference are briefly presented in Section 9. Finally, some further discussion is presented in Section 10.

## 2. Two simple information-combining methods

### 2.1. Convex combination

Suppose that  $\hat{\theta}_1$  and  $\hat{\theta}_2$  are two asymptotically unbiased estimators for  $\theta$  from two independent studies, and that they satisfy  $\hat{\theta}_i \sim N(\theta, \sigma_i^2)$ ,  $i = 1, 2$ . The most straightforward way of combining  $\hat{\theta}_1$  and  $\hat{\theta}_2$  is a convex combination,

$$\hat{\theta} = \alpha \hat{\theta}_1 + (1 - \alpha) \hat{\theta}_2, \quad 0 < \alpha < 1.$$

The asymptotic variance of  $\hat{\theta}$  is  $\sigma^2 = \alpha^2 \sigma_1^2 + (1 - \alpha)^2 \sigma_2^2$ , which takes its minimum at  $\alpha = \sigma_2^2 / (\sigma_1^2 + \sigma_2^2)$ . This suggests combining  $\hat{\theta}_1$  and  $\hat{\theta}_2$  by

$$\hat{\theta} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \hat{\theta}_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \hat{\theta}_2 = \frac{\hat{\theta}_1 / \sigma_1^2 + \hat{\theta}_2 / \sigma_2^2}{1 / \sigma_1^2 + 1 / \sigma_2^2},$$

an inverse-variance weighting estimator. In general,  $\sigma_1^2$  and  $\sigma_2^2$  are unknown; we may replace them by their estimators  $\hat{\sigma}_1^2$  and  $\hat{\sigma}_2^2$ , respectively, which leads to

$$\hat{\theta} = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2} \hat{\theta}_1 + \frac{\hat{\sigma}_1^2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2} \hat{\theta}_2 = \frac{\hat{\theta}_1 / \hat{\sigma}_1^2 + \hat{\theta}_2 / \hat{\sigma}_2^2}{1 / \hat{\sigma}_1^2 + 1 / \hat{\sigma}_2^2}.$$

As an alternative method, we may use the maximum likelihood method to argue that this is the best estimator. We can treat  $\hat{\theta}_i$  as a direct observation from  $\hat{\theta}_i | \theta \sim N(\theta, \sigma_i^2)$ ,  $i = 1, 2$ . Then, the log-likelihood (regarding  $\sigma_1^2$  and  $\sigma_2^2$  as known constants) is

$$-(\hat{\theta}_1 - \theta)^2 / (2\sigma_1^2) - (\hat{\theta}_2 - \theta)^2 / (2\sigma_2^2).$$

Maximizing this likelihood with respect to  $\theta$  or setting the score function to be zero, we end up with the same inverse-variance weighting estimator.

### 2.2. Random-effect meta-analysis

Dersimonian and Laird (1986) proposed a moment-based estimation method using a random-effect model for meta-analysis. Let  $\hat{\theta}_i$  be an estimator of  $\theta_i$  from the  $i$ -th study,  $i = 1, 2, \dots, K$ . For example,  $\hat{\theta}_i$  could be the estimated mean response from the  $i$ -th study. When the sample size  $n_i$  in the  $i$ -th study is reasonably large, we may assume that

$$\begin{aligned} \hat{\theta}_i | \theta_i &\sim N(\theta_i, w_i^{-1}), \quad \theta_i \sim N(\theta, \tau^2), \\ i &= 1, 2, \dots, K, \end{aligned}$$

where the  $w_i^{-1}$ 's are treated as known. Although the normal models hold to be true approximately, we assume that they are all true for ease of theoretical development. The goal here is to better estimate  $\theta$  by combining the results from all the studies.

Unconditionally, we have  $\hat{\theta}_i \sim N(\theta, w_i^{-1} + \tau^2)$ . Consider the following inverse-variance weighting estimator for  $\theta$ :

$$\hat{\theta} = \frac{\sum_{i=1}^K \hat{\theta}_i w_i}{\sum_{i=1}^K w_i}$$

with variance

$$\text{Var}(\hat{\theta}) = \sum_{i=1}^K w_i^2 (w_i^{-1} + \tau^2) / \left( \sum_{i=1}^K w_i \right)^2.$$

Define

$$Q = \sum_{i=1}^K w_i (\hat{\theta}_i - \hat{\theta})^2 = \sum_{i=1}^K w_i (\hat{\theta}_i - \theta)^2 - (\hat{\theta} - \theta)^2 \sum_{i=1}^K w_i.$$

We can easily check that

$$\mathbb{E}(Q) = (K - 1) + \tau^2 \left( \sum_{i=1}^K w_i - \sum_{i=1}^K w_i^2 / \sum_{j=1}^K w_j \right),$$

which implies that a natural estimator of  $\tau^2$  is

$$\hat{\tau}^2 = \frac{Q - (K - 1)}{\sum_{i=1}^K w_i - \sum_{i=1}^K w_i^2 / \sum_{j=1}^K w_j}.$$

For small sample sizes, there is no guarantee that this estimator is non-negative; one may replace it by  $\max(\hat{\tau}^2, 0)$ .

Alternatively, we may estimate  $\tau$  using the likelihood approach. The joint likelihood based on the  $\hat{\theta}_i$ s is

$$\ell(\theta, \tau) = -\frac{1}{2} \sum_{i=1}^K \frac{(\hat{\theta}_i - \theta)^2}{\tau^2 + w_i^{-1}} - \frac{1}{2} \sum_{i=1}^K \log(\tau^2 + w_i^{-1}).$$

Maximizing  $\ell$  with respect to  $\theta$  and  $\tau^2$  gives their maximum likelihood estimators (MLEs).

Lin and Zeng (2010) compared the relative efficiency of using summary statistics versus individual-level data in meta-analysis. They found that in general there was no information loss when using the summarized information compared with inference based on the original individual data when available.

### 3. Empirical likelihood and general estimating equations

In this section we briefly review Owen's (1988) EL and Qin and Lawless' (1994) estimating equations approaches, as those methods represent general tools for assembly of information from different sources. The maximum likelihood method for regular parametric models is among the most popular methods in statistical inference, as it has many nice properties. However,

model mis-specification is a major concern, as a mis-specified model may lead to biased results. For the case when the underlying distribution is multinomial, Hartely and Rao (1968) proposed a mean constrained estimator for the population total in survey sampling problems. To mimic the parametric likelihood but discard parametric model assumptions, Owen (1988) and Owen (1990) proposed the EL method, which is a natural generalization of the multinomial likelihood when the number of categories is equal to the sample size. The EL approach can be thought of as a bootstrap that does not resample, or as a likelihood without parametric assumptions (Owen, 2001).

#### 3.1. Definition of empirical likelihood

Suppose that  $X_1, \dots, X_n$  are  $n$  independent and identically distributed observations from  $X$ , with cumulative distribution  $F$ . For convenience, we assume there are no ties, i.e., any two observations are unequal to each other. The techniques developed below can be easily adapted to handle ties. Let  $dF(X_i)$ ,  $i = 1, 2, \dots, n$ , be the jumps of  $F(x)$  at the observed data points. The nonparametric likelihood is  $L(F) = \prod_{i=1}^n p_i$ . It is clear that if any  $p_i = 0$ , then  $L(F) = 0$ , and if  $\sum_{i=1}^n p_i < 1$ , then  $L(F) < L(F_*)$ , where  $F_*(x) = \sum_{i=1}^n p_i I(X_i \leq x) / \sum_{i=1}^n p_i$ . According to the likelihood principle (that parameters with larger likelihoods are preferable), one need only consider the distribution functions  $F(x)$  with  $p_i > 0$  and  $\sum_{i=1}^n p_i = 1$ .

If we maximize the log-likelihood

$$\ell(F) = \sum_{i=1}^n \log p_i \quad (1)$$

subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0, \quad (2)$$

then we obtain  $p_i = 1/n$ ,  $i = 1, 2, \dots, n$ . Therefore, the maximum EL estimator of  $F$  is  $F_n(x) = \sum_{i=1}^n p_i I(X_i \leq x) = n^{-1} \sum_{i=1}^n I(X_i \leq x)$ . This is why the empirical distribution is called the nonparametric MLE of  $F(x)$ .

Suppose we are interested in constructing a confidence interval for  $\mu = \mathbb{E}(X) = \int x dF(x)$ , the mean of  $X$ . Since we have discretized  $F$  at each of the observed data points, the integral becomes  $\mu = \sum_{i=1}^n p_i X_i$ . Next, we maximize the nonparametric log-likelihood subject to an extra constraint:

$$\sum_{i=1}^n p_i (X_i - \mu) = 0. \quad (3)$$

Maximizing the log-likelihood (1) subject to constraints (2) and (3), the Lagrange multiplier method

gives the profile log-likelihood of  $\mu$ ,

$$\ell_n(\mu) = - \sum_{i=1}^n \log\{1 + \lambda^\top(X_i - \mu)\} - n \log n, \quad (4)$$

where  $\lambda$  is the solution to  $\sum_{i=1}^n (X_i - \mu)/\{1 + \lambda^\top(X_i - \mu)\} = 0$ . We can treat  $\ell_n(\mu)$  as a parametric likelihood of  $\mu$ . Based on this likelihood, the maximum EL estimator of  $\mu$  is  $\hat{\mu} = \bar{X} = n^{-1} \sum_{i=1}^n X_i$ , which is exactly the sample mean. We define the likelihood ratio function as

$$R_n(\mu) = 2\{\max_{\mu} \ell_n(\mu) - \ell_n(\mu)\} = 2\{\ell_n(\bar{X}) - \ell_n(\mu)\}.$$

Under the regularity conditions specified in Owen (1988) and Owen (1990), as  $n$  goes to infinity,  $R_n(\mu_0)$  converges to the  $\chi^2$  distribution with  $p$  degrees of freedom, where  $p$  is the dimension of  $\mu$ , and  $\mu_0$  is the true value of  $\mu$ .

### 3.2. General estimating equations

The original EL was mainly used to make inference for linear functionals of the underlying population distribution such as the population mean (Owen, 1988, 1990). Qin and Lawless (1994) applied this method to general estimating models, which greatly broadened its applications. Specifically, suppose the population of interest satisfies a general estimating equation

$$\mathbb{E}\{g(X, \theta)\} = 0, \quad (5)$$

for a  $r \times 1$  vector-valued function  $g$  and some  $\theta$ , which is a  $p \times 1$  parameter to be estimated. We assume  $r \geq p$  as otherwise the true parameter value of  $\theta$  would be undefined.

For general estimating equations with  $r > p$  or over-identified models, Hansen (1982) proposed the celebrated GMM, which has become one of the most popular methods in the econometric community. In essence, the GMM minimizes

$$\left\{ \sum_{i=1}^n g(X_i, \theta) \right\}^\top \Sigma^{-1} \left\{ \sum_{i=1}^n g(X_i, \theta) \right\}$$

with respect to  $\theta$ , where  $\Sigma$  is the variance matrix of the estimating equation  $g(X, \theta)$ . If  $\Sigma$  is unknown, we may replace it by the sample variance  $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n g(X_i, \tilde{\theta})g^\top(X_i, \tilde{\theta})$ , where  $\tilde{\theta}$  is an initial and consistent estimate of  $\theta$ .

Instead of GMM, Qin and Lawless (1994) used the EL to make inferences for parameters defined by a general estimating equation. For discretized  $F(x)$  satisfying (2), Equation (5) becomes

$$\sum_{i=1}^n p_i g(X_i, \theta) = 0. \quad (6)$$

Maximizing the log-likelihood (1) subject to (2) and (6), we have the following profile log-likelihood of

$\theta$  (up to a constant):

$$\ell_n(\theta) = - \sum_{i=1}^n \log\{1 + \lambda^\top g(X_i, \theta)\},$$

where  $\lambda$  is the Lagrange multiplier determined by  $\sum_{i=1}^n g(X_i, \theta)/\{1 + \lambda^\top g(X_i, \theta)\} = 0$ . We then estimate  $\theta$  by the maximizer  $\hat{\theta} = \arg \max_{\theta} \ell_n(\theta)$ , whose limiting distribution is established in the following theorem. Hereafter, we use  $\nabla_{\theta}$  to denote the differentiation operator with respect to  $\theta$ .

**Theorem 3.1 (Qin & Lawless, 1994):** Denote  $g = g(X, \theta_0)$  and  $\nabla_{\theta^\top} g = \nabla_{\theta^\top} g(X, \theta_0)$ . Suppose that (1)  $\mathbb{E}(gg^\top)$  is positive definite, (2)  $\nabla_{\theta^\top} g(X, \theta)$  is continuous in a neighbourhood of  $\theta_0$ , (3)  $\|\nabla_{\theta^\top} g(X, \theta)\|$  and  $\|g(X, \theta)\|^3$  can be bounded by some integrable function  $G(X)$  in this neighbourhood, and (4)  $\mathbb{E}(\nabla_{\theta^\top} g)$  is of full rank. Then, as  $n \rightarrow \infty$ ,  $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V)$ , where  $\xrightarrow{d}$  means ‘convergence in distribution’ and

$$V = \{\mathbb{E}(\nabla_{\theta^\top} g) (\mathbb{E}gg^\top)^{-1} \mathbb{E}(\nabla_{\theta^\top} g)\}^{-1}. \quad (7)$$

### 3.3. Calculation of the information bound

Assuming that the parameter of interest satisfies the general estimating equation  $\mathbb{E}\{g(X, \theta)\} = 0$ , we next consider how well we can estimate  $\theta$  based on this model, and whether the maximum EL estimator is optimal. To answer these questions, we consider an *ideal* situation, where the probability function  $X$  has a parametric form  $f(x, \theta)$ , which is known up to  $\theta$ . We define

$$h(x, \eta, \theta) = \exp\{\eta^\top g(x, \theta)\} f(x, \theta) \int \exp\{\eta^\top g(t, \theta)\} f(t, \theta) dt,$$

implicitly assuming that  $\int \exp\{\eta^\top g(t, \theta)\} f(t, \theta) dt < \infty$ . Clearly,  $h(x, \eta, \theta)$  is an enlarged parametric model of  $f(x, \theta)$  as it reduces to  $f(x, \theta)$  when  $\eta = 0$ . As the parametric form  $f(x, \theta)$  is unknown in practice, we anticipate that any estimator based on the moment constraints  $\mathbb{E}\{g(X, \theta)\} = 0$  should have a variance that is no less than that of the MLE derived from the enlarged model. We show that even if the form of  $f(x, \theta)$  is available, the MLE of  $\theta$  based on  $h(x, \eta, \theta)$  has the same asymptotic variance as the maximum EL estimator.

With the parametric model  $h$ , we can estimate  $\theta$  by maximizing  $L(\theta, \eta) = \prod_{i=1}^n h(X_i, \eta, \theta)$  with respect to  $(\theta, \eta)$ . We denote the resulting MLE by  $(\tilde{\eta}, \tilde{\theta})$ . We show in Section 3.4 that under some regularity conditions on  $h$  (see, e.g., Theorems 14 and 23 of van de Vaart (2000)), as  $n \rightarrow \infty$ ,

$$\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} N(0, V), \quad (8)$$

where  $V$  is defined in (7). In general, the parametric form  $f(x, \theta)$  is unknown; hence, we expect that the best



estimator of  $\theta$  should have an asymptotic variance at least as large as  $V$ . As the maximum EL estimator of  $\theta$  of Qin and Lawless (1994) has asymptotic variance  $V$ , we conclude that it achieves the lower information bound.

**Remark 3.1:** If  $g(x, \theta)$  is an unbounded function of  $x$  for each  $\theta$ , we may construct a new density

$$h(x, \theta, \eta) = \frac{\psi\{\eta^\top g(x, \theta)\}f(x, \theta)}{\int \psi\{\eta^\top g(x, \theta)\}f(x, \theta)dx},$$

where  $\psi(x) = 2(1 + e^{-2x})^{-1}$  with  $\psi(0) = \psi'(0) = 1$ . Clearly,  $\psi$  is bounded. We may go through the same derivations to get the same conclusion.

**Remark 3.2:** Back and Brown (1992) established a similar result by constructing an exponential family. In particular, they defined  $h(x, \theta) = \exp\{\xi^\top(\theta)g(x, \theta_0) - a(\theta)\}f_0(x)$ , where  $f_0(x) = f(x, \theta_0)$  and  $\xi(\theta)$  is determined implicitly by the following conditions:  $\xi(\theta_0) = 0$ ,  $a(\theta_0) = 0$ ,  $\int \exp\{\xi^\top(\theta)g(x, \theta_0) - a(\theta)\}f_0(x)dx = 1$ , and  $\int g(x, \theta) \exp\{\xi^\top(\theta)g(x, \theta_0) - a(\theta)\}f_0(x)dx = 0$ . In Back & Brown's approach,  $\xi(\theta)$  is determined implicitly by the above constraint equation, whereas in our new approach,  $\eta$  is an independent parameter.

### 3.4. A sketched proof of (8)

The log-likelihood based on the enlarged model is  $\ell(\theta, \eta) = \sum_{i=1}^n \log\{h(X_i, \eta, \theta)\}$ , where

$$\begin{aligned} \log\{h(x, \eta, \theta)\} &= \eta^\top g(x, \theta) + \log f(x, \theta) \\ &\quad - \log \left[ \int \exp\{\eta^\top g(t, \theta)\}f(t, \theta)dt \right]. \end{aligned}$$

If  $\log\{h(x, \eta, \theta)\}$  satisfies the conditions of Theorem 14 of van de Vaart (2000) on  $m_\theta(x)$ , then  $(\tilde{\theta}, \tilde{\eta})$  is consistent with  $(\theta_0, 0)$ .

Result (8) follows from Theorem 23 of van de Vaart (2000). With tedious algebra, we find that

$$\begin{aligned} \nabla_\theta \log\{h(x, \theta_0, 0)\} &= 0, \\ \nabla_\eta \log\{h(x, \theta_0, 0)\} &= g(x, \theta_0), \\ \mathbb{E}\nabla_{\theta\theta^\top} \log\{h(X, \theta_0, 0)\} &= 0, \\ \mathbb{E}\nabla_{\eta\eta^\top} \log\{h(X, \theta_0, 0)\} &= -\mathbb{E}(gg^\top), \\ \mathbb{E}\nabla_{\theta\eta^\top} \log\{h(X, \theta_0, 0)\} &= \mathbb{E}\nabla_{\theta^\top} g(X, \theta_0) \\ &\quad - \mathbb{E}\{\nabla_{\theta^\top} g(X, \theta_0) + g(X, \theta_0)\nabla_{\theta^\top} \log f(X, \theta_0)\}. \end{aligned}$$

Under some mild assumptions, such as that  $\int g(x, \theta)f(x, \theta)dx = 0$  holds for  $\theta$  in a neighbourhood of  $\theta_0$ , differentiating both sides with respect to  $\theta$  leads to

$$\mathbb{E}\{\nabla_\theta g(X, \theta)\} + \mathbb{E}\{\nabla_\theta g(X, \theta)\nabla_\theta \log f(X, \theta)\} = 0,$$

which means  $\mathbb{E}\nabla_{\eta\theta^\top} \log\{h(X, \theta_0, 0)\} = \mathbb{E}\{\nabla_{\theta^\top} g(X, \theta_0)\}$ . As  $(\tilde{\theta}, \tilde{\eta})$  is consistent with  $(\theta_0, 0)$ , by

Theorem 5.23 of van de Vaart (2000), we have

$$\begin{aligned} \sqrt{n} \begin{pmatrix} \tilde{\theta} - \theta_0 \\ \tilde{\eta} - 0 \end{pmatrix} &= \begin{pmatrix} 0 & \mathbb{E}(\nabla_\theta g^\top) \\ \mathbb{E}(\nabla_{\theta^\top} g) & -\mathbb{E}(gg^\top) \end{pmatrix}^{-1} \\ &\quad \begin{pmatrix} 0 \\ n^{-1/2} \sum_{i=1}^n g(X_i, \theta_0) \end{pmatrix} + o_p(1). \end{aligned} \quad (9)$$

This, together with the fact that  $n^{-1/2} \sum_{i=1}^n g(X_i, \theta_0) \xrightarrow{d} N(0, \mathbb{E}(gg^\top))$  as  $n$  goes to infinity, implies (8).

### 3.5. Empirical entropy family

Again we assume that the available information is given by the estimating equation  $\mathbb{E}\{g(X, \theta)\} = 0$ . The enlarged parametric model  $h(x, \eta, \theta)$  satisfies

$$\int h(x, \eta, \theta)g(x, \theta)dx = 0$$

only if  $\eta = 0$ . Naturally, one may require  $\eta = \eta(\theta)$  to satisfy

$$\int g(x, \theta) \exp\{\eta^\top g(x, \theta)\}f(x, \theta)dx = 0.$$

It is often too restrictive to assume a known underlying parametric model  $f(x, \theta)$  in the construction of the enlarged parametric model  $h(x, \eta, \theta)$ . We may replace the cumulative distribution function  $F(x, \theta) = \int_{-\infty}^x f(t, \theta)dt$  by the empirical distribution  $F_n(x) = n^{-1} \sum_{i=1}^n I(X_i \leq x)$ . In this situation,  $\eta = \eta(\theta)$  is the solution to  $\sum_{i=1}^n g(x_i, \theta) \exp\{\eta^\top g(x_i, \theta)\} = 0$ .

Let  $H(x, \eta, \theta) = \int_{-\infty}^x h(t, \eta, \theta)dt$ . For fixed parameter values  $(\eta, \theta)$ , the jump of  $H$  at  $x = X_i$  is

$$\begin{aligned} dH(X_i, \eta, \theta) &= \exp\{\eta^\top(\theta)g(X_i, \theta)\} \Bigg/ \\ &\quad \left[ \sum_{j=1}^n \exp\{\eta^\top(\theta)g(X_j, \theta)\} \right], \end{aligned}$$

and the likelihood becomes

$$\prod_{i=1}^n dH(X_i, \eta, \theta) = \prod_{i=1}^n \frac{\exp\{\eta^\top(\theta)g(X_i, \theta)\}}{\sum_{j=1}^n \exp\{\eta^\top(\theta)g(X_j, \theta)\}}.$$

In fact, this is equivalent to the EL  $\prod_{i=1}^n p_i$ , where the  $p_i$ s minimize the Kullback–Leibler divergence (up to a constant) or minus the exponential titling likelihood  $\sum_{i=1}^n p_i \log(p_i)$  subject to the constraints  $\sum_{i=1}^n p_i = 1$ ,  $p_i \geq 0$ , and  $\sum_{i=1}^n p_i g(X_i, \theta) = 0$ . See Susanne (2007) for more details. We call this the empirical entropy family induced by the estimating equation  $\mathbb{E}\{g(X, \theta)\} = 0$ .

#### 4. Enhancing efficiency using auxiliary information

In this section, we discuss methods of incorporating auxiliary information to enhance estimation efficiency. This aspect was also investigated by Qin (2000). We assume a parametric model  $f(y|x, \beta)$  for the conditional density function of  $Y$  given  $X$  and leave the marginal distribution  $G(x)$  of  $X$  unspecified. We wish to make inferences for  $\beta$  when some auxiliary information is summarized through an estimating equation

$$\mathbb{E}\{\phi(X, \beta)\} = 0.$$

For example, if we know the mean  $\mu$  of  $Y$ , then we can construct an estimating equation  $\mathbb{E}(Y - \mu) = 0$ . We can take

$$\begin{aligned}\phi(X, \beta) &= \int (y - \mu)f(y|X, \beta) dy \\ &= \int yf(y|X, \beta) dy - \mu.\end{aligned}$$

Furthermore, we allow that the response  $Y$  may have missing values. Let  $D$  be the non-missingness indicator, which takes the value 1 if  $Y$  is available, and 0 otherwise. We assume a missing-at-random model

$$\begin{aligned}\text{pr}(D = 1 | Y = y, X = x) &= \text{pr}(D = 1 | X = x) \\ &= \pi(x),\end{aligned}$$

where  $\pi(x)$  depends only on  $x$ . We denote the observed data by  $(d_i, d_i y_i, x_i)$  ( $i = 1, 2, \dots, n$ ) and  $p_i = dG(x_i)$ . The likelihood of  $(\beta, G)$  is

$$\begin{aligned}L &= \prod_{i=1}^n \{\pi(x_i)f(y_i|x_i, \beta) dG(x_i)\}^{d_i} \\ &\quad \times \{[1 - \pi(x_i)]dG(x_i)\}^{1-d_i} \\ &= \prod_{j=1}^n \{\pi(x_j)\}^{d_j} \{1 - \pi(x_j)\}^{1-d_j} \\ &\quad \cdot \prod_{i=1}^n \{f(y_i|x_i, \beta)\}^{d_i} \cdot p_i.\end{aligned}$$

We can maximize this likelihood subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0, \quad \sum_{i=1}^n p_i \phi(x_i, \beta) = 0.$$

As  $\prod_{j=1}^n \{\pi(x_j)\}^{d_j} \{1 - \pi(x_j)\}^{1-d_j}$  is not a function of  $\beta$ , the profile hybrid empirical log-likelihood (up to a constant) is

$$\ell(\beta) = \sum_{i=1}^n [d_i \log f(y_i|x_i, \beta) - \log\{1 + \lambda^\top \phi(x_i, \beta)\}], \quad (10)$$

where  $\lambda$  is the Lagrange multiplier determined by

$$\sum_{i=1}^n \frac{\phi(x_i, \beta)}{1 + \lambda^\top \phi(x_i, \beta)} = 0. \quad (11)$$

For the special case where data are missing completely at random, i.e.,  $\pi(x)$  is a constant function of  $x$ , Qin (2000) established the following theorem.

**Theorem 4.1:** Let  $\beta_0$  be the true parameter value, let  $\hat{\beta}$  be the maximum hybrid EL estimator, i.e., the maximizer of (10), and let  $\hat{\lambda}$  be the corresponding Lagrange multiplier. Denote  $\phi = \phi(X, \beta_0)$ ,  $\nabla_\beta \phi = \nabla_\beta \phi(X, \beta_0)$ , and

$$\begin{aligned}J &= -\mathbb{E}\{d_i \nabla_{\beta\beta^\top} \log f(y_i|x_i, \beta_0)\} \\ &= \mathbb{V}\text{ar}\{d_i \nabla_\beta \log f(y_i|x_i, \beta_0)\}.\end{aligned}$$

Under some regularity conditions, when  $n$  goes to infinity, we have

$$\sqrt{n}((\hat{\beta} - \beta_0)^\top, \hat{\lambda}^\top)^\top \xrightarrow{d} N(0, \Sigma),$$

where  $\Sigma = \text{diag}(\Sigma_{11}, \Sigma_{22})$  with

$$\begin{aligned}\Sigma_{11} &= \{J + \mathbb{E}(\nabla_\beta \phi^\top)(\mathbb{E}\phi\phi^\top)^{-1}\mathbb{E}(\nabla_\beta \phi)\}^{-1}, \\ \Sigma_{22} &= \{\mathbb{E}(\nabla_\beta \phi^\top)J^{-1}\mathbb{E}(\nabla_\beta \phi) + \mathbb{E}(\phi\phi^\top)\}^{-1}.\end{aligned} \quad (12)$$

**Remark 4.1:** Imbens and Lancaster (1994) studied the same problem using GMM. In particular, they directly combined the conditional score estimating equation  $\nabla_\beta \log f(y|x, \beta)$  and  $\phi(x, \beta)$ . Even though the first-order large-sample results are the same, the hybrid EL based approach is more appealing as it respects the parametric conditional likelihood and replaces only the marginal likelihood with the EL. See Qin (2000) for numerical comparisons of results of the two methods.

#### 5. Combining summary information: a more flexible method for meta-analysis

Developing systematic methods for combining published information is one of the main goals of meta-analysis, which has become increasingly popular since little extra cost is needed. The main restriction in meta-analysis is that all studies must include the same variables in their analyses. The only difference allowed is in the sample sizes. Thus, studies must be discarded if they contain different variables from those in other studies.

Summarized information is often available from publications such as census reports and results of national health studies. For reasons including confidentiality, it is typically not possible to gain access to the original data, only the summarized reports. Suppose we are interested in conducting a new study that may contain some new variables of interest that are not available in the summarized information, for example, a genetic

study involving newly discovered biomarkers or genes. Below we discuss a more flexible method that could be used to combine published information and individual study data for enhanced inference in such cases. Chatterjee et al. (2016) discussed a related problem on the utilization of auxiliary information. As Han and Lawless (2016) pointed out, however, their methodology and theoretical results had already been developed by Imbens and Lancaster (1994) and Qin (2000) in the absence of selection bias in sampling.

We consider two cases. (I) The sample size for the summarized information is much larger than that of the new study. (II) Sample sizes from the two data sources are comparable. In Case I, we can treat the summarized information as known, i.e., the variation in the summarized data is negligible compared with the variation in the new study. In Case II, we have to take the variation in the summarized information into consideration as it is comparable to the variation in the new study. We focus on Case I in this section and study Case II in Section 6.

### 5.1. Setup and solution

Suppose that the summarized results were obtained from statistical analyses of response  $Y$  and covariate variables  $X$  (although the original data are not available), and that the new study includes an extra covariate  $Z$  in addition to  $(Y, X)$ . We are interested in fitting a parametric model  $f(y|x, z, \beta)$  for the conditional density function of  $Y$  given  $X$  and  $Z$ . Let  $(y_1^*, x_1^*), \dots, (y_N^*, x_N^*)$  be the historic data even though they are unavailable. The published information can be summarized in two ways:

- (I)  $\bar{h} = N^{-1} \sum_{i=1}^N h(y_i^*, x_i^*)$  is known; and
- (II)  $\gamma^*$  is the solution of an estimating equation  $\sum_{i=1}^N h(y_i^*, x_i^*, \gamma) = 0$ , where the function  $h(y, x, \gamma)$  is known up to  $\gamma$ .

Let  $(y_1, x_1, z_1), \dots, (y_n, x_n, z_n)$  be observed data from the new study. The basic assumption is that  $(y_i, x_i), i = 1, 2, \dots, n$ , and  $(y_i^*, x_i^*)$  have the same distribution. To utilize the summarized information, we can define estimating functions

$$g = (g_1, g_3), \quad g_1(y, x, z) = \nabla_\beta \log f(y|x, z, \beta),$$

$$g_3(y, x) = h(y, x) - \bar{h}$$

in Scenario (I), and

$$g = (g_1, g_3),$$

$$g_1(y, x, z) = \nabla_\beta \log f(y|x, z, \beta),$$

$$g_3(y, x) = h(y, x, \gamma^*)$$

in Scenario (II). We consider only the situation where  $n/N \rightarrow 0$ . In other words, the variation in the auxiliary information is negligible.

The EL approach amounts to maximizing  $\sum_{i=1}^n \log p_i$  subject to the constraint

$$\sum_{i=1}^n p_i g(y_i, x_i, z_i, \beta) = 0, \quad p_i \geq 0, \quad \sum_{i=1}^n p_i = 1.$$

According to Qin and Lawless (1994), the asymptotic variance of the maximum EL estimator  $\hat{\beta}$  based on estimating equation  $g$  is

$$[\mathbb{E}(\nabla_\beta g^\top) \{\mathbb{E}(gg^\top)\}^{-1} \mathbb{E}(\nabla_\beta^\top g)]^{-1},$$

where  $\nabla_\beta g = \partial g(y, x, z, \beta) / \partial \beta|_{\beta=\beta_0}$ ,  $g = g(y, x, z, \beta_0)$ , and  $\beta_0$  is the truth of  $\beta$ . we denote

$$A = \mathbb{E}(gg^\top) = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^\top & A_{22} \end{pmatrix},$$

$$A_{22.1} = A_{22} - A_{12}^\top A_{11}^{-1} A_{12}.$$

Equivalently, the asymptotic variance can be written as

$$[\mathbb{E}(\nabla_\beta g_1^\top) A_{11}^{-1} \mathbb{E}(\nabla_\beta^\top g_1) + \mathbb{E}(\nabla_\beta g_1^\top) A_{11}^{-1} A_{12} \\ \times A_{22.1}^{-1} A_{21} A_{11}^{-1} \mathbb{E}(\nabla_\beta^\top g_1)]^{-1},$$

or  $(J + A_{12} A_{22.1}^{-1} A_{21})^{-1}$ , where  $A_{11} = J$  is Fisher's information matrix.

In the above approach, the estimating equation  $g_3 = h(y, x) - \bar{h}$  does not involve the parameter  $\beta$ . However, there are ways to achieve higher efficiency. For example, we define  $g_2(x, z, \beta) = \psi(x, z, \beta)$  with

$$\psi(x, z, \beta) = \mathbb{E}\{h(Y, X) | X = x, Z = z\} - \bar{h} \\ = \int h(y, x) f(y|x, z, \beta) dy - \bar{h}.$$

Then,  $\mathbb{E}\{g_2(x, z, \beta)\} = 0$ . If we combine the empirical log-likelihood based on the estimating equation  $g_2$  and the log-likelihood  $\sum_{i=1}^n \log f(y_i | x_i, z_i, \beta)$  as in the previous section (see Equation (12)), then the asymptotic variance of the resulting MLE  $\hat{\beta}$  is given by

$$\{J + \mathbb{E}(\nabla_\beta \psi^\top) (\mathbb{E} \psi \psi^\top)^{-1} \mathbb{E}(\nabla_\beta^\top \psi)\}^{-1}.$$

In general, this approach can achieve better efficiency.

### 5.2. A comparison

Given two pairs of estimation functions,  $\{g_1, g_3\}$  and  $\{g_1, g_2\}$ , we may wonder combining which pair leads to a better estimator if we directly compare their asymptotic variance formulae. Alternatively, we may enquire whether we should combine all three constraints  $g = (g_1, g_2, g_3)$  together. Write  $g_{12} = g_{21}^\top = (g_1, g_2)$ ,  $a = \mathbb{E}\{h^\top(y, x) \nabla_\beta \log f(y|x, z, \beta)\}$ , and

$$\mathbb{E}(gg^\top) = \begin{pmatrix} J & 0 & a \\ 0 & \mathbb{E}(\psi \psi^\top) & \mathbb{E}(\psi \psi^\top) \\ a^\top & \mathbb{E}(\psi \psi^\top) & \mathbb{E}(h h^\top) \end{pmatrix}$$



$$= \begin{pmatrix} B_{11} & B_{12} \\ B_{12}^\top & B_{22} \end{pmatrix},$$

$$B_{11} = \begin{pmatrix} J & 0 \\ 0 & \mathbb{E}(\psi\psi^\top) \end{pmatrix}, \quad B_{12} = \begin{pmatrix} a \\ \mathbb{E}(\psi\psi^\top) \end{pmatrix}.$$

Using results from Qin and Lawless (1994) and

$$\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}^{-1} = \begin{pmatrix} I & -B_{11}^{-1}B_{12} \\ 0 & I \end{pmatrix} \\ \times \begin{pmatrix} B_{11}^{-1} & 0 \\ 0 & B_{22.1}^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -B_{21}B_{11}^{-1} & I \end{pmatrix}$$

with  $B_{22.1} = B_{11} - B_{12}^\top B_{11}^{-1} B_{12}$ , we find that the asymptotic variance of  $\hat{\beta}$  obtained by combining the three estimating equations and  $\sum_{i=1}^n \log f(y_i | x_i, z_i, \beta)$  is

$$[J + \mathbb{E}(\nabla_\beta \psi^\top) \{\mathbb{E}(\psi\psi^\top)\}^{-1} \mathbb{E}(\nabla_\beta \psi) \\ + \mathbb{E}(\nabla_\beta g_{21}) B_{11}^{-1} B_{12} B_{22.1}^{-1} B_{21} B_{11}^{-1} \mathbb{E}(\nabla_\beta g_{12})]^{-1}.$$

It can be shown that  $\mathbb{E}(\nabla_\beta g) = (-J, \mathbb{E}(\nabla_\beta \psi), 0)$  and  $\mathbb{E}(\nabla_\beta g_{12}) = (-J, a)$ . Immediately, we have

$$\mathbb{E}(\nabla_\beta g_{12}) B_{11}^{-1} B_{12} = (-J, a) \begin{pmatrix} J^{-1} & 0 \\ 0 & \{\mathbb{E}(\psi\psi^\top)\}^{-1} \end{pmatrix} \\ \times \begin{pmatrix} a \\ \mathbb{E}(\psi\psi^\top) \end{pmatrix} = 0,$$

which implies that the asymptotic variance in the case where  $g_1, g_2$ , and  $g_3$  are combined is the same as that in the case where  $g_1$  and  $g_2$  only are combined. This indicates that taking  $g_3$  into account leads to no efficiency gain in the estimation of  $\beta$ .

The method of combining  $g_2$  and the parametric likelihood  $\prod_{i=1}^n f(y_i | x_i, z_i, \beta)$  is better than that of combining  $g_1, g_3$ , and the parametric likelihood. To see this, recall that the asymptotical variances for the MLEs of  $\beta$  with the two methods are

$$V_1 = \{J + \mathbb{E}(\nabla_\beta \psi^\top) \{\mathbb{E}(\psi\psi^\top)\}^{-1} \mathbb{E}(\nabla_\beta \psi)\}^{-1}$$

and

$$V_2 = (J + A_{12} A_{22.1}^{-1} A_{21})^{-1}.$$

It suffices to show that  $V_2 - V_1 \geq 0$ , namely,  $V_2 - V_1$  is non-negative definite.

### 5.3. Proof of $V_2 - V_1 \geq 0$

For convenience, we assume that  $\mathbb{E}(h) = 0$ . As  $\mathbb{E}(\nabla_\beta \psi^\top) = A_{12}$  and  $\psi = \mathbb{E}(h | X, Z)$ , it suffices to show that

$$A_{22.1} - \mathbb{E}(\psi\psi^\top) = (A_{22} - A_{21} A_{11}^{-1} A_{12}) \\ - \mathbb{E}[\{\mathbb{E}(h | X, Z)\}^{\otimes 2}] \geq 0. \quad (13)$$

Let  $\mathbb{E}_*$  and  $\mathbb{Var}_*$  denote  $\mathbb{E}(\cdot | X, Z)$  and  $\mathbb{Var}(\cdot | X, Z)$ , respectively. As

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \mathbb{E} \left\{ \begin{pmatrix} g_1 \\ h \end{pmatrix}^{\otimes 2} \right\}$$

$$= \mathbb{E} \left\{ \mathbb{Var}_* \begin{pmatrix} g_1 \\ h \end{pmatrix} \right\} + \mathbb{Var} \left\{ \mathbb{E}_* \begin{pmatrix} g_1 \\ h \end{pmatrix} \right\}$$

and  $\mathbb{E}_*(g_1) = 0$ , it follows that

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \geq \mathbb{Var} \left\{ \mathbb{E}_* \begin{pmatrix} g_1 \\ h \end{pmatrix} \right\} \\ = \mathbb{E} \begin{pmatrix} 0 & 0 \\ 0 & \mathbb{E}_*(h) \mathbb{E}_*(h^\top) \end{pmatrix}.$$

Multiplying both sides by  $(-A_{21} A_{11}^{-1}, I)$  from the left and by  $(-A_{21} A_{11}^{-1}, I)^\top$  from the right, we arrive at

$$A_{22} - A_{21} A_{11}^{-1} A_{12} \geq \mathbb{E} \{ \mathbb{E}_*(h) \mathbb{E}_*(h^\top) \},$$

that is, inequality (13) holds, which implies  $V_2 - V_1 \geq 0$ .

## 6. Calibration of information from previous studies

We consider calibration of information using parametric likelihood, EL (Owen, 1988), and GMM (Hansen, 1982). When only summary information from previous studies is available, these three well-known methods can be used to calibrate such summary information and to make inferences about the unknown parameters of interest. We may wonder whether doing so results in efficiency loss compared with inferences based on the pooled data if they were all available. Zeng and Lin (2015) found that parametric-likelihood-based meta-analysis of summarized information retained first-order asymptotic efficiency compared with analysis based on individual data. We show here that EL and GMM also possess this property. This is extremely important, as individual data may involve privacy issues, whereas summarized information does not.

### 6.1. Efficiency comparison

Suppose that  $(Y_{ij}, X_{ij})$  ( $j = 1, 2, \dots, n_i; i = 1, 2, \dots, K$ ) are independent observations from the same population. We consider two scenarios according to the model's assumption about the population.

- (I) The conditional probability function (i.e., the probability density/mass function of a continuous/discrete random variable) of  $Y$  given  $X$  has a parametric form  $f(y | x, \beta)$ .
- (II) The population satisfies  $\mathbb{E}\{g(Y, X, \beta)\} = 0$ .

Here,  $\beta$  is a finite-dimensional unknown parameter, and  $\beta_*$  is its true value. Assume that data are available batch by batch, and that  $n_i/n = \rho_i \in (0, 1)$ , where  $n = \sum_{i=1}^K n_i$ . For the  $i$ -th batch ( $i = 1, 2, \dots, K$ ) of data:

- (a) under assumption (I), the parametric log-likelihood function of  $\beta$  is

$$\ell_{i,PL}(\beta) = \sum_{j=1}^{n_i} \log\{f(Y_{ij} | X_{ij}, \beta)\};$$

- (b) under assumption (II), we define an empirical log-likelihood function

$$\begin{aligned} \ell_{i,EL}(\beta) &= \sup \left\{ \sum_{j=1}^{n_i} \log(n_i p_j) : p_j \geq 0, \right. \\ &\quad \left. \sum_{j=1}^{n_i} p_j = 1, \right. \\ &\quad \left. \sum_{j=1}^{n_i} p_j g(Y_{ij}, X_{ij}; \beta) = 0 \right\} \\ &= - \sum_{j=1}^{n_i} \log\{1 + \lambda_i^\top g(Y_{ij}, X_{ij}; \beta)\} \\ &\quad - n_i \log(n_i), \end{aligned}$$

where  $\lambda_i$  satisfies  $\sum_{j=1}^{n_i} \frac{g(Y_{ij}, X_{ij}; \beta)}{1 + \lambda_i^\top g(Y_{ij}, X_{ij}; \beta)} = 0$ ;

- (c) under assumption (II), we define the objective function of the GMM method (GMM log-likelihood for short) as

$$\begin{aligned} \ell_{i,GMM}(\beta) &= - \left\{ \sum_{j=1}^{n_i} g(Y_{ij}, X_{ij}; \beta) \right\}^\top \\ &\quad \times \Omega^{-1} \left\{ \sum_{j=1}^{n_i} g(Y_{ij}, X_{ij}; \beta) \right\}, \end{aligned}$$

where  $\Omega = \text{Var}\{g(Y, X, \beta_*)\}$  and  $\beta_*$  is the true value of  $\beta$ . In practice,  $\beta_*$  is generally replaced by a consistent estimator of  $\beta$  in the expression for  $\Omega$ . Using the true value  $\beta_*$  of  $\beta$  does not affect the theoretical analysis presented in this section.

Let  $\ell_i(\beta) = \ell_{i,PL}(\beta)$ ,  $\ell_{i,EL}(\beta)$ , or  $\ell_{i,GMM}(\beta)$ . Under certain regularity conditions, it can be verified that for  $\beta = \beta_* + O_p(n^{-1/2})$ ,

$$\begin{aligned} \ell_i(\beta) &= U_i^\top \sqrt{n_i}(\beta - \beta_*) - \frac{n_i}{2}(\beta - \beta_*)^\top \\ &\quad \times V(\beta - \beta_*) + o_p(1). \end{aligned} \quad (14)$$

In Case (a),

$$\begin{aligned} U_i &= n_i^{-\frac{1}{2}} \sum_{j=1}^{n_i} \nabla_\beta \log\{f(Y_{ij} | X_{ij}, \beta_*)\}, \\ V &= \text{Var}[\nabla_\beta \log\{f(Y | X, \beta_*)\}]. \end{aligned}$$

In Case (b)

$$U_i = n_i^{-\frac{1}{2}} \sum_{j=1}^{n_i} g(Y_{ij}, X_{ij}; \beta_*), \quad V = A_{12} A_{22}^{-1} A_{21},$$

where

$$\begin{aligned} A &= \begin{pmatrix} 0 & \mathbb{E}\{\nabla_\beta g^\top(Y, X; \beta_*)\} \\ \mathbb{E}\{\nabla_\beta g(Y, X; \beta_*)\} & \mathbb{E}\{g(Y, X; \beta_*)g(Y, X; \beta_*)\} \end{pmatrix} \\ &\equiv \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}. \end{aligned}$$

In Case (c),

$$\begin{aligned} U_i &= -\{\mathbb{E}\nabla_\theta g^\top(Y, X, \beta_*)\} \Omega^{-1} n_i^{-\frac{1}{2}} \sum_{j=1}^{n_i} g(Y_{ij}, X_{ij}, \beta_*), \\ V &= \{\mathbb{E}\nabla_\theta g^\top(Y, X, \beta_*)\} \Omega^{-1} \{\mathbb{E}\nabla_\beta g(Y, X, \beta_*)\}. \end{aligned}$$

We denote the MLE of  $\beta$  based on the  $r$ -th batch of data by  $\hat{\beta}_i = \arg \max \ell_r(\beta)$ . The above approximation implies that

$$\sqrt{n_i}(\hat{\beta}_i - \beta_*) = V^{-1} U_i + o_p(1) \xrightarrow{d} N(0, V^{-1}).$$

When the  $K$ -th batch of individual data are available, we no longer have access to the individual data of the previous  $K-1$  batches but only have summarized information  $(\hat{\beta}_i, \hat{\Sigma}_i)$ ,  $i = 1, 2, \dots, K-1$ , where  $\hat{\beta}_i$  is the MLE based on the  $i$ -th batch of data and  $\hat{\Sigma}_i = V^{-1}/n_i + o(n^{-1})$ . We can define an augmented log-likelihood

$$\ell_A(\beta) = \ell_K(\beta) - \frac{1}{2} \sum_{i=1}^{K-1} (\hat{\beta}_i - \beta)^\top \hat{\Sigma}_i^{-1} (\hat{\beta}_i - \beta)$$

and the corresponding MLE  $\hat{\beta}_A = \arg \max \ell_A(\beta)$ . For  $\beta = \beta_* + O_p(n^{-1/2})$ , using the approximation in (14), we have

$$\begin{aligned} \ell_A(\beta) &= U_K^\top \sqrt{n_K}(\beta - \beta_*) - \frac{n_K}{2}(\beta - \beta_*)^\top V(\beta - \beta_*) \\ &\quad - \frac{1}{2} \sum_{i=1}^{K-1} n_i(\beta - \beta_*)^\top V(\beta - \beta_*) \\ &\quad + \sum_{i=1}^{K-1} n_i(\hat{\beta}_i - \beta_*)^\top V(\beta - \beta_*) + C + o_p(1) \\ &= n^{-1/2} \sum_{i=1}^K \sqrt{n_i} U_i^\top \cdot \sqrt{n}(\beta - \beta_*) \\ &\quad - \frac{n}{2}(\beta - \beta_*)^\top V(\beta - \beta_*) + C + o_p(1), \end{aligned}$$

where the constant  $C$  differs in different equations.

For comparison, based on the pooled data, in Case (a) we define the parametric log-likelihood as

$$\ell_{\text{PL}}(\beta) = \sum_{i=1}^K \sum_{j=1}^{n_i} \log\{f(Y_{ij} | X_{ij}, \beta)\};$$

in Case (b) we define the empirical log-likelihood function as

$$\begin{aligned} \ell_{\text{EL}}(\beta) &= \sup \left\{ \sum_{i=1}^K \sum_{j=1}^{n_i} \log(np_{ij}) : p_{ij} \geq 0, \right. \\ &\quad \times \sum_{i=1}^K \sum_{j=1}^{n_i} p_{ij} = 1, \\ &\quad \times \sum_{i=1}^K \sum_{j=1}^{n_i} p_{ij} g(Y_{ij}, X_{ij}; \beta) = 0 \left. \right\} \\ &= - \sum_{i=1}^K \sum_{j=1}^{n_i} \log\{1 + \lambda^\top g(Y_{ij}, X_{ij}; \beta)\} \\ &\quad - \sum_{i=1}^K n_i \log(n_i), \end{aligned}$$

where  $\lambda$  satisfies  $\sum_{i=1}^K \sum_{j=1}^{n_i} \frac{g(Y_{ij}, X_{ij}; \beta)}{1 + \lambda^\top g(Y_{ij}, X_{ij}; \beta)} = 0$ ; and in Case (c) we define the GMM log-likelihood as

$$\begin{aligned} \ell_{\text{GMM}}(\beta) &= - \left\{ \sum_{i=1}^K \sum_{j=1}^{n_i} g(Y_{ij}, X_{ij}; \beta) \right\}^\top \\ &\quad \times \Omega^{-1} \left\{ \sum_{i=1}^K \sum_{j=1}^{n_i} g(Y_{ij}, X_{ij}; \beta) \right\}. \end{aligned}$$

Let the log-likelihood based on the pooled data be  $\ell_{\text{pool}}(\beta) = \ell_{\text{PL}}(\beta)$ ,  $\ell_{\text{EL}}(\beta)$ , and  $\ell_{\text{GMM}}(\beta)$  in Cases (a), (b), and (c), respectively. Then, it can be shown that

$$\begin{aligned} \ell_{\text{pooled}}(\beta) &= n^{-1/2} \sum_{j=1}^K \sqrt{n_j} U_j^\top \cdot \sqrt{n}(\beta - \beta_*) \\ &\quad - \frac{n}{2} (\beta - \beta_*)^\top V (\beta - \beta_*) + C + o_p(1), \end{aligned}$$

for some constant  $C$ . Let  $\hat{\beta}_{\text{pooled}} = \arg \max \ell_{\text{pooled}}(\beta)$ . By comparing  $\ell_{\text{pooled}}(\beta)$  and  $\ell_A(\beta)$ , we obtain

$$\ell_{\text{pooled}}(\beta) = \ell_A(\beta) + C + o_p(1)$$

and

$$\begin{aligned} \sqrt{n}(\hat{\beta}_A - \beta_*) &= \sqrt{n}(\hat{\beta}_{\text{pooled}} - \beta_*) + o_p(1) \\ &= V^{-1} \cdot n^{-1/2} \sum_{j=1}^K \sqrt{n_j} U_j^\top + o_p(1) \\ &\xrightarrow{d} N(0, V^{-1}). \end{aligned}$$

This indicates that compared with the methods, including parametric likelihood, EL, and GMM, based on all individual data, the calibration method based on the last batch of individual data and all summary results of the previous batches has no efficiency loss.

## 6.2. When nuisance parameters are present

For batch  $i$ , assume that the data  $(Y_{ij}, X_{ij})$  ( $j = 1, 2, \dots, n_i$ ) satisfy either  $\text{pr}(Y_{ij} = y | X_{ij} = x) = f(y | x, \beta, \gamma_i)$  or  $\mathbb{E}\{g(Y, X, \beta, \gamma_i)\} = 0$ , where  $\beta$  is common but  $\gamma_i$  is a batch-specific parameter. We define  $\ell_r(\beta, \gamma_r)$  in the same way as  $\ell_r(\beta)$ . Let  $(\hat{\beta}_i, \hat{\gamma}_i)$  be the MLE of  $(\beta, \gamma_i)$  based on the  $i$ -th batch of data, and assume that approximately

$$((\hat{\beta}_i - \beta)^\top, (\hat{\gamma}_i - \gamma_i)^\top)^\top \sim N(0, \hat{\Sigma}_i)$$

with  $\hat{\Sigma}_i = (\hat{\Sigma}_{i,rs})_{1 \leq r,s \leq 2}$ .

We have two ways of combining information from previous studies. If we use all the previous summary information, we can define

$$\begin{aligned} \ell_A^{(1)}(\beta, \gamma_1, \dots, \gamma_K) &= \ell_K(\beta, \gamma_K) - \frac{1}{2} \sum_{i=1}^{K-1} ((\hat{\beta}_i - \beta)^\top, \\ &\quad (\hat{\gamma}_i - \gamma_i)^\top) \hat{\Sigma}_i^{-1} ((\hat{\beta}_i - \beta)^\top, (\hat{\gamma}_i - \gamma_i)^\top)^\top. \end{aligned} \quad (15)$$

As  $\hat{\beta}_i | \hat{\gamma}_i \sim N(\beta, \hat{\Sigma}_{i,11.2})$ , where  $\hat{\Sigma}_{i,11.2} = \hat{\Sigma}_{i,11} - \hat{\Sigma}_{i,12} \hat{\Sigma}_{i,22}^{-1} \hat{\Sigma}_{i,21}$ , using only this summary information, we can define

$$\begin{aligned} \ell_A^{(2)}(\beta, \gamma_K) &= \ell_K(\beta, \gamma_K) - \frac{1}{2} \sum_{i=1}^{K-1} (\hat{\beta}_i - \beta)^\top \\ &\quad \times \hat{\Sigma}_{i,11.2}^{-1} (\hat{\beta}_i - \beta). \end{aligned}$$

Below we show that the MLEs of  $\beta$  based on these two likelihoods are actually equal to each other. In other words, there is no efficiency loss when estimating  $\beta$  based on  $\ell_A^{(2)}(\beta, \gamma_K)$  instead of  $\ell_A^{(1)}(\beta, \gamma_1, \dots, \gamma_K)$ .

To see this, it suffices to show that

$$\sup_{\gamma_1, \dots, \gamma_{K-1}} \ell_A^{(1)}(\beta, \gamma_1, \dots, \gamma_K) = \ell_A^{(2)}(\beta, \gamma_K). \quad (16)$$

We denote the inverse matrix of  $\Sigma_i$  by  $\Sigma_i^{-1} = (\Sigma_i^{rs})_{1 \leq r,s \leq 2}$ , where

$$\begin{aligned} \Sigma_i^{11} &= \Sigma_{i,11.2}^{-1}, \quad \Sigma_i^{21} = -\Sigma_{i,22}^{-1} \Sigma_{i,21} \Sigma_{i,11.2}^{-1}, \\ \Sigma_i^{12} &= -\Sigma_{i,11.2}^{-1} \Sigma_{i,12} \Sigma_{i,22}^{-1}, \\ \Sigma_i^{22} &= \Sigma_{i,22}^{-1} + \Sigma_{i,22}^{-1} \Sigma_{i,21} \Sigma_{i,11.2}^{-1} \Sigma_{i,12} \Sigma_{i,22}^{-1}. \end{aligned}$$

It can be seen that

$$\begin{aligned} \ell_A^{(1)}(\beta, \gamma_1, \dots, \gamma_K) &= \ell_K(\beta, \gamma_K) - \frac{1}{2} \sum_{i=1}^{K-1} (\hat{\beta}_i - \beta)^\top \Sigma_i^{11} (\hat{\beta}_i - \beta) \end{aligned}$$

$$+ \sum_{i=1}^{K-1} (\hat{\beta}_i - \beta)^\top \Sigma_i^{12} (\gamma_i - \hat{\gamma}_i) \\ - \frac{1}{2} \sum_{i=1}^{K-1} (\gamma_i - \hat{\gamma}_i)^\top \Sigma_i^{22} (\gamma_i - \hat{\gamma}_i).$$

Setting  $\partial \ell_A^{(1)}(\beta, \gamma_1, \dots, \gamma_K) / \partial \gamma_i = 0$  ( $1 \leq i \leq K-1$ ) gives

$$(\gamma_i - \hat{\gamma}_i) = (\Sigma_i^{22})^{-1} \Sigma_i^{21} (\hat{\beta}_i - \beta).$$

Putting this back into  $\ell_A^{(1)}(\beta, \gamma_1, \dots, \gamma_K)$  gives

$$\begin{aligned} & \sup_{\gamma_1, \dots, \gamma_{K-1}} \ell_A^{(1)}(\beta, \gamma_1, \dots, \gamma_K) \\ &= \ell_K(\beta, \gamma_K) - \frac{1}{2} \sum_{i=1}^{K-1} (\hat{\beta}_i - \beta)^\top \\ & \quad \times \{ \Sigma_i^{11} - \Sigma_i^{12} (\Sigma_i^{22})^{-1} \Sigma_i^{21} \} (\hat{\beta}_i - \beta) + C \\ &= \ell_K(\beta, \gamma_K) - \frac{1}{2} \sum_{i=1}^{K-1} (\hat{\beta}_i - \beta)^\top \\ & \quad \times \Sigma_{i,11.2}^{-1} (\hat{\beta}_i - \beta) + C, \end{aligned}$$

where we used the definition of  $\Sigma_{i,11.2}$  in the last equation. We arrive at Equation (16) after comparing this with the definition of  $\ell_A^{(2)}(\beta, \gamma_K)$ .

## 7. Using covariate-specific disease prevalent information

As discussed in the previous section, summarized statistics from previous studies can sometimes be utilized to enhance the estimation efficiency in a current study. This is especially important in the big data era, when many types of information can be found through the internet. More specifically, suppose the prevalence of a disease is known at various levels of a known risk factor  $X$ . In this section, we combine this type of information in a case-control biased sampling setup.

### 7.1. Induced estimating equations under case-control sampling

Case-control sampling is among the most popular methods in cancer epidemiological studies. This is mainly because it is the most convenient, economic, and effective method. In the study of rare diseases in particular, one has to collect large samples in order to get a reasonable number of cases by using prospective sampling, which may not be practical. Using case-control sampling, a pre-specified number of cases ( $n_1$ ) and controls ( $n_0$ ) are collected retrospectively from case and control populations, respectively. Typically, this can be accomplished by sampling cases from hospitals and controls from the general disease-free population.

For a given risk factor  $X$ , let  $F_i(x) = \text{pr}(X \leq x | D = i)$  for  $i = 0, 1$ . Given  $X$  in a range  $(a, b]$ , the disease prevalence is

$$\text{pr}(D = 1 | a < X \leq b) = \phi(a, b),$$

where  $\phi(a, b)$  is known. Using Bayes' formula, we have

$$\begin{aligned} \phi(a, b) &= \frac{\pi \int_a^b dF_1(x)}{\text{pr}(a < X \leq b)}, \\ 1 - \phi(a, b) &= \frac{(1 - \pi) \int_a^b dF_0(x)}{\text{pr}(a < X \leq b)} \end{aligned}$$

with  $\pi = \text{pr}(D = 1)$ . It follows that

$$\int_a^b dF_1(x) = \frac{1 - \pi}{\pi} \frac{\phi(a, b)}{1 - \phi(a, b)} \int_a^b dF_0(x),$$

or

$$\begin{aligned} \mathbb{E}_1[I(a < X \leq b)] \\ = \frac{1 - \pi}{\pi} \frac{\phi(a, b)}{1 - \phi(a, b)} \mathbb{E}_0[I(a < X \leq b)], \end{aligned}$$

where  $\mathbb{E}_0$  and  $\mathbb{E}_1$  denote the expectation operators with respect to  $F_0$  and  $F_1$ , respectively.

We assume that given covariates  $X$  and  $Y$ , the underlying disease model is given by the conventional logistic regression

$$\text{pr}(D = 1 | x, y) = \frac{\exp(\alpha^* + x\beta + y\gamma + yx\xi)}{1 + \exp(\alpha^* + x\beta + y\gamma + yx\xi)}. \quad (17)$$

Let  $\alpha = \alpha^* - \eta$  with  $\eta = \log\{(1 - \pi)/\pi\}$ . It can be shown (see Qin, 2017) that this is equivalent to the exponential tilting model

$$\begin{aligned} f_1(x, y) &= f(x, y | D = 1) \\ &= \exp(\alpha + x\beta + y\gamma + yx\xi) f_0(x, y), \end{aligned}$$

where  $f_0(x, y) = f(x, y | D = 0)$ . As a consequence,

$$\begin{aligned} \mathbb{E}_0\{I(a < X \leq b) \exp(\eta + \alpha + \beta X + \gamma Y + \xi XY)\} \\ = \frac{1 - \pi}{\pi} \frac{\phi(a, b)}{1 - \phi(a, b)} \mathbb{E}_0[I(a < X \leq b)] \end{aligned}$$

or

$$\begin{aligned} \mathbb{E}_0\{I(a < X \leq b) \exp(\alpha + \beta X + \gamma Y + \xi XY) \\ - \frac{\phi(a, b)}{1 - \phi(a, b)} I(a < X \leq b)\} = 0. \end{aligned} \quad (18)$$

We denote

$$g_0(X, Y) = \exp(\eta + \alpha + \beta X + \gamma Y + \xi X Y) - 1$$

and the summarized auxiliary information equations as

$$g_i(X, Y) = I(a_{i-1} < X \leq a_i) \exp(\alpha + \beta X + \gamma Y$$

$$+ \xi XY) - \frac{\phi(a_{i-1}, a_i)}{1 - \phi(a_{i-1}, a_i)} \\ \times I(a_{i-1} < X \leq a_i)$$

with  $i = 1, 2, \dots, I$ . Then  $\mathbb{E}_0\{g(X, Y)\} = 0$ , where  $g(X, Y) = (g_0(X, Y), g_1(X, Y), \dots, g_I(X, Y))^\top$ .

## 7.2. Empirical likelihood approach

The log-likelihood is

$$\ell = \sum_{i=1}^n d_i(\eta + \alpha + \beta x_i + \gamma y_i + \xi x_i y_i) + \sum_{i=1}^n \log(p_i), \quad (19)$$

where  $p_i = dF_0(x_i)$ ,  $i = 1, 2, \dots, n$ , and the constraints are

$$p_i \geq 0, \quad \sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i g(x_i, y_i) = 0.$$

The profile log-likelihood is

$$\ell = \sum_{i=1}^n d_i(\eta + \alpha + \beta x_i + \gamma y_i + \xi x_i y_i) \\ - \sum_{i=1}^n \log\{1 + \lambda^\top g(x_i, y_i)\},$$

where the Lagrange multiplier  $\lambda$  is determined by

$$\sum_{i=1}^n \frac{g(x_i, y_i)}{1 + \lambda^\top g(x_i, y_i)} = 0.$$

Finally, the underlying parameters can be obtained by maximizing  $\ell$ .

If the overall disease prevalence probability  $\pi = \text{pr}(D = 1)$  is known, then  $\eta = \log\{(1 - \pi)/\pi\}$  is known. On the other hand, if it is unknown but  $I \geq 1$ , then  $\pi$  is identifiable. If  $I > 1$ , then we have an over-identified equation problem. This can be treated as a generalization of the EL method for estimating functions (Qin & Lawless, 1994) for biased sampling problems. Qin et al. (2015) considered the case where  $\eta$  is unknown and  $I \geq 1$ .

Let  $\omega = (\eta, \alpha, \beta, \gamma, \xi, \lambda)$  and let  $\hat{\omega}$  be its maximum EL estimator. As the first estimating function  $g_0$  corrects biased sampling in a case-control study, the remaining estimating functions  $g_1, \dots, g_I$  are used for improving efficiency. When  $n$  goes to infinity, it can be shown that the limit of  $\lambda$  is a  $(I + 1)$ -dimensional vector where the first component is  $\lim_{n \rightarrow \infty} (n_1/n)$  and the remainder are all zero. Qin et al. (2015) showed that if  $\rho = n_1/n_0$  remains constant as  $n \rightarrow \infty$  and  $\rho \in (0, 1)$ , then under suitable regularity conditions  $\sqrt{n}(\hat{\omega} - \omega_0)$  is asymptotically normally distributed with mean zero. Moreover, the estimation of the logistic regression parameters  $(\beta, \gamma, \xi)$  improves as the number  $I$  of estimating

functions increases. This means that a richer set of auxiliary information leads to better estimators. In practice, however, this consideration must be balanced with the numerical difficulty of solving a larger number of equations.

Notably, auxiliary information is informative for estimating  $\beta$  and  $\xi$  but not for estimating  $\gamma$ . This can be observed through the following equations:

$$\int I(a < x < b) \exp(\alpha + \beta x + \gamma y + \xi xy) dF_0(x, y) \\ = \int I(a < x < b) \exp\{\alpha + \beta x + s + \xi x(s/\gamma)\} \\ \times dF_0(x, s/\gamma).$$

As the underlying distribution  $F_0(x, y)$  is unspecified, we can treat  $F_0(x, s/\gamma)$  as a new underlying distribution  $F_0^*(x, s)$ . With  $F_0^*$  profiled out, the auxiliary information equation does not involve  $\gamma$  if  $\xi = 0$ . Hence, even if  $\xi \neq 0$ , the information for  $\gamma$  is minimal as  $\gamma$  and  $\xi$  cannot be separated.

## 7.3. Generalizations

The simulation results of Qin et al. (2015) indicate that when covariate-specific auxiliary information is employed, the estimator of the coefficient  $\beta$  of  $X$  has the maximum variance reduction, whereas the variance reductions for other coefficients are small. If the auxiliary information

$$\text{pr}(D = 1 | b_{j-1} < Y \leq b_j) = \psi_j, \quad j = 1, 2, \dots, J$$

is also available, we can combine them through estimating equations

$$g_i(X, Y) = I(a_{i-1} < X \leq a_i) e^{\alpha + \beta X + \gamma Y + \xi XY} \\ - \frac{\phi(a_{i-1}, a_i)}{1 - \phi(a_{i-1}, a_i)} I(a_{i-1} < X \leq a_i), \\ h_j(X, Y) = I(b_{j-1} < Y \leq b_j) e^{\alpha + \beta X + \gamma Y + \xi XY} \\ - \frac{\psi(b_{j-1}, b_j)}{1 - \psi(b_{j-1}, b_j)} I(b_{j-1} < Y \leq b_j).$$

It would be more informative if the auxiliary information  $\text{pr}(D = 1 | a < X < b, c < Y < d)$  is available.

## 7.4. More on the use of auxiliary information

Under a logistic regression model, the case and control densities are linked by the exponential tilting model

$$\text{pr}(x, y | D = 1) = \text{pr}(x, y | D = 0) \\ \times \exp(\alpha + x\beta + y\gamma + \xi xy). \quad (20)$$

Suppose that for the general population  $\mathbb{E}(X) = \mu_1$ ,  $\mathbb{E}(Y) = \mu_2$ , and  $\mathbb{E}(XY) = \mu_3$  are all known, and  $\pi =$



$\text{pr}(D = 1)$  is known or can be estimated using external data. Under the exponential tilting model (20), the density  $f(x, y)$  in the general population and the density  $\text{pr}(x, y | D = 0)$  in the control population are linked by

$$\begin{aligned} \text{pr}(x, y) &= \{\pi e^{\alpha + x\beta + y\gamma + \xi xy} + (1 - \pi)\} \\ &\times \text{pr}(x, y | D = 0). \end{aligned}$$

As a consequence

$$\mathbb{E}(X) = \mathbb{E}_0[X\{\pi e^{\alpha + X\beta + Y\gamma + \xi XY} + (1 - \pi)\}] = \mu_1,$$

where  $\mathbb{E}_0$  is an expectation with respect to  $\text{pr}(x, y | D = 0)$ . Let  $h(x, y) = (x - \mu_1, y - \mu_2, xy - \mu_3)$  with known  $\mu_1, \mu_2$ , and  $\mu_3$ . The log-likelihood under case-control data is still (19), where the  $p_i$ s satisfy the following constraints:

$$\begin{aligned} \sum_{i=1}^n p_i &= 1, \quad p_i \geq 0, \quad \sum_{i=1}^n p_i e^{\alpha + x_i\beta + y_i\gamma + x_i y_i \xi} = 1, \\ \sum_{i=1}^n p_i h(x_i, y_i) \{\pi e^{\alpha + x_i\beta + y_i\gamma + x_i y_i \xi} + (1 - \pi)\} &= 0. \end{aligned}$$

More generally, any information in the general population such as  $\mathbb{E}[\psi(Y, X)] = 0$  can be converted to an equation for the control population,

$$\mathbb{E}_0[\{\pi e^{\alpha + X\beta + Y\gamma + \xi XY} + (1 - \pi)\}\psi(Y, X)] = 0.$$

Therefore, the results developed by Qin et al. (2015) can be applied. The results of Chatterjee et al. (2016) for case-control data can be considered as a special case of Qin et al. (2015).

## 8. Communication-efficient distributed inference

In the era of big data, it is commonplace for data analyses to run on hundreds or thousands of machines, with the data distributed across those machines and no longer available in a single central location. Recently, parallel and distributed inference has become popular in the statistical literature in both frequentist and Bayesian settings. In essence, the data-parallel procedures are intended to break the overall dataset into subsets that are processed independently. To the extent that communication-avoiding procedures have been discussed explicitly, the focus has been on one-shot or embarrassingly parallel approaches that use only one round of communication in which estimators or posterior samples are first obtained in parallel on local machines, then communicated to a centre node, and finally combined to form a global estimator or approximation to the posterior distribution (Lee et al., 2017; Neiswanger et al., 2015; Wang & Dunson, 2015; Zhang et al., 2013). In the frequentist setting, most one-shot approaches rely on averaging (Zhang

et al., 2013), where the global estimator is the average of the local estimators. Lee et al. (2017) extend this idea to high-dimensional sparse linear regression by combining local debiased Lasso estimates (van de Geer et al., 2014). Recent work by Duchi et al. (2015) shows that under certain conditions, these averaging estimators can attain the information-theoretic complexity lower bound for linear regression, and at least  $O(dk)$  bits must be communicated in order to attain the minimax rate of parameter estimation, where  $d$  is the dimension of the parameter and  $k$  is the number of machines. This result holds even in the sparse setting (Braverman et al., 2016).

The method of Jordan et al. (2019) proceeds as follows. Suppose the big data consists of  $N$  observations and there are  $k$  machines. For the convenience of presentation, we assume that each machine has  $n$  observations, i.e.,  $N = nk$ . Denote the full-data likelihood by

$$\ell_N(\theta) = \frac{1}{k} \sum_{j=1}^k \ell_j(\theta),$$

where  $\ell_j(\theta)$  is the log-likelihood based on the data from the  $j$ -th machine. For  $\theta$  near its target value  $\bar{\theta}$ ,

$$\begin{aligned} \ell_N(\theta) &= \ell_N(\bar{\theta}) + \nabla_{\theta} \ell_N(\theta) \Big|_{\theta=\bar{\theta}} (\theta - \bar{\theta}) + R_N(\theta), \\ \ell_1(\theta) &= \ell_1(\bar{\theta}) + \nabla_{\theta} \ell_1(\theta) \Big|_{\theta=\bar{\theta}} (\theta - \bar{\theta}) + R_1(\theta), \end{aligned}$$

where  $R_N(\theta)$  and  $R_1(\theta)$  are remainders. Observing that  $R_N \approx R_1$ , define a surrogate log-likelihood

$$\begin{aligned} \bar{\ell}(\theta) &= \ell_N(\bar{\theta}) + (\theta - \bar{\theta})^\top \nabla_{\theta} \ell_N(\theta) \Big|_{\theta=\bar{\theta}} \\ &\quad + \left\{ \ell_1(\theta) - \ell_1(\bar{\theta}) - (\theta - \bar{\theta})^\top \nabla_{\theta} \ell_1(\theta) \Big|_{\theta=\bar{\theta}} \right\}. \end{aligned}$$

Ignoring the constant terms, the surrogate log-likelihood is

$$\bar{\ell}(\theta) = \ell_1(\theta) + \theta^\top \left\{ \nabla_{\theta} \ell_N(\theta) \Big|_{\theta=\bar{\theta}} - \nabla_{\theta} \ell_1(\theta) \Big|_{\theta=\bar{\theta}} \right\}.$$

The score equation based on the surrogate likelihood is

$$\begin{aligned} \nabla_{\theta} \bar{\ell}(\theta) &= \nabla_{\theta} \ell_1(\theta) + \left\{ \nabla_{\theta} \ell_N(\theta) \Big|_{\theta=\bar{\theta}} - \nabla_{\theta} \ell_1(\theta) \Big|_{\theta=\bar{\theta}} \right\} \\ &= 0. \end{aligned}$$

Let  $\hat{\theta}$  be the solution. Expanding it at  $\theta_0$  and using the fact that

$$\begin{aligned} N^{-1} \{ \nabla_{\theta} \ell_1(\theta_0) - \nabla_{\theta} \ell_N(\theta_0) \} \\ \rightarrow 0 \quad \text{in probability,} \end{aligned}$$

we can easily show that if  $\bar{\theta} - \theta_0 = O_p(N^{-1/2})$  then

$$(\hat{\theta} - \theta_0) = \{\nabla_{\theta\theta^\top} \ell_N(\theta_0)\}^{-1} \nabla_{\theta} \ell_N(\theta_0) + o_p(N^{-1/2}).$$

If we let  $\bar{\theta}$  be the MLE based on  $\ell_1(\theta)$ , the surrogate log-likelihood can be simplified to

$$\bar{\ell}(\theta) = \ell_1(\theta) + \theta^\top \nabla_{\theta} \ell_N(\bar{\theta}),$$

because  $\nabla_{\theta} \ell_1(\bar{\theta}) = 0$ .

If the dimension of  $\theta$  is high, one may add a penalty function in the surrogate log-likelihood and estimate  $\theta$  by  $\tilde{\theta} = \arg \max_{\theta \in \Theta} \{\bar{\ell}(\theta) - \lambda \|\theta\|_1\}$ , where  $\|\theta\|_1$  is the  $L_1$ -norm of  $\theta$ . Similarly, Bayesian inference can be adapted to the surrogate likelihood as well.

Duan et al. (2020) proposed distributed algorithms that account for heterogeneous distributions by allowing site-specific nuisance parameters. The proposed methods extend the surrogate likelihood approach (Jordan et al., 2019; Wang et al., 2017) to the heterogeneous setting by applying a novel density ratio tilting method to the efficient score function. Asymptotically, the approach described in Section 6.2 on nuisance parameters is equivalent to that of Duan et al. (2020).

## 9. Renewal estimation and incremental inference

Let  $U(D_1, \beta) = \nabla_{\beta} M(D_1, \beta)$  be a score function of  $\beta$  based on some objective function  $M(D_1, \beta)$  from the first batch of data, where  $M$  can be either the log-likelihood  $M(D_1, \beta) = \sum_{i=1}^{n_1} \log f(y_{1i} | x_{1i}, \beta)$  or a pseudo log-likelihood.

Let  $\hat{\beta}_1$  be the solution to  $U(D_1, \beta) = 0$ , when only the first batch of data  $D_1$  is available. Let  $D_2$  denote the second batch of data. If both of them are available, we let  $\hat{\beta}_2$  be the solution to the pooled score equation,  $U(D_1, \beta) + U(D_2, \beta) = 0$ . Clearly,  $\hat{\beta}_2$  is the most efficient estimator of  $\beta$  when  $D_1$  and  $D_2$  are both available.

If  $D_2$  is available but  $D_1$  is not, with only some summary information  $\hat{\beta}_1$  and  $\hat{\Sigma}_1$  in its place, how can we utilize the summary information efficiently? It is not feasible to estimate  $\beta$  by directly solving

$$U(\beta) \equiv U(D_1, \beta) + U(D_2, \beta) = 0,$$

which involves the individual data of the unavailable  $D_1$ . Luo (2020) considers expanding  $U(D_1, \beta)$  at  $\beta = \hat{\beta}_1$ , i.e.,

$$U(D_1, \beta) = U(D_1, \hat{\beta}_1) + (\beta - \hat{\beta}_1)^\top \nabla_{\beta} U(D_1, \hat{\beta}_1) + O(\|\beta - \hat{\beta}_1\|^2)$$

for  $\beta$  close to  $\hat{\beta}_1$ . As  $U(D_1, \hat{\beta}_1) = 0$ , it follows that

$$U(\beta) = U(D_2, \beta) + (\beta - \hat{\beta}_1)^\top \nabla_{\beta} U(D_1, \hat{\beta}_1) + O(\|\beta - \hat{\beta}_1\|^2).$$

Luo (2020) proposes obtaining an updated estimator of  $\beta$  by solving

$$(\beta - \hat{\beta}_1)^\top \nabla_{\beta} U(D_1, \hat{\beta}_1) + U(D_2, \beta) = 0. \quad (21)$$

Alternatively, we may understand this renewal estimation strategy in the manner of Zhang et al. (2020), who propose estimating  $\beta$  by maximizing

$$\sum_{i=1}^{n_2} \log f(y_{2i} | x_{2i}, \beta) - \frac{1}{2} n_1 (\hat{\beta}_1 - \beta)^\top \Sigma (\hat{\beta}_1 - \beta), \quad (22)$$

where  $\Sigma = \mathbb{E} \{\nabla_{\beta} \log f(Y | X, \beta) \nabla_{\beta}^\top \log f(Y | X, \beta)\}$  is the Fisher information. If both batches are available, the score for  $\beta$  is

$$S(\beta) = \sum_{i=1}^{n_1} \nabla_{\beta} \log f(y_{1i} | x_{1i}, \beta) + \sum_{i=1}^{n_2} \nabla_{\beta} \log f(y_{2i} | x_{2i}, \beta).$$

After recording  $\hat{\beta}_1$  and  $\Sigma$ , we no longer have the raw data  $\{(y_{1i}, x_{1i}), i = 1, 2, \dots, n_1\}$ . As

$$\hat{\beta}_1 - \beta = -n_1^{-1} \Sigma^{-1} \sum_{i=1}^{n_1} \nabla_{\beta} \log f(y_{1i} | x_{1i}, \beta) + o_p(n_1^{-1/2}),$$

differentiating (22) with respect to  $\beta$  gives

$$\begin{aligned} & \sum_{i=1}^{n_2} \nabla_{\beta} \log f(y_{2i} | x_{2i}, \beta) - n_1 \Sigma (\hat{\beta}_1 - \beta) \\ &= \sum_{i=1}^{n_1} \nabla_{\beta} \log f(y_{1i} | x_{1i}, \beta) \\ &+ \sum_{i=1}^{n_2} \nabla_{\beta} \log f(y_{2i} | x_{2i}, \beta) + o_p(n^{1/2}). \end{aligned}$$

Here, we have assumed that  $n_1 = O(n_2) = O(n)$ . This indicates that estimating  $\beta$  by maximizing (22) results in no efficiency loss asymptotically compared with the MLE based on all individual data, where the latter is infeasible in the current situation.

## 10. Concluding remarks

Rapid growth in hardware technology has made data collection much easier and more effective. In many applications, data often arrive in streams and chunks, which leads to batch-by-batch data or streaming data. For example, web sites served by widely distributed web servers may need to coordinate many distributed clickstream analyses, e.g., to track heavily accessed web

pages as part of their real-time performance monitoring. Other examples include financial applications, network monitoring, security, telecommunications data management, manufacturing, and sensor networks (Babcock et al., 2002; Nguyen et al., 2021). The continuous arrival of such data in multiple, rapid, time-varying, possibly unpredictable and unbounded streams not only yields many fundamentally new research problems but provides contains various forms of auxiliary information.

Assembling information from different data sources has become indispensable in big data and artificial intelligence research. Statistical tools play an essential part in updating information. In this paper, we have presented a selective review of several traditional statistical methods, including meta-analysis, calibration information methods in survey sampling, and EL together with over-identified estimating equations and GMM. We have also briefly reviewed some recently developed statistical methods, including communication-efficient distributed statistical inference and renewal estimation and incremental inference, which can be regarded as the latest developments of calibration information methods in the era of big data. Although these methods were developed in different fields and in different statistical frameworks, in principle, they are asymptotically equivalent to well-known methods developed for meta-analysis. These methods result in almost no or little information loss compared with the case when full data are available.

Finally, we apologize to people whose work has inadvertently have been left out of our reference list.

## Acknowledgments

The authors thank the editor and two referees for constructive comments and suggestions that led to significant improvements in this paper.

## Funding

This research was supported by the National Natural Science Foundation of China [grant numbers 71931004, 12171157, and 32030063], the 111 Project [grant number B14019], the Development Fund for Shanghai Talents and the Natural Sciences and Engineering Research Council of Canada (grant number RGPIN-2020-04964).

## References

- Babcock, B., Babu, S., Datar, M., Motwani, R., & Widom, J. (2002). Models and issues in data stream systems. In *Proceedings of the 21 ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems* (pp. 1–16). ACM.
- Back, K., & Brown, D. P. (1992). GMM, maximum likelihood, and nonparametric efficiency. *Economics Letters*, 39(1), 23–28. [https://doi.org/10.1016/0165-1765\(92\)90095-G](https://doi.org/10.1016/0165-1765(92)90095-G)
- Braverman, M., Garg, A., Ma, T., Nguyen, H., & Woodruff, D. (2016). Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the 48th annual ACM symposium on theory of computing* (pp. 1011–1020). ACM.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2009). *Introduction to meta-analysis*. Wiley.
- Chatterjee, N., Chen, Y.-H., Maas, P., & Carroll, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111(513), 107–117. <https://doi.org/10.1080/01621459.2015.1123157>
- Chaudhuri, S., Handcock, M. S., & Rendall, M. S. (2008). Generalized linear models incorporating population level information: an empirical likelihood based approach. *Journal of the Royal Statistical Society: Series B*, 70(2), 311–328. <https://doi.org/10.1111/rssb.2008.70.issue-2>
- Chen, J., & Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80(1), 107–116. <https://doi.org/10.1093/biomet/80.1.107>
- Chen, J., Sitter, R., & Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89(1), 230–237. <https://doi.org/10.1093/biomet/89.1.230>
- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). Wiley.
- Dersimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3), 177–188. [https://doi.org/10.1016/0197-2456\(86\)90046-2](https://doi.org/10.1016/0197-2456(86)90046-2)
- Duan, R., Ning, Y., & Chen, Y. (2020). *Heterogeneity-aware and communication-efficient distributed statistical inference*. arXiv:1912.09623v1.
- Duchi, J., Jordan, M., Wainwright, M., & Zhang, Y. (2015). *Optimality guarantees for distributed statistical estimation*. arXiv:1405.0782.
- Han, P., & Lawless, J. (2016). Comment. *Journal of the American Statistical Association*, 111(513), 118–121. <https://doi.org/10.1080/01621459.2016.1149399>
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4), 1029–1054. <https://doi.org/10.2307/1912775>
- Hartely, H. O., & Rao, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55(3), 547–557. <https://doi.org/10.1093/biomet/55.3.547>
- Imbens, G., & Lancaster, T. (1994). Combining micro and macro data in microeconomic models. *Review of Economic Studies*, 61(4), 655–680. <https://doi.org/10.2307/2297913>
- Jordan, M. I., Lee, J. D., & Yang, Y. (2019). Communication-efficient distribution statistical inference. *Journal of the American Statistical Association*, 114(526), 668–681. <https://doi.org/10.1080/01621459.2018.1429274>
- Lee, J., Liu, Q., Sun, Y., & Taylor, J. (2017). Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18, 1–30. <http://jmlr.org/papers/v18/16-002.html>
- Lin, D. Y., & Zeng, D. (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, 97(2), 321–332. <https://doi.org/10.1093/biomet/asq006>
- Luo, L. (2020). Renewable estimation and incremental inference in generalized linear models with streaming data sets. *Journal of the Royal Statistical Society, Series B*, 82(1), 69–97. <https://doi.org/10.1111/rssb.12352>
- Neiswanger, W., Wang, C., & Xing, E. (2015). Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the 30th conference on uncertainty in artificial intelligence* (pp. 623–632). AUAI Press.

- Nguyen, T. D., Shih, M. H., Srivastava, D., Tirthapura, S., & Xu, B. (2021). Stratified random sampling from streaming and stored data. *Distributed and Parallel Databases*, 39(3), 665–710. <https://doi.org/10.1007/s10619-020-07315-w>
- Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2), 237–249. <https://doi.org/10.1093/biomet/75.2.237>
- Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *Annals of Statistics*, 18(1), 90–120. <https://doi.org/10.1214/aos/1176347494>
- Owen, A. B. (2001). *Empirical likelihood*. CRC.
- Qin, J. (2000). Combining parametric and empirical likelihoods. *Biometrika*, 87(2), 484–490. <https://doi.org/10.1093/biomet/87.2.484>
- Qin, J. (2017). *Biased sampling, over-identified parameter problems and beyond*. Springer.
- Qin, J., & Lawless, J. (1994). Empirical likelihood and general equations. *Annals of Statistics*, 22(1), 300–325. <https://doi.org/10.1214/aos/1176325370>
- Qin, J., Zhang, H., Li, P., Albanes, D., & Yu, K. (2015). Using covariate specific disease prevalence information to increase the power of case-control study. *Biometrika*, 102(1), 169–180. <https://doi.org/10.1093/biomet/asu048>
- Susanne, M. S. (2007). Point estimation with exponentially tilted empirical likelihood. *Annals of Statistics*, 35(2), 634–672. <https://doi.org/10.1214/0090536060000001208>
- Tian, L., & Gu, Q. (2016). *Communication-efficient distributed sparse linear discriminant analysis*. arXiv:1610.04798.
- van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high dimensional models. *Annals of Statistics*, 42(3), 1166–1202. <https://doi.org/10.1214/14-AOS1221>
- van de Vaart, V. W. (2000). *Asymptotic statistics*. Cambridge University Press.
- Wang, X., & Dunson, D. (2015). *Parallelizing MCMC via Weierstrass sampler*. arXiv:1312.4605.
- Wang, J., Kolar, M., Srebro, N., & Zhang, T. (2017). Efficient distributed learning with sparsity. In *Proceedings of the 34th international conference on machine learning*, Sydney, Australia, PMLR 70 (pp. 3636–3645).
- Wu, C., & Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453), 185–193. <https://doi.org/10.1198/016214501750333054>
- Wu, C., & Thompson, M. E. (2020). *Sampling theory and practice*. Springer.
- Zeng, D. & Lin, D. Y. (2015). On random-effects meta-analysis. *Biometrika*, 102(2), 281–294.
- Zhang, Y., Duchi, J., & Wainwright, M. (2013). Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14, 3321–3363.
- Zhang, H., Deng, L., Schiffman, M., Qin, J., & Yu, K. (2020). Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika*, 107(3), 689–703. <https://doi.org/10.1093/biomet/asaa014>