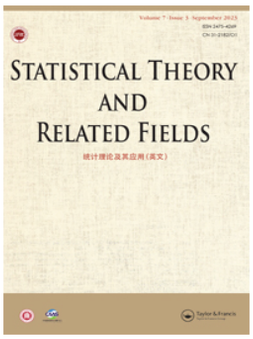

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

MLE with datasets from populations having shared parameters

Jun Shao & Xinyan Wang

To cite this article: Jun Shao & Xinyan Wang (2023) MLE with datasets from populations having shared parameters, Statistical Theory and Related Fields, 7:3, 213-222, DOI: 10.1080/24754269.2023.2180185

To link to this article: https://doi.org/10.1080/24754269.2023.2180185

Published online: 04 Mar 2023.

STATISTICAL THEORY AND RELATED FIELDS
2023, VOL. 7, NO. 3, 213–222
https://doi.org/10.1080/24754269.2023.2180185

Taylor & Francis
Taylor & Francis Group

MLE with datasets from populations having shared parameters

Jun Shao[a] and Xinyan Wang[b]

[a]School of Statistics, East China Normal University, Shanghai, People's Republic of China; [b]Department of Statistics, University of Wisconsin, Madison, WI, USA

ABSTRACT
We consider maximum likelihood estimation with two or more datasets sampled from different populations with shared parameters. Although more datasets with shared parameters can increase statistical accuracy, this paper shows how to handle heterogeneity among different populations for correctness of estimation and inference. Asymptotic distributions of maximum likelihood estimators are derived under either regular cases where regularity conditions are satisfied or some non-regular situations. A bootstrap variance estimator for assessing performance of estimators and/or making large sample inference is also introduced and evaluated in a simulation study.

ARTICLE HISTORY
Received 4 August 2022
Revised 27 January 2023
Accepted 1 February 2023

KEYWORDS
Accuracy; asymptotic relative efficiency; bootstrap; population heterogeneity; regularity conditions

1. Introduction

With advanced technologies in data collection and storage, in modern statistical analyses we often have multiple datasets as independent samples from different populations having shared parameters. Typically, one of these multiple datasets is primary with carefully collected data from a population of interest. The other datasets are from external sources, such as data from other studies, administrative records and publicly available information from internet.

On one hand, the fact that populations share common parameters provides a great opportunity for increasing statistical accuracy by utilizing multiple datasets instead of a single dataset. On the other hand, because of the difference in data collection, study purpose and/or time of investigation, heterogeneity often exists among populations so that we cannot simply combine all datasets into a single large dataset to run analysis, but must develop or modify statistical methodology to correctly utilize multiple datasets. The research on analysis with multiple datasets fits into a general framework of data integration (Kim et al., 2021; Lohr & Raghunathan, 2017; Merkouris, 2004; Rao, 2021; Yang & Kim, 2020; Zhang et al., 2017; Zieschang, 1990).

In this article, we study maximum likelihood estimation (MLE) for independent datasets with parametric populations sharing some (not necessarily all) parameters. For simplicity of presentation, we focus on the case of two independent datasets. The main idea and result can be extended to multiple datasets. Our research can also be extended to semi-parametric estimation, such as empirical likelihood or Cox regression for survival data.

Throughout, we consider two independent random samples. One random sample of size n, resulting a dataset $\{X_1, \ldots, X_n\}$, is sampled from a parametric population with probability density $f(x, \theta, \phi)$ (for either continuous or discrete x), where f is a known function and θ and ϕ are unknown parameter vectors. Another random sample of size m, resulting a dataset $\{Y_1, \ldots, Y_m\}$, is sampled from a population with probability density $g(y, \theta, \varphi)$, where g is a known function and θ and φ are unknown parameter vectors. Note that X_i and Y_j can be vectors. The shared parameter θ can be either the main parameter vector of interest or a nuisance parameter vector, and ϕ and φ are other parameter vectors in two populations.

Let ϑ denote the vector with θ, ϕ, and φ as sub-vectors. In Section 2, we derive the maximum likelihood estimator (MLE) of ϑ based on two datasets, which is expected to be asymptotically more efficient than each MLE based on a single dataset, since more data are used for estimating the shared parameter θ, a component of ϑ. The asymptotic normality of MLE of ϑ is established when densities f and g satisfy regularity conditions that are typically assumed for MLE. Applications to location-scale problems are discussed in Section 3, where we also present a situation in which f or g does not satisfy the regularity conditions. Section 4 contains an example in which regularity conditions do not hold and MLE is not asymptotically normal. The common mean of a discrete data problem is considered in

CONTACT Xinyan Wang xwang2587@wisc.edu Department of Statistics, University of Wisconsin, Madison, WI 53706, USA

Section 5. Section 6 is devoted to the scenario where an additional uncertainty exists in the second population density g. To handle the situation where asymptotic normality of the MLE of ϑ is not available, we introduce a bootstrap variance estimator in Section 7 and provide some simulation results to examine finite sample performances.

2. MLEs with two datasets

The following are regularity conditions for probability density $p(x,\vartheta)$ (with a fixed ϑ) of a continuous or discrete random variable/vector X, typically assumed for MLEs in parametric populations (Shao, 2003).

(R1) For every x in the range of X, $p(x,\vartheta)$ is twice continuously differentiable with respect to ϑ in an open set of the Euclidean space with a fixed dimension.

(R2) $\frac{\partial}{\partial\vartheta}\int p(x,\vartheta)\mathrm{d}x = \int \frac{\partial}{\partial\vartheta}p(x,\vartheta)\mathrm{d}x$ and $\frac{\partial}{\partial\vartheta}\int \frac{\partial}{\partial\vartheta^\top}p(x,\vartheta)\mathrm{d}x = \int \frac{\partial^2}{\partial\vartheta\,\partial\vartheta^\top}p(x,\vartheta)\mathrm{d}x$, where $C^\top$ denotes the transpose of a vector or matrix C and the integral should be replaced by an appropriate summation when X is discrete.

(R3) The Fisher information matrix $-E\{\frac{\partial^2}{\partial\vartheta\,\partial\vartheta^\top}\log p(X,\vartheta)\}$ exists and is positive definite,

(R4) For any given ϑ, there exists a positive number c_ϑ and a positive function h_ϑ such that $E\{h_\vartheta(X)\} < \infty$ and $\sup_{\gamma:\|\gamma-\vartheta\|<c_\vartheta}\|\frac{\partial^2\log p(x,\gamma)}{\partial\gamma\,\partial\gamma^\top}\| \le h_\vartheta(x)$ for all x in the range of X, where $\|A\| = \sqrt{\mathrm{trace}(A^\top A)}$ for any matrix A.

In this section, we assume that both f and g satisfy regularity conditions (R1) –(R4). When some regularity conditions are not satisfied, we have to deal with the problem case by case. See, for example, the problem of normal and Laplace distributions in Section 3.2 and the problem of two truncation distributions in Section 4.

The log likelihood function of ϑ is

$$\ell(\vartheta) = \sum_{i=1}^n \log f(X_i,\theta,\phi) + \sum_{j=1}^m \log g(Y_j,\theta,\varphi)$$

and the score function is

$$s(\vartheta) = \frac{\partial\ell(\vartheta)}{\partial\vartheta} = \begin{pmatrix} \sum_{i=1}^n \frac{\partial\log f(X_i,\theta,\phi)}{\partial\theta} + \sum_{j=1}^m \frac{\partial\log g(Y_j,\theta,\varphi)}{\partial\theta} \\ \sum_{i=1}^n \frac{\partial\log f(X_i,\theta,\phi)}{\partial\phi} \\ \sum_{j=1}^m \frac{\partial\log g(Y_j,\theta,\varphi)}{\partial\varphi} \end{pmatrix}.$$

If $\widehat{\vartheta}$ is a solution to the score equation $s(\vartheta)=0$, then we call $\widehat{\vartheta}$ an MLE of ϑ, although traditionally an MLE is defined as a maximizer of $\ell(\vartheta)$ over the range of ϑ and $\widehat{\vartheta}$ satisfying $s(\widehat{\vartheta})=0$ may not be a maximizer.

A solution to the score equation often does not have an explicit form, even when each MLE of a single population has an explicit solution.

Under regularity conditions (R1)-(R4), $E\{s(\vartheta)\}=0$ and

$$\mathrm{Var}\{s(\vartheta)\} = -E\left\{\frac{\partial s(\vartheta)}{\partial\vartheta^\top}\right\} = n\mathscr{I}(\vartheta)$$

is the Fisher information matrix of information contained in two samples. Let

$$\mathscr{I}_{\theta\theta}(\theta,\phi) = -E\left\{\frac{\partial^2\log f(X_i,\theta,\phi)}{\partial\theta\,\partial\theta^\top}\right\}, \quad \mathscr{I}_{\theta\theta}(\theta,\varphi) = -E\left\{\frac{\partial^2\log g(Y_j,\theta,\varphi)}{\partial\theta\,\partial\theta^\top}\right\},$$

$$\mathscr{I}_{\theta\phi}(\theta,\phi) = -E\left\{\frac{\partial^2\log f(X_i,\theta,\phi)}{\partial\theta\,\partial\phi^\top}\right\}, \quad \mathscr{I}_{\theta\varphi}(\theta,\varphi) = -E\left\{\frac{\partial^2\log g(Y_j,\theta,\varphi)}{\partial\theta\,\partial\varphi^\top}\right\},$$

$$\mathscr{I}_{\phi\phi}(\theta,\phi) = -E\left\{\frac{\partial^2\log f(X_i,\theta,\phi)}{\partial\phi\,\partial\phi^\top}\right\}, \quad \mathscr{I}_{\varphi\varphi}(\theta,\varphi) = -E\left\{\frac{\partial^2\log g(Y_j,\theta,\varphi)}{\partial\varphi\,\partial\varphi^\top}\right\}.$$

Then

$$\mathscr{I}(\vartheta) = \begin{pmatrix} \mathscr{I}_{\theta\theta}(\theta,\phi) + a\mathscr{I}_{\theta\theta}(\theta,\varphi) & \mathscr{I}_{\theta\phi}(\theta,\phi) & a\mathscr{I}_{\theta\varphi}(\theta,\varphi) \\ \mathscr{I}_{\theta\phi}(\theta,\phi)^\top & \mathscr{I}_{\phi\phi}(\theta,\phi) & 0 \\ a\mathscr{I}_{\theta\varphi}(\theta,\varphi)^\top & 0 & a\mathscr{I}_{\varphi\varphi}(\theta,\varphi) \end{pmatrix}$$

is positive definite, where $a = m/n$ and without loss of generality we assume that $m = an$ for a fixed $a > 0$. It can be seen that $\mathscr{I}(\vartheta)$ is increasing in a in the sense that $A \geq B$ for two non-negative definite matrices A and B if and only if $A-B$ is non-negative definite.

Using the standard argument in asymptotic theory, e.g., Theorem 4.17 in Shao (2003), we obtain the following result.

Theorem 2.1: *Assume (R1)–(R4) and that $m = an$ with a remaining fixed as n increases. Then, with probability tending to 1 as $n \to \infty$, there exists $\widehat{\vartheta}$ (depending on n) such that $P\{s(\widehat{\vartheta}) = 0\} \to 1$ and*

$$\sqrt{n}(\widehat{\vartheta} - \vartheta) \xrightarrow{d} N\left(0, \{\mathscr{I}(\vartheta)\}^{-1}\right), \tag{1}$$

where $\xrightarrow{d}$ denotes convergence in distribution and $N(C, D)$ is the normal distribution with mean C and covariance matrix D.

The asymptotic result (1) enables us to assess performance of $\widehat{\vartheta}$ and to carry out large sample statistical inference on parameter ϑ or any of its components θ, ϕ, and φ. When some of regularity conditions (R1) –(R4) are not satisfied, however, we may apply the bootstrap method (see Section 3.2 and Section 7 for the normal and Laplace problem) or directly derive the asymptotic distribution of $\widehat{\vartheta}$ (see Section 4 for the problem of two truncation distributions).

3. Application to location-Scale problems

An application of our general result in Section 2 is to the case where $f(x,\theta,\phi) = \frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$ and $g(y,\theta,\varphi) = \frac{1}{\tau} g(\frac{x-\nu}{\tau})$ for two continuous probability density functions f and g on real line, i.e., both populations are in location-scale families. We have several scenarios.

(1) Two location-scale families sharing the same location and scale parameters: $\mu = \nu, \sigma = \tau, \theta = (\mu,\sigma)^\top$, and both ϕ and φ are constants.
(2) Two location-scale families sharing the same location parameter but having different scale parameters: $\mu = \nu$, $\theta = \mu, \phi = \sigma$, and $\varphi = \tau$.
(3) Two location-scale families sharing the same scale parameter but having different location parameters: $\sigma = \tau$, $\theta = \sigma, \phi = \mu$, and $\varphi = \nu$.

Under any location-scale problem, it is often true that $\mathscr{I}_{\theta\phi}(\theta,\phi) = 0$ and $\mathscr{I}_{\theta\varphi}(\theta,\varphi) = 0$ and, hence, the inverse of $\mathscr{I}(\vartheta)$ can be easily obtained. For example, if both f and g are continuously differentiable functions symmetric about 0, then it follows from Example 3.9 in Shao (2003) that both $\mathscr{I}_{\theta\phi}(\theta,\phi)$ and $\mathscr{I}_{\theta\varphi}(\theta,\varphi)$ varnish.

In the following we consider a special case in details.

3.1. Normal and Laplace densities with a single scale parameter

Suppose that $f(x,\theta) = \frac{1}{\sqrt{2\pi}\theta} e^{-x^2/(2\theta^2)}$, $x \in (-\infty,\infty)$, which is the normal distribution $N(0,\theta^2)$, and that $g(y,\theta) = \frac{1}{2\theta} e^{-|y|/\theta}$, $y \in (-\infty,\infty)$, which is the Laplace distribution (also called double exponential distribution) with mean zero and standard deviation $\sqrt{2}\theta$. The two densities share the common scale parameter $\theta > 0$.

The MLEs of θ based on data from f and g, respectively, are

$$\widehat{\theta}_N = \sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2} \qquad \text{and} \qquad \widehat{\theta}_E = \frac{1}{m}\sum_{j=1}^m |Y_j|.$$

In this particular case, we can obtain an explicit form of the MLE $\widehat{\theta}$ of θ based on all data from two samples. The log likelihood based on two samples is

$$\ell(\theta) = -\sum_{i=1}^n \frac{X_i^2}{2\theta^2} - \sum_{j=1}^m \frac{|Y_j|}{\theta} - \log\{(2\pi)^{n/2} 2^m \theta^{n+m}\}.$$

The score function is

$$s(\theta) = \frac{1}{\theta^3}\sum_{i=1}^{n} X_i^2 + \frac{1}{\theta^2}\sum_{j=1}^{m} |Y_j| - \frac{n+m}{\theta}.$$

Setting $s(\theta) = 0$ and using the form of MLE from each sample, we obtain that

$$\theta^2 - \left(\frac{m\widehat{\theta}_E}{n+m}\right)\theta - \left(\frac{n\widehat{\theta}_N^2}{n+m}\right) = 0.$$

Since $\theta > 0$ and only one root is positive, we obtain that the MLE of θ is

$$\widehat{\theta} = \frac{1}{2}\left\{\frac{a\widehat{\theta}_E}{a+1} + \sqrt{\left(\frac{a\widehat{\theta}_E}{a+1}\right)^2 + \frac{4\widehat{\theta}_N^2}{a+1}}\right\}, \qquad \text{where } a = m/n. \tag{2}$$

Note that $\widehat{\theta}$ is a nonlinear function of $\widehat{\theta}_N$ and $\widehat{\theta}_E$. In general, the MLE of the shared parameter based on two datasets is not a simple function of separate MLEs based on each single dataset.

To derive the asymptotic distribution of $\widehat{\theta}$, we can use the general result (1), because regularity conditions (R1) –(R4) are satisfied for f and g. Since $\widehat{\theta}$ has an explicit form, we can also simply derive it. Because X_i's and Y_j's are independent and $a = m/n$,

$$\sqrt{n}\begin{pmatrix}\widehat{\theta}_N - \theta \\ \sqrt{a}\widehat{\theta}_E - \sqrt{a}\theta\end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix}\theta^2/2 & 0 \\ 0 & \theta^2\end{pmatrix}\right).$$

Define

$$g(t,s) = \frac{1}{2}\left\{\frac{\sqrt{a}s}{a+1} + \sqrt{\frac{as^2}{(a+1)^2} + \frac{4t^2}{a+1}}\right\}.$$

Then,

$$g\left(\widehat{\theta}_N, \sqrt{a}\widehat{\theta}_E\right) = \widehat{\theta} \quad \text{and} \quad g\left(\theta, \sqrt{a}\theta\right) = \theta.$$

Hence, by the delta method, e.g., Theorem 1.12 in Shao (2003),

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N\left(0,\ \nabla g^\top \begin{pmatrix}\theta^2/2 & 0 \\ 0 & \theta^2\end{pmatrix} \nabla g\right),$$

where ∇g is the derivative vector of g at $(t,s) = (\theta, \sqrt{a}\theta)$, i.e.,

$$\frac{\partial g}{\partial t} = \frac{2t}{a+1}\Big/\sqrt{\frac{as^2}{(a+1)^2} + \frac{4t^2}{a+1}},$$

$$\frac{\partial g}{\partial s} = \frac{1}{2}\left\{\frac{\sqrt{a}}{a+1} + \frac{as}{(a+1)^2}\Big/\sqrt{\frac{as^2}{(a+1)^2} + \frac{4t^2}{a+1}}\right\},$$

$$\nabla g = \left(\frac{2}{a+2}, \frac{\sqrt{a}}{a+2}\right)^\top.$$

This leads to the following result.

Corollary 3.1: *Assume that $m = an$ with a remaining fixed as n increases. Then, as $n \to \infty$,*

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{\theta^2}{a+2}\right).$$

The asymptotic relative efficiency of $\widehat{\theta}_N$ with respect to $\widehat{\theta}$ is $2/(a+2)$, which is decreasing in a and bounded between 0 and 1. The asymptotic relative efficiency of $\widehat{\theta}_E$ with respect to $\widehat{\theta}$ is $a/(a+2)$, which is increasing in a and bounded between 0 and 1.

3.2. Normal and Laplace densities with shared scale and location parameters

Consider a more general case where f and g share a scale parameter and a location parameter. That is, $f(x,\theta,\mu) = \frac{1}{\sqrt{2\pi}\theta}e^{-(x-\mu)^2/(2\theta^2)}$, $x \in (-\infty,\infty)$, which is the normal distribution $N(\mu,\theta^2)$, and $g(y,\theta,\mu) = \frac{1}{2\theta}e^{-|y-\mu|/\theta}$, $y \in (-\infty,\infty)$, which is the Laplace distribution with mean μ and standard deviation $\sqrt{2}\theta$. Note that regularity conditions (R1) –(R4) are not satisfied for g, since g is not always differentiable in μ.

For parameter vector $\vartheta = (\mu,\theta)^\top$, the log likelihood is

$$\ell(\vartheta) = -\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{2\theta^2} - \sum_{j=1}^{m}\frac{|Y_j-\mu|}{\theta} - \log\{(2\pi)^{n/2}2^m\theta^{n+m}\}.$$

Although $\ell(\vartheta)$ is not always differentiable in μ, it is concave in μ and, hence, the MLE $\widehat{\mu}$ of μ exists though it does not have an explicit form. The MLE of θ is given by (2) with $\widehat{\theta}_N$ and $\widehat{\theta}_E$ replaced by, respectively,

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\widehat{\mu})^2} \qquad \text{and} \qquad \frac{1}{m}\sum_{j=1}^{m}|Y_j-\widehat{\mu}|.$$

The asymptotic distribution of $\widehat{\vartheta} = (\widehat{\mu},\widehat{\theta})^\top$ cannot be obtained from (1), since g does not satisfy conditions (R1) –(R4). For assessing performance of $\widehat{\vartheta}$ and/or making inference, we recommend a bootstrap method, which is discussed in Section 7 and studied by simulation.

4. Application to two truncation distributions

Let $f(x,\theta)$ and $g(y,\theta)$ be positive density functions on the interval $(0,\theta)$ and zero outside $(0,\theta)$, where $\theta > 0$ is an unknown scale parameter common for both populations, and f and g are known when θ is known. The likelihood is

$$\prod_{i=1}^{n}f(X_i,\theta)I_{\{X_i<\theta\}}\prod_{j=1}^{m}g(Y_j,\theta)I_{\{Y_j<\theta\}} = \left\{\prod_{i=1}^{n}f(X_i,\theta)\prod_{j=1}^{m}g(Y_j,\theta)\right\}I_{\{X_{(n)}<\theta\}}I_{\{Y_{(m)}<\theta\}},$$

where I_A is the indicator of event A, $X_{(n)} = \max(X_1,\ldots,X_n)$ and $Y_{(m)} = \max(Y_1,\ldots,Y_m)$. This likelihood is not always differentiable in θ, but it can be seen that the MLE of θ is $\widehat{\theta} = \max(X_{(n)}, Y_{(m)})$, a maximizer of the likelihood.

This is an example in which regularity conditions (R1) –(R4) in Section 2 are not satisfied so that result (1) does not hold. The MLE $\widehat{\theta}$ is not even asymptotically normal. In the following we directly derive the asymptotic distribution of $\widehat{\theta}$.

It follows from the result in Example 2.34 of Shao (2003), the independence of X_i's and Y_j's, and $m = an$ that

$$n\begin{pmatrix}\theta - X_{(n)}\\ \theta - Y_{(m)}\end{pmatrix} \xrightarrow{d} \begin{pmatrix}\dfrac{\varepsilon_1}{f(\theta,\theta)}\\ \dfrac{\varepsilon_2}{a\,g(\theta,\theta)}\end{pmatrix},$$

where ε_1 and ε_2 are independent random variables with the same exponential distribution having density e^{-x}, $x>0$. Because

$$\min\left\{n(\theta - X_{(n)}),\, n(\theta - Y_{(m)})\right\} = n\left\{\theta - \max(X_{(n)},\, Y_{(m)})\right\} = n(\theta-\widehat{\theta}),$$

we obtain that

$$n(\theta-\widehat{\theta}) \xrightarrow{d} \min\left\{\frac{\varepsilon_1}{f(\theta,\theta)},\, \frac{\varepsilon_2}{a\,g(\theta,\theta)}\right\}.$$

From the independence of ε_1 and ε_2, for any $t>0$,

$$\begin{aligned}P\left\{\min\left\{\frac{\varepsilon_1}{f(\theta,\theta)},\, \frac{\varepsilon_2}{ag(\theta,\theta)}\right\} > t\right\} &= P\left\{\frac{\varepsilon_1}{f(\theta,\theta)} > t\right\}P\left\{\frac{\varepsilon_2}{ag(\theta,\theta)} > t\right\}\\ &= \exp\left\{-t\{f(\theta,\theta)+ag(\theta,\theta)\}\right\}.\end{aligned}$$

This leads to the following result.

Theorem 4.1: *Under the assumed conditions on f and g in this section,*

$$n(\theta - \widehat{\theta}) \xrightarrow{d} E(\theta, a),$$

where $E(\theta, a)$ *is the exponential distribution with scale parameter* $1/\{f(\theta,\theta) + ag(\theta,\theta)\}$.

Inference on θ can be made using this asymptotic result.

The asymptotic relative efficiency of the MLE $X_{(n)}$ based on the first dataset with respect to the MLE $\widehat{\theta}$ based on two datasets is $\{1 + ag(\theta,\theta)/f(\theta,\theta)\}^{-2}$, which is increasing in a and bounded between 0 and 1. The asymptotic relative efficiency of the MLE $Y_{(m)}$ based on the second dataset with respect to the MLE $\widehat{\theta}$ based on two datasets is $\{1 + a^{-1}f(\theta,\theta)/g(\theta,\theta)\}^{-2}$, which is decreasing in a and bounded between 0 and 1.

5. Application to Poisson and binomial samples

Here we consider a discrete data problem, where X_i has the Poisson distribution with mean θ, Y_j is binary with $P(Y_j = 1) = \theta$, and $\theta \in (0, 1)$ is the shared parameter. Let $\overline{X}$ be the sample mean of X_i's and $\overline{Y}$ be the sample mean of Y_j's. The score function based on two samples is

$$s(\theta) = n\left(\frac{\overline{X}}{\theta} - 1 + \frac{a\overline{Y}}{\theta} - \frac{1 - a\overline{Y}}{1 - \theta}\right), \qquad \text{where } a = m/n.$$

Setting $s(\theta) = 0$, we obtain the score equation

$$\theta^2 - (1 + a + \overline{X})\theta + \overline{X} + a\overline{Y} = 0.$$

Since the score equation is a quadratic equation, it has two solutions if and only if

$$(1 + a + \overline{X})^2 - 4(\overline{X} + a\overline{Y}) > 0.$$

By the law of large numbers, as $n \to \infty$, both $\overline{X}$ and $\overline{Y}$ converge to θ almost surely and

$$(1 + a + \overline{X})^2 - 4(\overline{X} + a\overline{Y}) \to (1 + a + \theta)^2 - 4(1 + a)\theta = (1 + a - \theta)^2 > 0$$

almost surely. This shows that, with probability tending to 1 as $n \to \infty$, the score equation has two real solutions,

$$\left\{1 + a + \overline{X} \pm \sqrt{(1 + a + \overline{X})^2 - 4(\overline{X} + a\overline{Y})}\right\}/2.$$

The solution with $+$ sign in front of the squared root is always larger than 1, out of the range $(0, 1)$ for θ in this problem. Hence, we conclude that the MLE of θ is

$$\widehat{\theta} = \min\left\{1, \frac{1 + a + \overline{X} - \sqrt{\left(1 + a + \overline{X}\right)^2 - 4(\overline{X} + a\overline{Y})}}{2}\right\}.$$

The minimum is taken because $0 < \theta < 1$. Again, the MLE $\widehat{\theta}$ is a nonlinear function of the separate MLEs, $\overline{X}$ and $\overline{Y}$.

The asymptotic distribution of $\widehat{\theta}$ can be derived using the delta-method, but because regularity conditions (R1) –(R4) are satisfied, it is a corollary of Theorem 2.1 in Section 2.

Corollary 5.1: *Under the Poisson and binary assumptions for two datasets and* $m = an$, *as* $n \to \infty$,

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{\theta(1 - \theta)}{1 - \theta + a}\right).$$

The asymptotic relative efficiency of the MLE $\overline{X}$ based on the first dataset with respect to the MLE $\widehat{\theta}$ based on two datasets is $(1 - \theta)/(1 - \theta + a)$, which is decreasing in a and bounded between 0 and 1. The asymptotic relative efficiency of the MLE $\overline{Y}$ based on the second dataset with respect to the MLE $\widehat{\theta}$ based on two datasets is $a/(1 - \theta + a)$, which is increasing in a and bounded between 0 and 1.

6. MLEs with two samples and an additional uncertainty

In this section, we consider a scenario in which the first sample is obtained under a controlled study so that we know the form of probability density $f(x,\theta,\phi)$, but the form of $g(y,\theta,\varphi)$ for the second sample has an additional uncertainty, because the second sample may be obtained through a past study and/or public records. We assume that the additional uncertainty comes from an unknown parameter ζ taking two possible values, 0 and 1, i.e., the probability density of the second sample is $g(y,\theta,\varphi,\zeta)$, where $\zeta = 0$ or 1 and g is still a known density when θ, φ, and ζ are known.

How do we derive the MLE of $\vartheta = (\theta^\top, \phi^\top, \varphi^\top)^\top$? If ζ is known, then the MLE can be obtained using the method in Section 2. Since ζ takes only two values, if $\widehat{\zeta}$ is a consistent estimator of ζ, i.e.,

$$\lim_{n\to\infty} P\left(\widehat{\zeta} = \zeta\right) = 1, \tag{3}$$

then we obtain the MLE of ϑ as

$$\widehat{\vartheta} = \begin{cases} \widehat{\vartheta}(0), & \widehat{\zeta} = 0, \\ \widehat{\vartheta}(1), & \widehat{\zeta} = 1, \end{cases}$$

where $\widehat{\vartheta}(0)$ and $\widehat{\vartheta}(1)$ are MLEs under $\zeta = 0$ and $\zeta = 1$, respectively.

A suggested consistent estimator of ζ is the MLE of ζ based on the second sample, Y_j's. Let $\widehat{\theta}(\zeta)$ and $\widehat{\varphi}(\zeta)$ be the MLEs of θ and φ, respectively, based on Y_j's, when the value of ζ is fixed. Then the MLE of ζ is

$$\widehat{\zeta} = \begin{cases} 0, & \prod_{j=1}^{m} g(Y_j, \widehat{\theta}(0), \widehat{\varphi}(0), 0) \geq \prod_{j=1}^{m} g(Y_j, \widehat{\theta}(1), \widehat{\varphi}(1), 1), \\ 1, & \prod_{j=1}^{m} g(Y_j, \widehat{\theta}(0), \widehat{\varphi}(0), 0) < \prod_{j=1}^{m} g(Y_j, \widehat{\theta}(1), \widehat{\varphi}(1), 1). \end{cases}$$

The following result gives the asymptotic distribution of the MLE $\widehat{\vartheta}$.

Theorem 6.1: *If* (3) *holds and regularity conditions (R1)–(R4) are satisfied when* $\zeta = 0$ *or* 1, *and if* $m = an$ *with* a *remaining fixed as* n *increases, then*

$$\sqrt{n}(\widehat{\vartheta} - \vartheta) \xrightarrow{d} N\left(0,\ \{\mathcal{I}(\vartheta,\zeta)\}^{-1}\right),$$

where $\mathcal{I}(\vartheta,\zeta)$ *is the Fisher information as defined in Section 2 under the true value of* ζ.

Condition (3) has to be checked for each particular problem. The following is an example.

Suppose that $f(x,\theta)$ is the density of $N(0,\theta^2)$, $g(y,\theta,0)$ is the same normal density for $N(0,\theta^2)$ but $g(y,\theta,1)$ is the Laplace distribution with zero mean and standard deviation $\sqrt{2}\theta$ given in Section 3.1. In other words, sample one is from the main study whereas sample two is from an external source in which the data may follow the same distribution as sample one but may deviate from sample one. The parameters ϕ and φ are constant (non-existing).

In this example, when $\widehat{\zeta} = 0$, we can simply combine the two samples and the MLE of θ is $\sqrt{(\sum_{i=1}^{n} X_i^2 + \sum_{j=1}^{m} Y_j^2)/(n+m)}$; on the other hand, when $\widehat{\zeta} = 1$, the MLE of θ is given by (2). To check (3), note that

$$\widehat{\theta}(0) = \sqrt{\frac{1}{m}\sum_{j=1}^{m} Y_j^2} \qquad \text{and} \qquad \widehat{\theta}(1) = \frac{1}{m}\sum_{j=1}^{m} |Y_j|.$$

Then,

$$\log\left\{\prod_{j=1}^{m} g(\widehat{\theta}(0), 0)\right\} = -\frac{m}{2} - m\log\widehat{\theta}(0) - \frac{m\log(2\pi)}{2}$$

and

$$\log\left\{\prod_{j=1}^{m} g(\widehat{\theta}(1),1)\right\} = -m - m\log\widehat{\theta}(1) - m\log 2.$$

When $\zeta = 0$, $\widehat{\theta}(0) \xrightarrow{p} \theta$ and $\widehat{\theta}(1) \xrightarrow{p} (2/\pi)^{1/2}\theta$, where $\xrightarrow{p}$ denotes convergence in probability as $n \to \infty$. Hence

$$\frac{1}{m}\log\left\{\prod_{j=1}^{m} g(\widehat{\theta}(0),0)\right\} - \frac{1}{m}\log\left\{\prod_{j=1}^{m} g(\widehat{\theta}(1),1)\right\} \xrightarrow{p} \frac{1}{2} + \log\frac{2}{\pi} > 0,$$

which implies that $P(\widehat{\zeta} = 0) \to 1$. On the other hand, when $\zeta = 1$, $\widehat{\theta}(0) \xrightarrow{p} \sqrt{2}\theta$, $\widehat{\theta}(1) \xrightarrow{p} \theta$, and

$$\frac{1}{m}\log\left\{\prod_{j=1}^{m} g(\widehat{\theta}(0),0)\right\} - \frac{1}{m}\log\left\{\prod_{j=1}^{m} g(\widehat{\theta}(1),1)\right\} \xrightarrow{p} \frac{1}{2} - \frac{\log\pi}{2} < 0,$$

which implies that $P(\widehat{\zeta} = 1) \to 1$. This shows that (3) always holds in this example.

Still in this example, the results here and in Section 3.1 indicate that

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \begin{cases} N\left(0, \frac{\theta^2}{2a+2}\right), & \zeta = 0, \\ N\left(0, \frac{\theta^2}{a+2}\right), & \zeta = 1. \end{cases}$$

The result can obviously be extended to the situation where the second sample is from a population that is one of k populations with $k \geq 3$.

7. Bootstrap variance estimation

In situations where regularity conditions (R1) –(R4) are not satisfied for f or g, the asymptotic distribution of MLE $\widehat{\vartheta}$ may not be available, either it does not exist or it is not established. Here, we introduce a bootstrap variance estimator which can be used for assessing performance of $\widehat{\vartheta}$ or making large sample inference. A description about the general bootstrap methodology can be found, for example, in Efron and Tibshirani (1993) and Shao (2003).

Let $\{X_1^{*b}, \ldots, X_n^{*b}\}$ and $\{Y_1^{*b}, \ldots, Y_m^{*b}\}$ be two independent simple random samples with replacement from $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$, respectively, and let $\widehat{\vartheta}^{*b}$ be the MLE of ϑ based on dataset $\{X_1^{*b}, \ldots, X_n^{*b}, Y_1^{*b}, \ldots, Y_m^{*b}\}$. If we independently repeat this for $b = 1, \ldots, B$, where B is called the bootstrap replication size and is typically large, then the bootstrap variance estimator for $\widehat{\vartheta}$ is the sample covariance matrix of $\widehat{\vartheta}^{*b}$, $b = 1, \ldots, B$.

We carry out a simulation study to examine the performance of this bootstrap variance estimator in the normal-Laplace problem considered in Section 3.2. At the same time, we also check the performance of MLE $(\widehat{\mu}, \widehat{\theta})$ based on two datasets, X_i's and Y_i's, and compare it with $(\overline{X}, \widehat{\theta}_X)$ and $(\widetilde{Y}, \widehat{\theta}_Y)$, which are the MLEs based on the single dataset of X_i's and single dataset of Y_i's, respectively, where $\overline{X}$ = sample mean of X_i's, $\widetilde{Y}$ = sample median of Y_j's, $\widehat{\theta}_X = \{\sum_{i=1}^{n}(X_i - \overline{X})^2/n\}^{1/2}$, and $\widehat{\theta}_Y = \sum_{i=1}^{m}|Y_i - \widetilde{Y}|/m$. The bootstrap is applied to obtain $\widehat{\mathrm{SD}}$ for the standard deviation (SD) of any fixed point estimator.

The simulation results with 1000 replications are shown in Table 1. A summary is given as follows.

(1) The MLE's, $\widehat{\mu}$, $\overline{X}$, and $\widetilde{Y}$, all have almost no bias as estimators of μ (= 1 in simulation). In terms of the SD, The MLE $\widehat{\mu}$ is the best among the three. The sample median based on Y_j's is substantially worse than the other two, although asymptotically it is as efficient as the sample mean $\overline{X}$ of X_i's.

Table 1. Results from 1000 simulations for the normal-Laplace problem with location $\mu = 1$ and scale $\theta = 1$ ($n = m = 100$, SD = standard deviation, $(\widehat{\mu}, \widehat{\theta})$ = the MLE of (μ, θ) based on X_i's and Y_j's, $(\overline{X}, \widehat{\theta}_X)$ = the MLE of (μ, θ) based on X_i's, $(\widetilde{Y}, \widehat{\theta}_Y)$ = the MLE of (μ, θ) based on Y_j's, and $\widehat{\mathrm{SD}}$ is by bootstrap with $B = 500$).

	$\widehat{\mu}$	$\overline{X}$	$\widetilde{Y}$	$\widehat{\theta}$	$\widehat{\theta}_X$	$\widehat{\theta}_Y$
Mean by simulation	1.0024	1.0012	1.0034	1.1397	0.9956	1.3979
SD by simulation	0.0856	0.0969	0.1579	0.0675	0.0721	0.1367
Mean of $\widehat{\mathrm{SD}}$ by simulation	0.0863	0.0996	0.1593	0.0668	0.0687	0.1378
SD of $\widehat{\mathrm{SD}}$	0.0124	0.0079	0.0376	0.0069	0.0094	0.0194

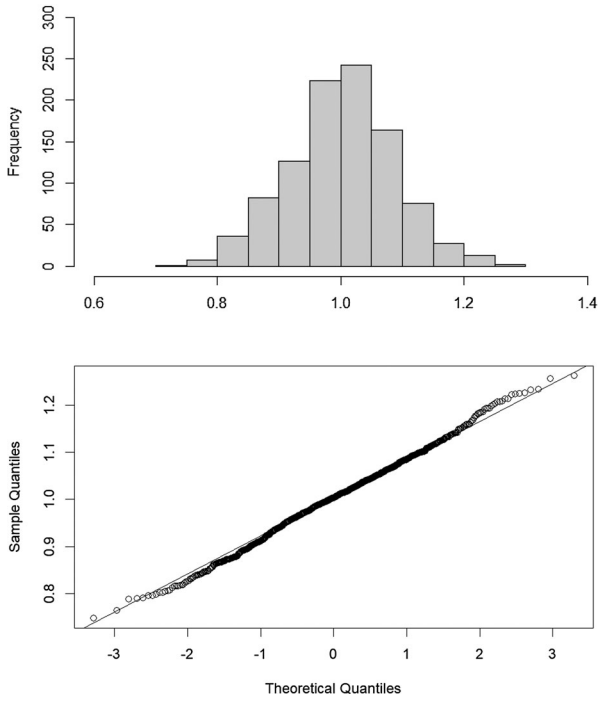

Figure 1. Histogram and Q–Q plot of 1000 simulated values of $\widehat{\mu}$.

(2) The MLE $\widehat{\theta}$ of θ does not have a negligible bias, although its performance is acceptable with sample size $n+m=200$ and its SD is slightly smaller than the SD of $\widehat{\theta}_X$. The large bias of the MLE $\widehat{\theta}$ mainly comes from the large bias of the MLE $\widehat{\theta}_Y$ for the Laplace dataset, as it has large bias and SD.
(3) The bootstrap SD estimator $\widehat{\text{SD}}$ performs very well for all estimators (see the rows under "SD by simulation" and "mean of $\widehat{\text{SD}}$ by simulation" in Table 1), even when the point estimator has non-negligible bias.

The histogram of 1000 values of $\widehat{\mu}$ from simulation is shown in Figure 1, together with a Q–Q plot. The result suggests $\widehat{\mu}$ is asymptotically normal, although such a theoretical result has not been established.

Acknowledgments

The authors would like to thank two anonymous referees for helpful comments and suggestions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Jun Shao's research was partially supported by the National Natural Science Foundation of China [Grant Number 11831008] and the U.S. National Science Foundation [Grant Number DMS-1914411].

References

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Halll/CRC.
Kim, H. J., Wang, Z., & Kim, J. K (2021). Survey data integration for regression analysis using model calibration. *arXiv 2107.06448*.
Lohr, S. L., & Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science*, *32*(2), 293–312. https://doi.org/10.1214/16-STS584
Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, *99*(468), 1131–1139. https://doi.org/10.1198/016214504000000601
Rao, J. N. K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, *83*(1), 242–272. https://doi.org/10.1007/s13571-020-00227-w
Shao, J. (2003). *Mathematical statistics*. 2nd ed. Springer.
Yang, S., & Kim, J. K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, *3*(2), 625–650. https://doi.org/10.1007/s42081-020-00093-w

Zhang, Y., Ouyang, Z., & Zhao, H. (2017). A statistical framework for data integration through graphical models with application to cancer genomics. *The Annals of Applied Statistics*, *11*(1), 161–184. https://doi.org/10.1214/16-AOAS998

Zieschang, K. D. (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association*, *85*(412), 986–1001. https://doi.org/10.1080/01621459.1990.10474969