Statistica Sinica Preprint No: SS-2023-0056	
Title	Empirical Likelihood Using External Summary Information
Manuscript ID	SS-2023-0056
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0056
Complete List of Authors	Lyu Ni, Jun Shao, Jinyi Wang and Lei Wang
Corresponding Authors	Lei Wang
E-mails	lwangstat@nankai.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

Empirical Likelihood Using External Summary Information

Lyu Ni[1], Jun Shao[2], Jinyi Wang[2] and Lei Wang[3]

[1]*East China Normal University,* [2] *University of Wisconsin-Madison*
and [3]*Nankai University*

Abstract: Statistical analysis in modern scientific research nowadays has opportunities to utilize external summary information from similar studies to gain efficiency. However, the population generating data for current study, referred to as internal population, is typically different from the external population for summary information, although they share some common characteristics that make efficiency improvement possible. The existing population heterogeneity is a challenging issue especially when we have only summary statistics but not individual-level external data. In this paper, we apply an empirical likelihood approach to estimating internal population distribution, with external summary information utilized as constraints for efficiency gain under population heterogeneity. We show that our approach produces an asymptotically more efficient estimator of internal population distribution compared with the customary empirical likelihood without using any external information, under the condition that the external information is based on a dataset with size larger than that

Address for correspondence: Lei Wang, School of Statistics and Data Science, KLMDASR, LEBPS and LPMC, Nankai University. E-mail: lwangstat@nankai.edu.cn.

of the dataset from internal population. Some simulation results are given to supplement asymptotic theory. A real data example is also illustrated.

Key words and phrases: constraints, data integration, population heterogeneity, quantile estimation, shared parameters, summary statistics.

1. Introduction

Consider the estimation of a population distribution $F_{X,Z}$ defined on the k-dimensional Euclidean space $\mathscr{R}^k$, where X and Z are vectors with dimensions l and $k-l$, respectively, based on a random sample $\{X_i, Z_i, i = 1, ..., n\}$ from $F_{X,Z}$. Nowadays we often also have information in the form of summary statistics, not necessarily individual-level data, from external sources (such as past similar studies), which can be utilized to increase statistical accuracy in estimating $F_{X,Z}$ and its characteristics. Specifically, there is an external sample $\{X_i^E, i = 1, ..., m\}$, independent of $\{X_i, Z_i, i = 1, ..., n\}$, from an external population distribution F_X^E, where X^E and X measure the same quantity and have the same dimension l, but F_X^E is not necessarily the same as F_X, the distribution of X. When $l < k$, the vector Z is not measured externally due to progress of new technology and/or new scientific relevance or other practical reasons. In what follows, $\{X_i, Z_i, i = 1, ..., n\}$ and $F_{X,Z}$ are referred to as internal data and internal population, respectively, to distinguish external data $\{X_i^E, i = 1, ..., m\}$ and external population F_X^E.

The purpose of our study is to develop estimation methodology using internal data and external summary statistics (functions of $X_1^E, ..., X_m^E$), when individual-level external data $X_1^E, ..., X_m^E$ are not available due to some practical reasons. This problem has been studied in Chatterjee et al. (2016) and Zhang et al. (2020) when $F_{X,Z}$ follows a parametric model, whereas we study the estimation of $F_{X,Z}$ with the nonparametric empirical likelihood approach. This research fits into a general framework of data integration (Merkouris, 2004; Lin and Zeng, 2010; Lohr and Raghunathan, 2017; Zhang et al., 2017; Kundu et al., 2019; Yang et al., 2023; Yang and Kim, 2020; Wang et al., 2023; Rao, 2021; Li et al., 2022; Tian and Feng, 2022).

Our research takes into consideration of the heterogeneity between internal and external populations F_X and F_X^E, although they share some common part as a link that makes it possible to improve the estimation of $F_{X,Z}$ using external information. To the best of our knowledge, population heterogeneity is not well addressed in coupling internal data and external summary information. For example, Chatterjee et al. (2016) and Zhang et al. (2020) assume $F_X = F_X^E$, when only external summary statistic is available.

To present the main ideas, we focus on one external dataset, since extensions to multiple external datasets are straightforward. Our main method is empirical likelihood using the external summary statistic in a constraint. We establish asymptotic normality of estimators of $F_{X,Z}$ with explicit formulas of asymptotic covariance matrices, which can be used to compare their asymptotic efficiency with the customary estimator without using any external information and to make inference on $F_{X,Z}$ or its characteristics. Some simulation results are presented as complementary to asymptotic theory. A real data sample is also illustrated.

2. Empirical Likelihood with External Information

We follow the notation developed in Section 1. To link the internal and external populations F_X and F_X^E for the purpose of increasing the accuracy in estimating $F_{X,Z}$, where F_X is the l-dimensional marginal of the k-dimensional internal population $F_{X,Z}$ of interest, we assume that there is a p-dimensional parameter vector θ shared by both F_X and F_X^E and defined by

$$\int u(x,\theta)dF_X(x) = \int u(x,\theta)dF_X^E(x) = 0, \tag{2.1}$$

where $u(\cdot,\cdot)$ is a known vector function from $\mathscr{R}^l \times \mathscr{R}^p$ to $\mathscr{R}^p$ with continuous partial derivatives with respect to θ. For example, $u(x,\theta) = x - \theta$, in which case $p = l$ and θ is the common mean of F_X and F_X^E.

Let $\widehat{\theta}^E$ be a p-dimensional estimator of θ in (2.1) based on external data via a generalized estimation equation, i.e.,

$$\frac{1}{m}\sum_{i=1}^{m} u(X_i^E, \widehat{\theta}^E) = 0, \tag{2.2}$$

an empirical analog of $\int u(x,\theta)dF_X^E(x) = 0$ in (2.1) based on X_i^E's. For example, in the common mean example where $u(x,\theta) = x - \theta$, $\widehat{\theta}^E$ is the sample mean $\bar{X}^E$ of $X_1^E, ..., X_m^E$. Note that in the current paper we only have the value of $\widehat{\theta}^E$ as external summary statistic (information), not individual-level values $X_1^E, ..., X_m^E$.

To make use of external information $\widehat{\theta}^E$, we require that any estimate $\widehat{F}_{X,Z}$ of $F_{X,Z}$ based on internal data has property

$$\int u(x, \widehat{\theta}^E) d\widehat{F}_X(x) = 0, \tag{2.3}$$

an empirical analog of $\int u(x,\theta)dF_X(x) = 0$ in (2.1), where $\widehat{F}_X$ is the l-dimensional marginal of $\widehat{F}_{X,Z}$ for X. Using the method of empirical likelihood (Owen, 1988, 2001; Qin and Lawless, 1994), we use (2.3) as a constraint in the estimation of $F_{X,Z}$ based on internal data, treating $\widehat{\theta}^E$ as known. That is, we estimate $F_{X,Z}$ by $\widehat{F}_{X,Z}$ as a maximizer of

$$\prod_{i=1}^{n} p_i \quad \text{subject to} \;\; p_i > 0, \quad i = 1, ..., n, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i u(X_i, \widehat{\theta}^E) = 0, \tag{2.4}$$

where p_i is a point mass of any distribution whose support consists of n points (X_i, Z_i), $i = 1, ..., n$, and $\widehat{\theta}^E$ is treated as fixed and known in maximization. The resulting estimator $\widehat{F}_{X,Z}$ satisfies (2.3).

The customary estimator without using any external information, the empirical distribution $\bar{F}_{X,Z}$ putting mass n^{-1} at each (X_i, Z_i), is a maximizer of (2.4) only when $u \equiv 0$, because, when $u \not\equiv 0$, $\int u(x, \widehat{\theta}^E) d\bar{F}_X(x) = n^{-1} \sum_{i=1}^{n} u(X_i, \widehat{\theta}^E)$ is typically not 0 although $\int u(x, \theta) dF_X^E(x) = 0$, where $\bar{F}_X$ is the l-dimensional marginal of $\bar{F}_{X,Z}$ for X. This is the reason why $\widehat{F}_{X,Z}$ can be more efficient than $\bar{F}_{X,Z}$.

Using the Lagrange multiplier method, we can show that the maximizer of (2.4) is the distribution

$$\widehat{F}_{X,Z} \;\; \text{putting mass} \;\; \widehat{p}_i = \frac{1}{n\{1 + \lambda^\top u(X_i, \widehat{\theta}^E)\}} \quad \text{at each } (X_i, Z_i), \tag{2.5}$$

where $\lambda \in \mathscr{R}^p$ is the Lagrange multiplier satisfying

$$\sum_{i=1}^{n} \widehat{p}_i u(X_i, \widehat{\theta}^E) = \frac{1}{n} \sum_{i=1}^{n} \frac{u(X_i, \widehat{\theta}^E)}{1 + \lambda^\top u(X_i, \widehat{\theta}^E)} = 0,$$

and $a^\top$ denotes the transpose of vector a.

Note that

$$\frac{\partial}{\partial\lambda}\left[\frac{1}{n}\sum_{i=1}^{n}\log\{1+\lambda^{\top}u(X_i,\widehat{\theta}^E)\}\right]=\frac{1}{n}\sum_{i=1}^{n}\frac{u(X_i,\widehat{\theta}^E)}{1+\lambda^{\top}u(X_i,\widehat{\theta}^E)}$$

and

$$\frac{\partial^2}{\partial\lambda\partial\lambda^{\top}}\left[\frac{1}{n}\sum_{i=1}^{n}\log\{1+\lambda^{\top}u(X_i,\widehat{\theta}^E)\}\right]=-\frac{1}{n}\sum_{i=1}^{n}\frac{u(X_i,\widehat{\theta}^E)u(X_i,\widehat{\theta}^E)^{\top}}{\{1+\lambda^{\top}u(X_i,\widehat{\theta}^E)\}^2}<0$$

(negative definite) if $u\not\equiv 0$. Hence, there is a unique sequence of $\{\lambda=\lambda(X_1,...,X_n,\widehat{\theta}^E),n=1,2,...\}$ such that

$$\lim_{n\to\infty}P\left(\frac{1}{n}\sum_{i=1}^{n}\frac{u(X_i,\widehat{\theta}^E)}{1+\lambda^{\top}u(X_i,\widehat{\theta}^E)}=0\right)=1\qquad\text{and}\qquad\lambda=o_p(1),\quad(2.6)$$

where $o_p(1)$ denotes a term converging to 0 in probability as $n\to\infty$. Therefore, with probability tending to 1, $\widehat{F}_{X,Z}$ is uniquely defined.

For any s fixed $t_1,...,t_s$ in $\mathscr{R}^k$, we define $\mathcal{F}=\big(F_{X,Z}(t_1),...,F_{X,Z}(t_s)\big)^{\top}$, $\widehat{\mathcal{F}}=\big(\widehat{F}_{X,Z}(t_1),...,\widehat{F}_{X,Z}(t_s)\big)^{\top}$ for estimator $\widehat{F}_{X,Z}$ in (2.5), and define $\bar{\mathcal{F}}=\big(\bar{F}_{X,Z}(t_1),...,\bar{F}_{X,Z}(t_s)\big)^{\top}$ for the empirical distribution $\bar{F}_{X,Z}$. Also, define $\bar{u}=n^{-1}\sum_{i=1}^{n}u(X_i,\theta)$, $U=\text{Var}\{u(X,\theta)\}$ (assumed to be non-singular), and $W=\big(W(t_1),...,W(t_s)\big)$, $W(t_j)=E\{u(X,\theta)I(\binom{X}{Z}\le t_j)\}$, where $I(A)$ denotes the indicator function of event A and $a\le b$ for vectors a and b means that every component of a is no larger than the corresponding component of b. Following the argument in the proof of Theorem 5.4 in Shao (2003), we obtain that

$$\sqrt{n}(\widehat{\mathcal{F}}-\mathcal{F})=\sqrt{n}(\bar{\mathcal{F}}-\mathcal{F}-\bar{u}^{\top}U^{-1}W-\tilde{u}^{\top}M^{-\top}L^{\top}U^{-1}W)+o_p(1),\quad(2.7)$$

where $L = \int\{\partial u(x,\theta)/\partial\theta\}dF_X(x)$, $M = \int\{\partial u(x,\theta)/\partial\theta\}dF_X^E(x)$ (assumed to be non-singular), $M^{-\top} = (M^{-1})^\top$, $\widetilde{u} = m^{-1}\sum_{j=1}^m u(X_j^E,\theta)$, and the last equality follows from

$$\widehat{\theta}^E - \theta = -M^{-1}\widetilde{u} + m^{-1/2}o_p(1) \tag{2.8}$$

by (2.2) and Taylor's expansion. The covariance matrix

$$\begin{aligned} & n\mathrm{Var}(\bar{\mathcal{F}} - \bar{u}^\top U^{-1}W - \widetilde{u}^\top M^{-\top}L^\top U^{-1}W) \\ = {} & \Lambda - W^\top U^{-1}W + m^{-1}nW^\top U^{-1}LM^{-1}VM^{-\top}L^\top U^{-1}W, \end{aligned}$$

where $\Lambda = n\mathrm{Var}(\bar{\mathcal{F}})$, the $k \times k$ matrix whose (i,j)th element is equal to $P\big(\{(X,Z) \le t_i\} \cap \{(X,Z) \le t_j\}\big) - F_{X,Z}(t_i)F_{X,Z}(t_j)$, $V = m\mathrm{Var}(\widetilde{u})$, and the equality follows from

$$n\mathrm{Var}(\bar{u}^\top U^{-1}W) = n\mathrm{Cov}(\bar{\mathcal{F}},\ \bar{u}^\top U^{-1}W)$$

and

$$n\mathrm{Var}(\widetilde{u}^\top M^{-\top}L^\top U^{-1}W) = nW^\top U^{-1}LM^{-1}\mathrm{Var}(\widetilde{u})M^{-\top}L^\top U^{-1}W.$$

Consequently, by the central limit theorem, we obtain the following result.

Theorem 1. *Assume (2.1), $\widehat{\theta}^E$ defined by (2.2) satisfies (2.8) as $m \to \infty$, and matrices L, M, U, and V are non-singular. Then, for any s fixed* *distinct* $t_1, ..., t_s$ *in* $\mathscr{R}^k$, *as* $n \to \infty$ *and* $m \to \infty$,

$$\sqrt{n}\{\big(\widehat{F}_{X,Z}(t_1), ..., \widehat{F}_{X,Z}(t_s)\big)^\top - \big(F_{X,Z}(t_1), ..., F_{X,Z}(t_s)\big)^\top\} \xrightarrow{d} N(0, \Sigma),$$

$$\Sigma = \Lambda - W^\top U^{-1} W + r\, W^\top U^{-1} L M^{-1} V M^{-\top} L^\top U^{-1} W, \tag{2.9}$$

where $\xrightarrow{d}$ *denotes convergence in distribution,* $N(0, \Sigma)$ *is the normal distribution with mean 0 and covariance matrix* Σ, *and* r *is the limit of* n/m.

Result (2.9) indicates how statistical accuracy can be affected through using external information provided by $\widehat{\theta}^E$, since Λ is the asymptotic covariance matrix for the customary empirical distribution $\bar{F}_{X,Z}$ without using any external information. If the sample size of external dataset m dominates the sample size of internal dataset, i.e., $r = 0$, then Σ in (2.9) is $\Lambda - W^\top U^{-1} W$, smaller than Λ (in the order for nonnegative definite matrices) and, hence, $\widehat{F}_{X,Z}$ in (2.5) is asymptotically more efficient than $\bar{F}_{X,Z}$. If $r > 0$, then whether $\widehat{F}_{X,Z}$ is better depends on the magnitude of the last two terms in Σ in (2.9) involving the quality of external information. In the special case where $L = M$ (e.g., when $u(x, \theta) = x - \theta$) and $V = U$, Σ in (2.9) reduces to $\Lambda - (1 - r) W^\top U^{-1} W$ and, thus, $\widehat{F}_{X,Z}$ is better than $\bar{F}_{X,Z}$ if and only if $r < 1$ (the external dataset has a large size than the internal dataset).

If we estimate a characteristic of $F_{X,Z}$ given as $\psi(F_{X,Z})$, a functional of $F_{X,Z}$, then $\psi(\widehat{F}_{X,Z})$ is asymptotically more efficient than $\psi(\bar{F}_{X,Z})$ when $\widehat{F}_{X,Z}$ is more efficient than $\bar{F}_{X,Z}$. Specific examples are given in Section 4.

Result (2.9) is useful for large sample inference on characteristics of population $F_{X,Z}$. To make inference, we need to estimate the covariance matrix Σ in (2.9), which requires some additional external information for the variability of $\widehat{\theta}^E$. From (2.8) and $V = m\mathrm{Var}(\widetilde{u})$, the asymptotic covariance matrix for $\sqrt{m}(\widehat{\theta}^E - \theta)$ is $\Xi = M^{-1}VM^{-\top}$. Assume that, together with $\widehat{\theta}^E$ in (2.2), we also have an external summary statistic $\widehat{\Xi}$ as a covariance matrix estimator of Ξ for $\widehat{\theta}^E$. In the case where $\widehat{\theta}^E$ is the sample mean $\bar{X}^E$ of $X_1^E, ..., X_m^E$, for example, $\widehat{\Xi}$ is the sample covariance matrix of $X_1^E, ..., X_m^E$. Assuming that the sample size m of the external dataset is known, we can estimate r by n/m. Matrices Λ, W, U, and L in (2.9) can all be estimated using internal data. Therefore, Σ in (2.9) can be estimated by substitution.

When the estimator $\psi(\widehat{F}_{X,Z})$ under consideration is complex, for example, a quantile of $\widehat{F}_{X,Z}$, we may apply the following bootstrap method to estimate the variance of $\psi(\widehat{F}_{X,Z})$. Independently for $b = 1, ..., B$, let (X_i^{*b}, Z_i^{*b}), $i = 1, ..., n$, be selected with replacement from (X_i, Z_i), $i = 1, ..., n$, and let $\widehat{\theta}^{E*b} \sim N(\widehat{\theta}^E, \widehat{\Xi}/m)$. Let $\psi(\widehat{F}^{*b}_{X,Z})$ be $\psi(\widehat{F}_{X,Z})$ with (X_i, Z_i)'s and $\widehat{\theta}^E$ replaced by (X_i^{*b}, Z_i^{*b})'s and $\widehat{\theta}^{E*b}$, respectively. Then, the bootstrap variance estimator for $\psi(\widehat{F}_{X,Z})$ is the sample variance of $\psi(\widehat{F}^{*b}_{X,Z})$, $b = 1, ..., B$. This bootstrap method is used in the example presented in Section 6.

3. Guaranteed Efficiency Gain

It is interesting to know whether we can construct an estimator of $F_{X,Z}$ that is almost always better than the empirical distribution $\bar{F}_{X,Z}$, i.e., utilizing external information has a guaranteed efficiency gain. The discussion in Section 2 indicates that $\widehat{F}_{X,Z}$ in (2.5) does not always achieve this, especially when $r \geq 1$, due to the uncertainty in external information.

To reach a guaranteed efficiency gain, we replace $\widehat{\theta}^E$ in (2.2) by the following shared parameter estimator that uses not only external information but also internal data,

$$\widetilde{\theta} = \frac{n\widehat{\theta} + m\widehat{\theta}^E}{n+m}, \tag{3.10}$$

where $\widehat{\theta}$ is a generalized estimation equation estimator of θ based on internal data,

$$\frac{1}{n}\sum_{i=1}^{n} u(X_i, \widehat{\theta}\,) = 0. \tag{3.11}$$

We define a new estimator of $F_{X,Z}$ as

$$\widetilde{F}_{X,Z} \text{ given by (2.5) with } \widehat{\theta}^E \text{ replaced by } \widetilde{\theta} \text{ in (3.10)-(3.11).} \tag{3.12}$$

By the same argument in Section 2 we can show that

$$\sqrt{n}(\widetilde{\mathcal{F}} - \mathcal{F}) = \sqrt{n}(\bar{\mathcal{F}} - \mathcal{F} - \tfrac{1}{r+1}\bar{u}^\top U^{-1}W - \tfrac{1}{r+1}\widetilde{u}^\top M^{-\top}L^\top U^{-1}W) + o_p(1), \tag{3.13}$$

where $\widetilde{\mathcal{F}} = \big(\widetilde{F}_{X,Z}(t_1), ..., \widetilde{F}_{X,Z}(t_s)\big)^\top$, and

$$\begin{aligned} & n\mathrm{Var}(\bar{\mathcal{F}} - \mathcal{F} - \tfrac{1}{r+1}\bar{u}^\top U^{-1}W + \tfrac{1}{r+1}\widetilde{u}^\top M^{-\top}L^\top U^{-1}W) \\ = {} & \Lambda + \tfrac{W^\top U^{-1}W}{(r+1)^2} - \tfrac{2W^\top U^{-1}W}{r+1} - \tfrac{nW^\top U^{-1}LM^{-1}VM^{-\top}L^\top U^{-1}W}{m(r+1)^2}. \end{aligned}$$

Consequently, we obtain the following result.

Theorem 2. *Under the same conditions in Theorem 1 and (3.10)-(3.11),*

$$\begin{aligned} \sqrt{n}\{\big(\widetilde{F}_{X,Z}(t_1), ..., \widetilde{F}_{X,Z}(t_s)\big)^\top - \big(F_{X,Z}(t_1), ..., F_{X,Z}(t_s)\big)^\top\} & \xrightarrow{d} N(0, \widetilde{\Sigma}), \\ \widetilde{\Sigma} = \Lambda - \tfrac{2r+1}{(r+1)^2}W^\top U^{-1}W + \tfrac{r}{(r+1)^2}W^\top U^{-1}LM^{-1}VM^{-\top}L^\top U^{-1}W, & \end{aligned} \tag{3.14}$$

as $n \to \infty$ and $m \to \infty$, for any s fixed distinct $t_1, ..., t_s$ in $\mathscr{R}^k$.

In the special case where $L = M$ and $V = U$, $\widetilde{\Sigma}$ in (3.14) is equal to $\Lambda - \frac{1}{r+1}W^\top U^{-1}W$, which means that $\widetilde{F}_{X,Z}$ is always better than the empirical distribution $\bar{F}_{X,Z}$; also, using $\widetilde{\theta}$ in (3.10)-(3.11) is better than using $\widehat{\theta}^E$ in (2.2). Note that V is the covariance matrix of $u(X^E, \theta)$ under external population whereas U is the covariance matrix of $u(X, \theta)$ under internal population. If $V = cU$ (the external $u(X^E, \theta)$ is c times as variable as the internal $u(X,\theta)$) and $L = M$, then $\widetilde{\Sigma}$ in (3.14) is $\Lambda - \frac{(2-c)r+1}{(r+1)^2} W^\top U^{-1} W$. Hence, whether $\widetilde{F}_{X,Z}$ is better than the empirical distribution $\bar{F}_{X,Z}$ depends on the size of external dataset and the variability in external data, i.e., on the sign of $(2-c)r+1$.

For large sample inference, result (3.14) can be used with $\widetilde{\Sigma}$ estimated by the same method as described in Section 2 for the estimation of Σ in (2.9).

4. Common Mean of X

In this section, we consider the case where $u(X,\theta) = X - \theta$, i.e., θ is the common mean vector shared by the internal and external populations. A specific example is the situation where X is the vector of covariates and responses under some treatments, and there are $k - l > 0$ new treatments in the study of internal population resulting in Z-data, and these treatments and data are not in the external study.

Since $\partial u(X,\theta)/\partial\theta = -I$, where I is the identity matrix, $L = M$ in (2.9) or (3.14) and the result in Theorem 1 or 2 simplifies. Further, in Theorem 1 or 2, $U = \text{Var}(X)$ and $V = \text{Var}(X^E)$, and $U = V$ if the internal and external covariance matrices of X are the same, i.e., the shared parameter is not only the mean but also the covariance matrix of X.

If we use (2.2), then $\widehat{\theta}^E = \bar{X}^E$, the sample mean of external $X_1^E, ..., X_m^E$, although individual values of $X_1^E, ..., X_m^E$ are not available. If we use $\widetilde{\theta}$ given by (3.10)-(3.11), then

$$\widetilde{\theta} = \frac{n}{n+m}\bar{X} + \frac{m}{n+m}\bar{X}^E, \tag{4.15}$$

where $\bar{X}$ is the sample mean of internal data X_i's. It can be seen in this case $\widetilde{\theta}$ is better than $\widehat{\theta}^E$.

4.1 Estimation of population means

Consider the estimation of population mean vector $\mu = \int t dF_{X,Z}(t)$. The first l components of μ is the shared parameter vector θ.

With $F_{X,Z}$ estimated by $\widehat{F}_{X,Z}$ in (2.5) and $\widehat{\theta}^E = \bar{X}^E$, μ is estimated by

$$\widehat{\mu} = \int t d\widehat{F}_{X,Z}(t) = \sum_{i=1}^{n} \widehat{p}_i \begin{pmatrix} X_i \\ Z_i \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+\lambda^\top(X_i - \bar{X}^E)} \begin{pmatrix} X_i \\ Z_i \end{pmatrix}.$$

From (2.6), with probability tending to 1,

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i - \bar{X}^E}{1+\lambda^\top(X_i - \bar{X}^E)} = \sum_{i=1}^{n} \widehat{p}_i X_i - \sum_{i=1}^{n} \widehat{p}_i \bar{X}^E = \sum_{i=1}^{n} \widehat{p}_i X_i - \bar{X}^E,$$

since $\sum_{i=1}^{n} \widehat{p}_i = 1$. This means that the first l components of $\widehat{\mu}$ is $\bar{X}^E$, a function of external data only, which may be fine if the external sample size m is much larger than the internal sample size n, but is not very good otherwise.

With $F_{X,Z}$ estimated by $\widetilde{F}_{X,Z}$ in (3.12) with $\widetilde{\theta}$ given by (3.10)-(3.11), i.e., (4.15), μ is estimated by

$$\widetilde{\mu} = \int t d\widetilde{F}_{X,Z}(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+\lambda^\top(X_i - \widetilde{\theta})}\binom{X_i}{Z_i},$$

and the first l components of $\widetilde{\mu}$ is $\widetilde{\theta}$, which is a more reasonable estimator of the first l components of μ, the shared parameter vector, especially when m is not much larger than n. This supports the use of $\widetilde{\theta}$ in Section 3 to replace $\widehat{\theta}^E$ in Section 2.

The following result for the asymptotic normality of $\widehat{\mu}$ and $\widetilde{\mu}$ can be shown using the same argument in the derivation of Theorem 1 or 2, or applying the mean functional to the result in Theorem 1 or 2.

Corollary 1. *Assume the conditions in Theorem 2 with $u(X, \theta) = X - \theta$ and finiteness of the second-order moments of (X, Z). Then,*

$$\sqrt{n}(\widehat{\mu} - \mu) \xrightarrow{d} N\left(0,\ \mathrm{Var}\binom{X}{Z} - H^\top U^{-1} H + r H^\top U^{-1} V U^{-1} H\right), \qquad (4.16)$$

$$\sqrt{n}(\widetilde{\mu} - \mu) \xrightarrow{d} N\left(0,\ \mathrm{Var}\binom{X}{Z} - \tfrac{2r+1}{(r+1)^2} H^\top U^{-1} H + \tfrac{r}{(r+1)^2} H^\top U^{-1} V U^{-1} H\right), \qquad (4.17)$$

where r is the limit of n/m and $H = E\left\{(X - \theta)(X^\top, Z^\top)\right\}$ under the internal population.

More details about the asymptotic covariance matrices in (4.16)-(4.17) can be obtained. Let $D = \mathrm{Var}(Z)$ and $C = \mathrm{Cov}(X, Z)$. Then

$$\mathrm{Var}(T) = \begin{pmatrix} U & C \\ C^\top & D \end{pmatrix}, \qquad H = \begin{pmatrix} U & C \end{pmatrix},$$

and, hence, the asymptotic covariance matrix of $\sqrt{n}(\widehat{\mu} - \mu)$ in (4.16) is

$$\begin{pmatrix} rV & rVU^{-1}C \\ rC^\top U^{-1}V & D - C^\top U^{-1}C + rC^\top U^{-1}VU^{-1}C \end{pmatrix},$$

and the asymptotic covariance matrix of $\sqrt{n}(\widetilde{\mu} - \mu)$ in (4.17) is

$$\begin{pmatrix} \frac{r^2}{(r+1)^2}U + \frac{r}{(r+1)^2}V & \frac{r^2}{(r+1)^2}C + \frac{r}{(r+1)^2}VU^{-1}C \\ \frac{r^2}{(r+1)^2}C^\top + \frac{r}{(r+1)^2}C^\top U^{-1}V & D - \frac{2r+1}{(r+1)^2}C^\top U^{-1}C + \frac{r}{(r+1)^2}C^\top U^{-1}VU^{-1}C \end{pmatrix}.$$

For the estimation of the first l components of μ (the mean of X), the comparison between $\widehat{\mu}$ and $\widetilde{\mu}$ is actually the comparison between $\widehat{\theta}^E = \bar{X}^E$ and $\widetilde{\theta}$ in (4.15). If $r = 0$, then the result shows that the convergence rate of $\widehat{\mu}$ and $\widetilde{\mu}$ is faster than $1/\sqrt{n}$.

For estimating the last $k - l$ components of μ, the convergence rate is $1/\sqrt{n}$ even if $r = 0$. The comparison between $\widehat{\mu}$ and $\widetilde{\mu}$ is between two matrices $D - C^\top U^{-1}C + rC^\top U^{-1}VU^{-1}C$ and $D - \frac{2r+1}{(r+1)^2}C^\top U^{-1}C + \frac{r}{(r+1)^2}C^\top U^{-1}VU^{-1}C$, which is similar to the comparison between Σ in (2.9) and $\widetilde{\Sigma}$ in (3.14). In particular, if $V = U$, then the former is $D - (1 - r)C^\top U^{-1}C$ but the latter is $D - \frac{1}{r+1}C^\top U^{-1}C$ and, hence, $\widetilde{\mu}$ is always better than $\widehat{\mu}$ asymptotically.

4.2 Estimation of population quantiles

Consider the estimation of quantile vector $Q = \big(F^{-1}(\pi_1), ..., F^{-1}(\pi_s)\big)^\top$, where F is a particular marginal of $F_{X,Z}$, $\pi_1, ..., \pi_s$ are s distinct known points in $(0,1)$, and $F^{-1}(\pi) = \inf\{t : F(t) \geq \pi\}$. If we do not use any external information, then a customary estimator is the vector of sample quantiles. If we estimate $F_{X,Z}$ by $\widehat{F}_{X,Z}$ or $\widetilde{F}_{X,Z}$, then our estimator is $\widehat{Q} = \big(\widehat{F}^{-1}(\pi_1), ..., \widehat{F}^{-1}(\pi_s)\big)^\top$ or $\widetilde{Q} = \big(\widetilde{F}^{-1}(\pi_1), ..., \widetilde{F}^{-1}(\pi_s)\big)^\top$, where $\widehat{F}$ and $\widetilde{F}$ are the corresponding marginals of $\widehat{F}_{X,Z}$ and $\widetilde{F}_{X,Z}$, respectively. Using the same argument in the proof of Bahadur's representation (see, e.g., Theorem 5.11 of Shao (2003)), we can show that

$$\sqrt{n}(\widehat{Q} - Q) = \sqrt{n}\left(\tfrac{F(F^{-1}(\pi_1)) - \widehat{F}(F^{-1}(\pi_1))}{f(F^{-1}(\pi_1))}, ..., \tfrac{F(F^{-1}(\pi_s)) - \widehat{F}(F^{-1}(\pi_s))}{f(F^{-1}(\pi_s))}\right)^\top + o_p(1),$$

where $f(F^{-1}(\pi_j))$ is the derivative of F at $F^{-1}(\pi_j)$ assumed to be positive, $j = 1, ..., s$. The same result holds with $(\widehat{F}, \widehat{Q}\,)$ replaced by $(\widetilde{F}, \widetilde{Q}\,)$. This representation together with result (2.9) or (3.14) show that $\widehat{Q}$ or $\widetilde{Q}$ is typically asymptotically more efficient than the sample quantile vector without using external information.

For inference about quantiles, the bootstrap method introduced in the end of Section 2 can be applied.

5. Simulation

In this section we present some simulation results under the scenario in Section 4. Consider $k = 3$, $l = 2$, a two-dimensional X and a univariate Z, i.e., $u(X, \theta) = X - \theta$, and the mean of X, $\theta = E(X)$, is the two-dimensional shared parameter vector for internal and external populations.

5.1 Simulation with a continuous Z

We consider the following four cases for internal and external populations.

A. For both internal and external populations, X is bivariate normal,

$$X \sim N\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}\right). \tag{5.18}$$

For internal population, conditional on X, Z is normal with mean $\alpha + \beta^\top X$ and variance 0.25. For external population, conditional on X, Z is normal with the same mean as in the internal population but a different variance = 1.

B. For internal and external populations, X is generated according to (5.18). For internal population, conditional on X, Z is the same as that in case A. For external population, conditional on X, Z has the double exponential distribution with mean $\alpha + \beta^\top X$ and scale parameter 0.5.

C. In internal population, X is generated according to (5.18). For external population, X is generated according to (5.18) but with the covariance matrix replaced by $\begin{pmatrix} 2 & 0.5 \\ 0.5 & 1 \end{pmatrix}$. Conditional on X, Z is generated the same as in case B, for both internal and external populations.

D. X is generated the same as in case C. For internal population, conditional on X, $Z - \alpha - \beta^\top X$ has a probability density $f(t)$ that is the normal density with mean 0 and variance 0.25 when $t < 0$, and is the double exponential density with mean 0 and scale parameter 0.5 when $t \geq 0$. For external population, conditional on X, $Z - \alpha - \beta^\top X$ has probability density $f(-t)$.

In all cases, $(\alpha, \beta^\top) = (1.5, 0.4, -0.8)$. In cases A and B, the internal and external populations of X are the same, whereas in cases C and D, the internal and external populations of X are different although they share the same mean $E(X)$. Conditional on X, the internal and external distributions of Z are always different, normal distributions with difference variances in case A, normal versus double exponential distributions in cases B and C, and asymmetric distributions in case D.

We consider the estimation of two parameters in internal population, the mean $E(Z)$ and Q_{75} = the 0.75 quantile of Z, with internal sample size $n = 100$ and external sample size $m = 100$, 200, 500, 1000, and 10000, which ranges from comparable with n to much larger than n.

Based on 2000 simulation runs, Table 1 presents the simulation bias and standard deviation (SD) of the following estimators.

1. $\bar{Z}$ = the sample mean and $\bar{Q}_{75}$ = sample 0.75 quantile, based on internal Z data without using any external information.

2. $\widetilde{Z}$ = the mean and $\widetilde{Q}_{75}$ = the 0.75 quantile of the third marginal of $\widetilde{F}_{X,Z}$ in (3.12) (the estimated distribution of Z), where $\widetilde{\theta} = \dfrac{n}{n+m}\bar{X} + \dfrac{m}{n+m}\bar{X}^E$, $\bar{X}$ is the sample mean for internal data, and $\bar{X}^E$ is the external summary statistic, the sample mean for external X-data.

3. $\widehat{Z}$ = the mean and $\widehat{Q}_{75}$ = the 0.75 quantile of the third marginal of $\widehat{F}_{X,Z}$ in (2.5), where $\widehat{\theta} = \bar{X}^E$ is the external summary statistic.

The following is a summary of the results in Table 1 based on 2000 simulations.

1. All biases are negligible, even for the case with $m = n = 100$. For $\bar{Z}$ and $\bar{Q}_{75}$ without using external information, the bias and SD have small variations within each setting due to simulation error.

2. For the estimation of mean $E(Z)$, the efficiency gain of using external summary information is very substantial, when $\widetilde{\theta}$ in (3.10) is used.

The efficiency gain ranges from 18% to 43% when m ranges from 100 to 10^4 for cases A-C. The efficiency gain is slightly smaller for case D when the distribution of Z is asymmetric. When $\widehat{\theta}^E$ in (2.2) is used, the efficiency gain of using external summary information is negligible when $m = 100$, becomes appreciable when $m = 200$, and is comparable with the use of $\widetilde{\theta}$ when $m \geq 1000$.

3. Although the shared parameter θ is the mean vector (of X), the quantile estimation also has substantial gain when external summary information is utilized. The efficiency gain for 0.75 quantile estimation can still be between 10-20% for cases A-C and 7-17% for case D, when $\widetilde{\theta}$ in (3.10) is used. When $\widehat{\theta}^E$ in (2.2) is used, we need $m \geq 500$ in order to see substantial efficiency gain for quantile estimation.

5.2 Sensitivity of assumption (2.1)

Assumption (2.1) of shared parameter is a bridge between internal and external data for utilizing external information. If (2.1) is violated, then our proposed estimators may be biased. Here, we perform a simulation under case A of Section 5.1 to see the sensitivity of (2.1). Specifically, we add a positive constant δ to the two mean values of X in (5.18) for the external population; that is, the internal X has mean vector $(1, 0)$ but external X^E has mean vector $(1-\delta,-\delta)$. The remaining parameters are unchanged.

Figures 1 and 2 present the root of mean squared error (RMSE) of mean and quantile estimation based on the three different estimators considered in Section 5.1, with δ varying from 0 to 0.25. As expected, when δ increases, the RMSE of the estimator using internal data only keeps stable; the RMSE of two estimators using external information increases. It can be seen that the proposed method using $\widetilde{\theta}$ still has much smaller RMSEs than the one using internal data only unless $\delta > 0.2$ and $m = 500$, and using $\widetilde{\theta}$ is better than using $\widehat{\theta}$.

5.3 Simulation with a binary Z

We consider a binary Z with X being a 15-dimension multivariate normally distributed vector having mean $(1, 0.5, 0, ..., 0)^\top$ and covariance matrix whose diagonal elements are equal to 1 and off-diagonal elements are equal to 0.3, for both internal and external populations. Conditional on X, Z follows a Bernoulli distribution with probability π satisfying $\log \pi(1-\pi) = \alpha + \beta^T X$, where $(\alpha, \beta^T) = (-0.5, 1, -1, 0.5, ..., 0.5)$ for the internal population and $(\alpha, \beta^T) = (-0.3, 1, -1, 0.5, ..., 0.5)$ for the internal population. For this binary Z, we consider the estimation of mean $E(Z)$ of the internal population with internal sample size $n = 100$ and external sample size $m = 100, 200, 500, 1,000$, and $10,000$. Based on 2000 simulation runs, Table 2 presents the simulation bias and standard deviation (SD) of the estimators $\bar{Z}$, $\widetilde{Z}$, and $\widehat{Z}$ as defined in Section 5.1. Similar conclusions to those in Section 5.1 can be obtained from the simulation results in Table 2

.

6. An Example

An important part of agriculture around the world, particularly in Turkey, is about dry beans. The Turkish Standards Institution classifies dry beans according to their physical features that can help farmers to identify dry beans and monitor their quality. Two physical features are the major axis length X, which is the length of longest straight line that can be drawn from a bean, and the area Z of a bean. We consider a dataset available on the website https://www.muratkoklu.com/datasets/ with two type of dry bean, BARBUNYA and HOROZ. The dataset for BARBUNYA is used as the internal dataset and the dataset for HOROZ is treated as the external dataset. Figure 3 shows the boxplots of areas and major axis lengths of two dry beans, and Table 3 provides some basic statistics. It can be seen that these two types of dry beans share almost the same major axis length, but their areas differ greatly. We would like to estimate the mean, 0.25, 0.5, and 0.75 quantiles of the area of BARBUNYA using information from major axis lengths of BARBUNYA and HOROZ to improve efficiency.

For the mean $E(Z)$, quantiles Q_{25}, Q_{50}, and Q_{75} of area of BARBUNYA, we compute the following three types of estimates.

1. The sample mean and quantiles based on only internal data for dry bean BARBUNYA.

2. The mean and quantiles of last marginal of $\widetilde{F}_{X,Z}$ in (3.12) using $\widetilde{\theta} = \dfrac{n}{n+m}\bar{X} + \dfrac{m}{n+m}\bar{X}^E$ with $\bar{X}$ = the sample mean of major axis length for internal data, $\bar{X}^E$ = the sample mean of major axis length for external data (HOROZ), $n = 1322$, and $m = 1928$.

3. The mean and quantiles of last marginal of $\widehat{F}_{X,Z}$ in (2.5) using $\widehat{\theta}^E = \bar{X}^E$.

For each point estimate, we compute the bootstrap standard error as the square root of the bootstrap variance estimator described in the end of Section 2 with $B = 2000$. The results are given in Table 4.

From Table 4, the point estimates for the same parameter are close to each other. The estimates using external information with $\widetilde{F}_{X,Z}$ in (3.12) and $\widetilde{\theta}$ in (3.10) have smaller standard errors than those using internal data only; the relative efficiency to internal data only in terms of standard error is substantial for the estimation of mean, Q_{50}, and Q_{75}, and is slight for the estimation of Q_{25}. Comparing two methods of using external information, we find that the method using $\widetilde{F}_{X,Z}$ in (3.12) is much better than the method using $\widehat{F}_{X,Z}$ in (2.5); in fact the method using $\widehat{F}_{X,Z}$ in (2.5) is comparable with the method without using external information as $r = 0.686$ is not very small. Thus, using external information is worthwhile especially for the method using $\widetilde{F}_{X,Z}$ in (3.12).

7. Discussion

Using external summary information, we improve the nonparametric empirical distribution $\bar{F}_{X,Z}$ based on internal data only by the nonparametric empirical likelihood estimator $\widetilde{F}_{X,Z}$ in (3.12) with $\widetilde{\theta}$ given by (3.10) using both internal estimator $\widehat{\theta}$ and external estimator $\widehat{\theta}^E$ of the shared parameter θ defined by (2.1).

If $\bar{F}_{X,Z}$ is replaced by the semi-parametric empirical likelihood estimator that maximizes

$$\prod_{i=1}^{n} p_i \quad \text{subject to} \quad p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i g(X_i, \beta) = 0$$

over p_i's and β, where g is a known vector function and β is an unknown parameter vector with dimension smaller than the dimension of g, then our method can be extended to $\widetilde{F}_{X,Z}$ that maximizes

$$\prod_{i=1}^{n} p_i \quad \text{subject to} \quad p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i g(X_i, \beta) = \sum_{i=1}^{n} p_i u(X_i, \widetilde{\theta}) = 0$$

over p_i's and β. Properties of this $\widetilde{F}_{X,Z}$ can be derived similarly.

An anonymous referee suggests an alternative to $\widetilde{F}_{X,Z}$ in (3.12); that is, $\check{F}_{X,Z}$ putting mass

$$\check{p}_i = \frac{n}{n+m}\widehat{p}_{EL,i} + \frac{m}{n+m}\widehat{p}_i$$

to each (X_i, Z_i), $i = 1, ..., n$, where $\widehat{p}_i$ is given in (2.5) and $\widehat{p}_{EL,i}$ is the mass of nonparametric empirical likelihood estimator without using any external information. Because the dimensions of u and θ are the same, $\widehat{p}_{EL,i} = n^{-1}$ for all i and, hence,

$$\check{F}_{X,Z} = \frac{n}{n+m}\bar{F}_{X,Z} + \frac{m}{n+m}\widehat{F}_{X,Z}.$$

Define $\check{\mathcal{F}} = \left(\check{F}_{X,Z}(t_1), ..., \check{F}_{X,Z}(t_s)\right)^\top$. Following the notation in Sections 2-3, we have

$$\begin{aligned}\sqrt{n}(\check{\mathcal{F}} - \mathcal{F}) &= \sqrt{n}(\bar{\mathcal{F}} - \mathcal{F}) + \sqrt{n}\frac{m}{n+m}(\widehat{\mathcal{F}} - \bar{\mathcal{F}}) \\ &= \sqrt{n}(\bar{\mathcal{F}} - \mathcal{F}) - \frac{\sqrt{n}}{r+1}\bar{u}^\top U^{-1}W - \frac{\sqrt{n}}{r+1}\bar{u}^\top U^{-1}W + o_p(1) \\ &= \sqrt{n}(\widetilde{\mathcal{F}} - \mathcal{F}) + o_p(1),\end{aligned}$$

where the second equality follows from (2.7) and $r =$ the limit of n/m, and the last equality follows from (3.13). Therefore, $\widetilde{F}_{X,Z}$ and $\check{F}_{X,Z}$ are asymptotically equivalent. This asymptotic result is confirmed by simulation not reported in this paper.

Acknowledgements

We would like to thank two anonymous referees for helpful comments and suggestions. Lyu Ni's research was partially supported by the Shanghai Sailing Program (22YF1411300) and the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East ChinaNormal University), Ministry of Education. Jun Shao's research was partially supported by the National Natural Science Foundation of China (11831008) and the U.S. National Science Foundation (DMS-1914411). Lei Wang's research was supported by the Fundamental Research Funds for the Central Universities and the National Natural Science Foundation of China (12271272).

References

Chatterjee, N., Y. H. Chen, P. Maas, and R. J. Carroll (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.

Kundu, P., R. Tang, and N. Charterjee (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika 106*(3), 567–585.

Li, S., T. T. Cai, and H. Li (2022). Estimation and inference with proxy data and its genetic applications. *arXiv preprint arXiv:2201.03727*.

Lin, D. and D. Zeng (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika 97*(2), 321–332.

Lohr, S. L. and T. E. Raghunathan (2017). Combining survey data with other data sources. *Statistical Science 32*(2), 293–312.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association 99*(468), 1131–1139.

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

Owen, A. B. (2001). *Empirical Likelihood.* Chapman and Hall/CRC Press, New York.

Qin, J. and J. Lawless (1994, 03). Empirical likelihood and general estimating equations. *The Annals of Statistics 22*(1), 300–325.

Rao, J. N. K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B 83*(1), 242–272.

Shao, J. (2003). *Mathematical Statistics* (2nd ed.). Springer, New York.

Tian, Y. and Y. Feng (2022). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association, to appear.*.

Wang, Z., H. J. Kim, and J. K. Kim (2023). Survey data integration for regression analysis using model calibration. *Survey Methodology, to appear*.

Yang, S., C. Gao, D. Zeng, and X. Wang (2023). Elastic integrative analysis of randomized trial and real-world data for treatment heterogeneity estimation. *Journal of Royal Statistical Society, Series B: Statistical Methodology 85*, 575–596.

Yang, S. and J. K. Kim (2020). Statistical data integration in survey sampling: a review. *Japanese Journal of Statistics and Data Science 3*(2), 625–650.

Zhang, H., L. Deng, M. Schiffman, J. Qin, and K. Yu (2020). Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika 107*(3), 689–703.

Zhang, Y., Z. Ouyang, and H. Zhao (2017). A statistical framework for data integration through graphical models with application to cancer genomics. *The Annals of Applied Statistics 11*(1), 161–184.

Lyu Ni

School of Data Science and Engineering & KLATASDS-MOE, East China Normal University, Shanghai, China

E-mail: lni@dase.ecnu.edu.cn

Jun Shao

Department of Statistics, University of Wisconsin-Madison, Madison, WI, U.S.A

E-mail: jshao@wisc.edu

Jinyi Wang

Department of Statistics, University of Wisconsin-Madison, Madison, WI, U.S.A

E-mail: jwang2242@wisc.edu

Lei Wang

School of Statistics and Data Science, KLMDASR, LEBPS and LPMC,
Nankai University, Tianjin, China.

E-mail: lwangstat@nankai.edu.cn

Table 1: Simulation results (2000 replications) for estimation of mean $E(Z)$ and 0.75 quantile Q_{75} when internal sample size $n = 100$

Case A		estimation of $E(Z) = 1.90$			estimation of $Q_{75} = 2.52$		
		$\bar{Z}$	$\widetilde{Z}$	$\widehat{Z}$	$\bar{Q}_{75}$	$\widetilde{Q}_{75}$	$\widehat{Q}_{75}$
$m = 100$	bias	0.0021	-0.0009	-0.0040	0.0064	0.0037	0.0038
	SD	0.0941	0.0758	0.0917	0.1283	0.1158	0.1264
$m = 200$	bias	0.0017	-0.0002	-0.0011	-0.0018	-0.0031	-0.0026
	SD	0.0917	0.0669	0.0744	0.1244	0.1091	0.1152
$m = 500$	bias	0.0008	0.0014	0.0016	0.0006	0.0029	0.0036
	SD	0.0952	0.0601	0.0617	0.1272	0.1056	0.1059
$m = 10^3$	bias	-0.0018	-0.0013	-0.0013	-0.0045	-0.0022	-0.0016
	SD	0.0901	0.0549	0.0557	0.1240	0.1056	0.1059
$m = 10^4$	bias	0.0028	-0.0009	-0.0010	0.0014	-0.0014	-0.0012
	SD	0.0938	0.0516	0.0516	0.1277	0.1012	0.1011
Case B		estimation of $E(Z) = 1.90$			estimation of $Q_{75} = 2.52$		
		$\bar{Z}$	$\widetilde{Z}$	$\widehat{Z}$	$\bar{Q}_{75}$	$\widetilde{Q}_{75}$	$\widehat{Q}_{75}$
$m = 100$	bias	0.0001	-0.0008	-0.0019	-0.0008	-0.0004	0.0016
	SD	0.0914	0.0747	0.0932	0.1249	0.1124	0.1249
$m = 200$	bias	-0.0001	0.0016	0.0024	-0.0026	0.0006	0.0019
	SD	0.0896	0.0661	0.0738	0.1239	0.1091	0.1143
$m = 500$	bias	-0.0008	-0.0012	-0.0013	0.0003	0.0025	0.0026
	SD	0.0934	0.0605	0.0622	0.1242	0.1035	0.1047
$m = 10^3$	bias	-0.0009	0.0002	0.0003	-0.0010	0.0015	0.0020
	SD	0.0903	0.0546	0.0554	0.1256	0.1045	0.1047
$m = 10^4$	bias	-0.0018	-0.0017	-0.0017	-0.0037	-0.0007	-0.0007
	SD	0.0938	0.0526	0.0526	0.1266	0.1026	0.1026

Table 1: continued

Case C			estimation of $E(Z) = 1.90$			estimation of $Q_{75} = 2.52$		
			$\bar{Z}$	$\widetilde{Z}$	$\widehat{Z}$	$\bar{Q}_{75}$	$\widetilde{Q}_{75}$	$\widehat{Q}_{75}$
	$m = 100$	bias	-0.0017	-0.0017	-0.0019	-0.0011	-0.0011	0.0015
		SD	0.0911	0.0751	0.0955	0.1240	0.1122	0.1286
	$m = 200$	bias	0.0018	0.0025	0.0029	-0.0004	0.0011	0.0037
		SD	0.0919	0.0678	0.0750	0.1241	0.1091	0.1157
	$m = 500$	bias	0.0001	-0.0010	-0.0012	-0.0051	-0.0037	-0.0031
		SD	0.0934	0.0606	0.0629	0.1267	0.1047	0.1059
	$m = 10^3$	bias	0.0011	0.0005	0.0004	-0.0002	0.0022	0.0024
		SD	0.0914	0.0547	0.0555	0.1214	0.1008	0.1006
	$m = 10^4$	bias	0.0007	-0.0019	-0.0019	0.0018	0.0025	0.0026
		SD	0.0923	0.0525	0.0526	0.1230	0.1003	0.1003
Case D			estimation of $E(Z) = 1.95$			estimation of $Q_{75} = 2.58$		
			$\bar{Z}$	$\widetilde{Z}$	$\widehat{Z}$	$\bar{Q}_{75}$	$\widetilde{Q}_{75}$	$\widehat{Q}_{75}$
	$m = 100$	bias	0.0005	0.0010	0.0016	-0.0031	-0.0011	0.0029
		SD	0.0975	0.0816	0.1017	0.1342	0.1250	0.1417
	$m = 200$	bias	-0.0023	-0.0006	0.0002	-0.0075	-0.0050	-0.0008
		SD	0.0991	0.0749	0.0810	0.1345	0.1222	0.1285
	$m = 500$	bias	0.0002	0.0016	0.0018	-0.0021	0.0014	0.0032
		SD	0.0995	0.0681	0.0693	0.1357	0.1152	0.1167
	$m = 10^3$	bias	0.0015	0.0013	0.0013	0.0006	0.0030	0.0030
		SD	0.0992	0.0653	0.0660	0.1367	0.1170	0.1184
	$m = 10^4$	bias	0.0052	0.0015	0.0014	0.0053	0.0031	0.0030
		SD	0.1000	0.0614	0.0614	0.1347	0.1118	0.1119

Figure 1: The RMSE values different δ on the estimation of mean under $m = 100$, 200 and 500.

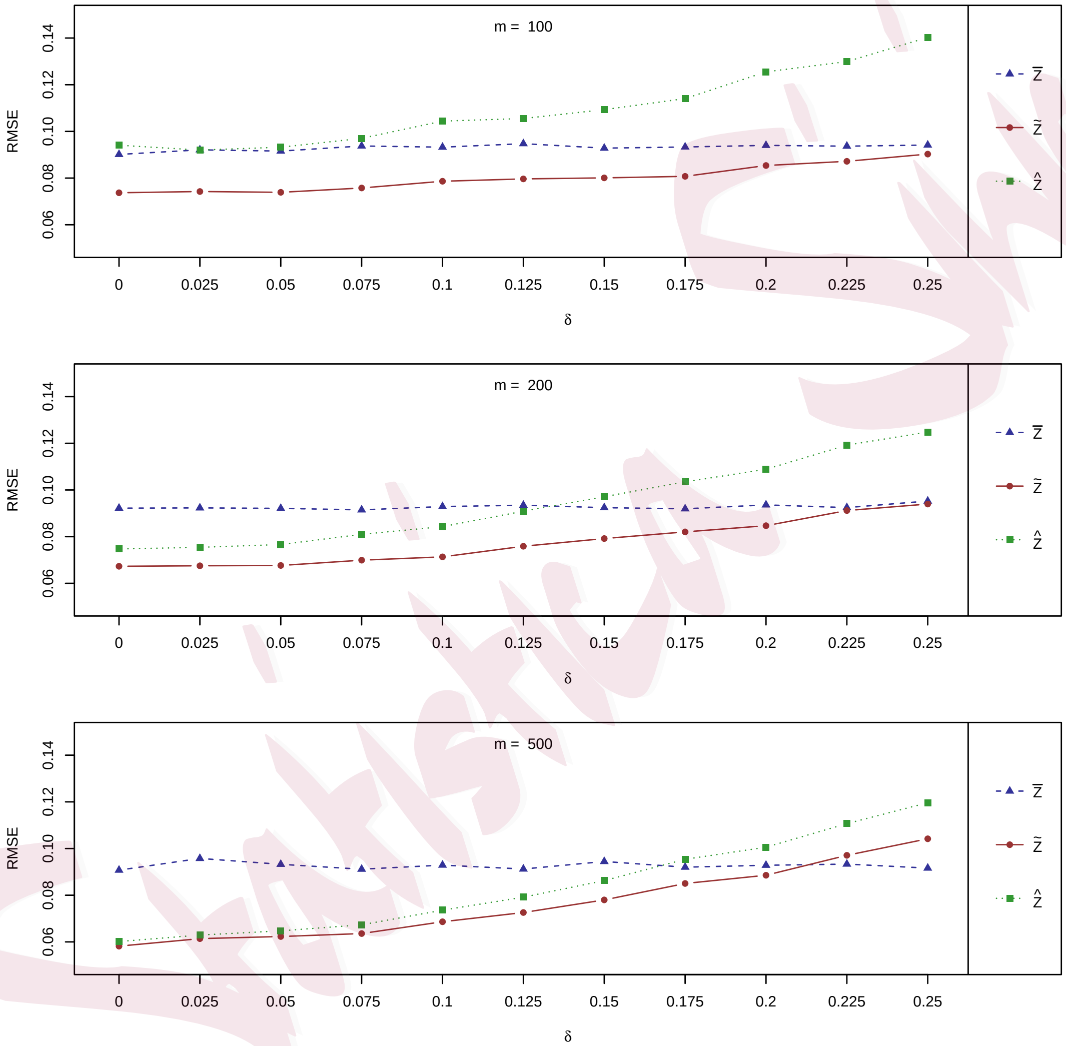

Figure 2: The RMSE values different δ on the estimation of quantile under $m = 100$, 200 and 500.

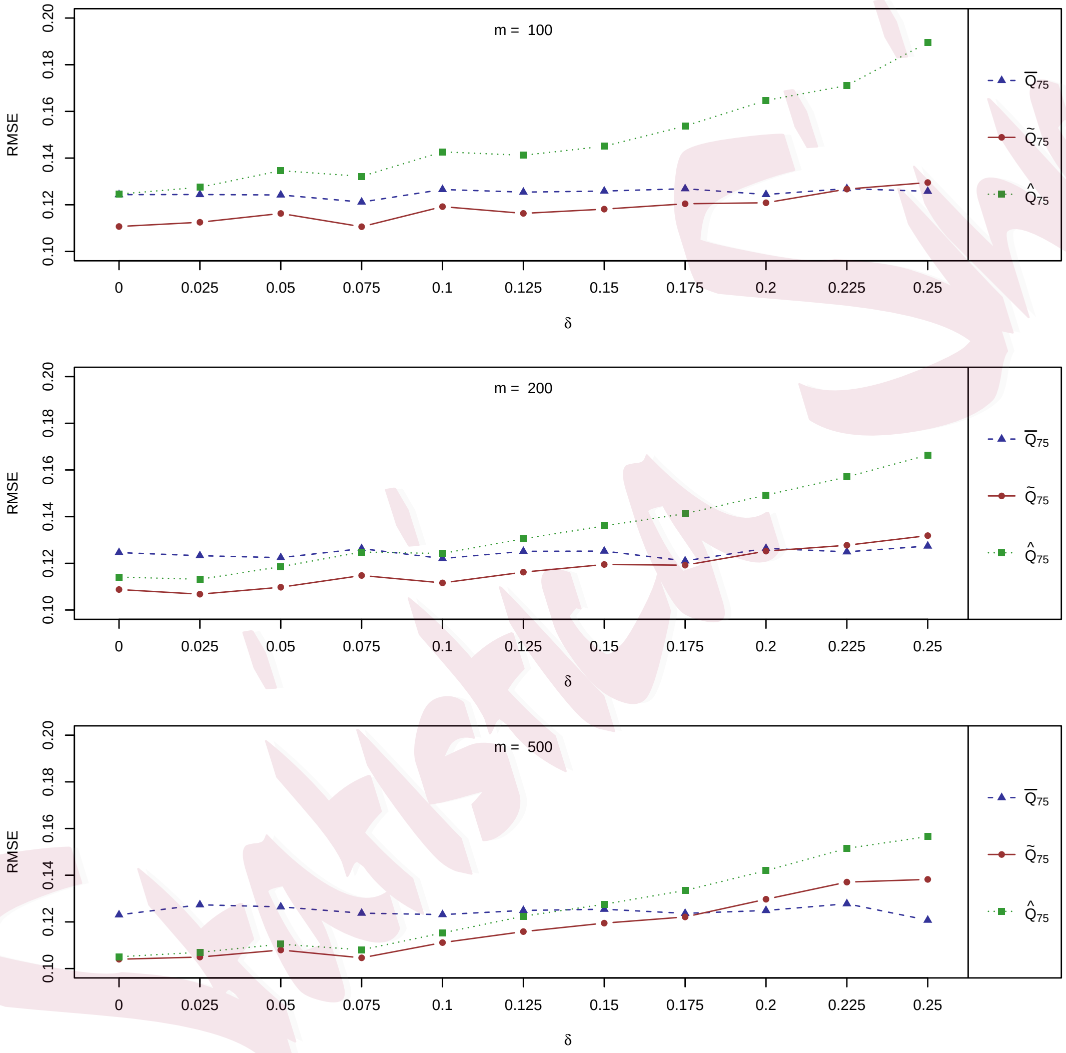

Table 2: Simulation results (2000 replications) for estimation of mean $E(Z)$ for a Bernoulli variable Z

		estimation of $E(Z) = 0.5$		
		$\bar{Z}$	$\widetilde{Z}$	$\widehat{Z}$
$m = 100$	bias	0.0016	0.0014	0.0016
	SD	0.0511	0.0448	0.0567
$m = 200$	bias	0.0008	-0.0005	-0.0012
	SD	0.0484	0.0425	0.0494
$m = 500$	bias	-0.0022	-0.0005	-0.0002
	SD	0.0500	0.0410	0.0432
$m = 10^3$	bias	0.0001	-0.0010	-0.0012
	SD	0.0497	0.0389	0.0390
$m = 10^4$	bias	0.0020	0.0006	0.0006
	SD	0.0504	0.0380	0.0381

Figure 3: The boxplots of two types of dry beans

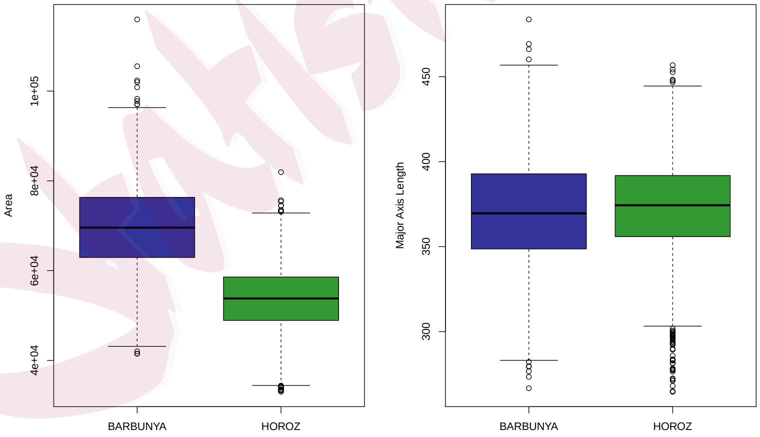

Table 3: The basic information of two dry bean datasets

	BARBUNYA	HOROZ
Sample size	1322	1928
Sample mean of X (length)	370.0	372.6
Sample SD of X	32.3	30.2
Sample mean of Z (area)	69804.1	53648.5
Sample SD of Z	10265.4	7341.4
Correlation coefficient of Z and X	0.88	0.91

Table 4: The estimates of mean and quantiles for the area of BARBUNYA with bootstrap SE

Estimate	$E(Z)$	Q_{25}	Q_{50}	Q_{75}
Using internal data only	69804.1	62930.0	69580.0	76307.0
Bootstrap SE	286.7	386.1	323.9	383.7
Using $\widetilde{F}_{X,Z}$ in (3.12) and $\widetilde{\theta}$ in (3.10)	70224.1	63357.0	69963.0	76729.0
Bootstrap SE	215.4	380.2	258.8	322.8
Relative efficiency to internal only	1.3310	1.0155	1.2515	1.1887
Using $\widehat{F}_{X,Z}$ in (2.5) and $\widehat{\theta}^E$ in (2.3)	70512.7	63733.0	70135.0	76904.0
Bootstrap SE	264.4	383.4	313.7	391.6
Relative efficiency to internal only	1.0843	1.0070	1.0325	0.9798