Statistica Sinica **35** (2025), 1-19
doi:https://doi.org/10.5705/ss.202023.0056

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 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 $\{X_i, Z_i, i = 1, \ldots, n\}$ 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 $\{X_i^E, i = 1, \ldots, m\}$, independent of $\{X_i, Z_i, i = 1, \ldots, n\}$, 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

*Corresponding author. E-mail: lwangstat@nankai.edu.cn