Statistica Sinica **34** (2024), 1-23
doi:https://doi.org/10.5705/ss.202021.0446

KERNEL REGRESSION UTILIZING EXTERNAL INFORMATION AS CONSTRAINTS

Chi-Shian Dai and Jun Shao

University of Wisconsin-Madison and East China Normal University

Abstract: With advancements in data collection and storage technology, data analysis in modern scientific research and practice has shifted from analyzing single data sets to coupling several data sets. Here, we consider a nonparametric kernel regression in an internal data set analysis, using constraints for auxiliary information from an external data set with summary statistics. Under several conditions, we show that the proposed constrained kernel regression estimator is asymptotically normal, and outperforms the standard kernel regression without external information in terms of the asymptotic mean integrated square error. Furthermore, we consider the situation in which the internal and external data have different populations. Simulation results confirm our theory and quantify the improvements from using external data. Lastly, we demonstrate the proposed method using a real-data example.

Key words and phrases: Asymptotic mean integrated square error, constraints, data integration, external summary statistics, two-step kernel regression.

1. Introduction

With advancements in data collection and storage technology, many modern statistical analyses have access to both primary individual-level data and information from independent external data sets, which typically may be large, but often contain relatively crude information, such as summary statistics, owing to practical and ethical reasons. Sources of external data sets include those from a population-based census, administrative data sets, and databases from past investigations. In what follows, primary individual-level data are referred to as internal data. An internal data set addresses specific scientific questions, and so may contain additional measured covariates from each sampled subject and, consequently, is much smaller than external data sets, owing to cost considerations. Thus, there is a growing need for internal data analysis that also uses summary information from external data sets. This line of research fits into a more general framework of data integration (Kim, Wang and Kim (2021); Lohr and Raghunathan (2017); Merkouris (2004); Rao (2021); Yang and Kim (2020); Zhang, Ouyang and Zhao (2017); Zieschang (1990)), and differs from traditional meta-analysis, which is based on multiple data sets with summary

*Corresponding author.