Statistica Sinica **36** (2026), 1-26
doi:https://doi.org/10.5705/ss.202023.0302

INFORMATION-BASED OPTIMAL SUBDATA SELECTION FOR CLUSTERWISE LINEAR REGRESSION

Yanxi Liu, John Stufken and Min Yang*

AbbVie Inc., George Mason University and University of Illinois at Chicago

Abstract: Mixture-of-Experts (MoE) models are commonly used when there exist distinct clusters with different relationships between the independent and dependent variables. Fitting such models for large datasets, however, is computationally virtually impossible. An attractive alternative is to use a subdata selected by "maximizing" the Fisher information matrix. A major challenge is that no closed-form expression for the Fisher information matrix is available for such models. Focusing on clusterwise linear regression models, a subclass of MoE models, we develop a framework that overcomes this challenge. We prove that the proposed subdata selection approach is asymptotically optimal, i.e., no other method is statistically more efficient than the proposed one when the full data size is large.

Key words and phrases: D-optimality, information matrix, latent indicator, massive data, MLE.

1. Introduction

Modern information technologies, such as cloud computing, The Internet of Things, and the social networking, are drivers for exponential growth of the size of datasets. Size may now be measured by TB or even PB instead of MB and GB (Cai and Zhu, 2015). While the extraordinary amount of data offers unprecedented opportunities for scientific discoveries and advancement, it also poses unprecedented challenges for analysis. These challenges are typically amplified by the complexity of the data and the speed with which it must be analyzed. A critical question for the statistics community is how to detect statistical relationships within high volumes of data with a complicated structure and turn it into actionable knowledge (Bühlmann et al., 2016).

With large datasets, relationships between input and output variables may no longer be homogeneous. Linear models and generalized linear models, which are effective when relationships are homogeneous, may be inadequate in the era of big data. One strategy for dealing with heterogeneity is through Mixture-of-Experts (MoE) models. The rationale for MoE models is to uncover hidden clusters within the data, such that within each cluster relationships between input and output variables can be adequately modeled by a single regression or classification model. While any such regression or classification model may be inadequate for

*Corresponding author. E-mail: minyang.stat@gmail.com