Statistica Sinica **33** (2023), 1-23
doi:https://doi.org/10.5705/ss.202020.0361

DATA INTEGRATION IN HIGH DIMENSION WITH MULTIPLE QUANTILES

Guorong Dai[1], Ursula U. Müller[2] and Raymond J. Carroll[2]

[1]*Fudan University and* [2]*Texas A&M University*

Abstract: In this study, we focus on the analysis of high-dimensional data that come from multiple sources ("experiments"), and thus have different, possibly correlated responses, but share the same set of predictors. The measurements of the predictors may be different across experiments. We introduce a new regression approach, using multiple quantiles to select those predictors that affect any of the responses at any quantile level and to estimate the nonzero parameters. Our approach differs from established methods by being able to handle heterogeneity in data sets and heavy-tailed error distributions, two difficulties that are often encountered in complex data scenarios. Our estimator minimizes a penalized objective function that aggregates the data from the different experiments. We establish the model selection consistency and asymptotic normality of the estimator. In addition, we present an information criterion that can be used for consistent model selection. Simulations and two data applications illustrate the advantages of our method in recovering the underlying regression models. These advantages come from taking the group structure induced by the predictors across experiments and the quantile levels into account.

Key words and phrases: Data integration, high dimensional data, information criterion, penalized quantile regression.

1. Introduction

To set the stage for this work on data integration (DI), consider K data sets from K different populations, where K is some fixed number, with linear regression models

$$Y_k = X_k^{\mathrm{T}} \alpha_k^* + U_k \quad (k = 1, \ldots, K). \tag{1.1}$$

Here, Y_k is a scalar response, X_k is a p-dimensional predictor, α_k^* is a p-dimensional parameter vector, and U_k is the error term. Zellner (1962) referred to this set of models as *seemingly unrelated regressions* and proposed the idea of estimating the regression parameters simultaneously using a generalized least squares method. The responses in model (1.1) are different, but dependent. The predictors are

Corresponding author: Guorong Dai, Department of Statistics and Data Science, School of Management, Fudan University, Shanghai 200433, China. E-mail: guorongdai@fudan.edu.cn.