Nonparametric Estimation for Symbolic Data — Measurement Error Models

Jiayang SUN, *Case Western Reserve University, USA*, E-mail: jiayang@case.edu
Xiaofeng Wang, *Case Western Reserve University, USA*

Key words: Symbolic data, inhomogeneous measurement errors, errors in covariates, deconvolution, nonparametric density and regression estimation

Mathematical Subject Classification: 62G07, 62G08, 62G20, 62P10, 62P35

Abstract: Many interesting practical problems can be formulated as studies about data with measurement errors. On one side, data measured with errors are typical examples of "symbolic" data recently defined by Diday and others. Symbolic data extend standard data to the case of variables not restricted to being numerical or categorical but more generally being any "symbolic value", such as intervals, probability distributions, functions, finite set of values etc. On the another side, the density estimation problem from the data with measurement errors is related to the long-challenging deconvolution problem, and the regression estimation with errors in covariates is related to the regression problem in a general linear model with repeated measures designs. Most existing procedures assume that the errors are homogeneous. In this talk, we present new procedures for the deconvolution problem and the regression problem, respectively, when the errors can be *inhomogeneous* and fairy generally distributed. The estimator is "suboptimal", stable and easy to compute; in particular, NO Fourier transformation (as it is for most deconvolution estimators) is needed in the calculation. The idea starts from the case when the errors are uniformly distributed. It then proceeds to the case when errors are a mixture of uniforms, hence approximating a large class of error distributions (including the normal distribution), much like how most random number generators are built up. This opens a new line of attacks in deconvolution problems. Based on a neat representation of our new density estimator for data with measurement errors, we show how a new nonparametric regression estimator can be established based on a "density-clone". The new procedures are consistent and computationally efficient. They are compared with some of existing estimators. We shall also talk about their implications to symbolic data analysis.