Semiparametric and Nonparametric Density Deconvolution for Data with Measurement Error; Applications to Nutrition Data

Xiao Chen Yu
Supervisors:Ray Carroll, James Brown and Stephen Woodcock
PhD Thesis Mathematics
University of Technology Sydney
Faculty of Science
Year of Submission:2021

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Xiao Chen Yu declare that this thesis, is submitted in fulfillment of the requirements for the award of PhD thesis: Mathematics, in the School of Mathematical and Physical Sciences at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature:

Production Note:
Signature removed prior to publication.

Date: Mar 2021

Contents

Part I

Abstract

Our inspiration behind this thesis is nutritional data, more specifically nutritional data collected through short term methods such as the 24HR recall. These collection methods obtain results that are quite accurate in what a subject consumed in a day, but is not an accurate representation of what a subject's consumption pattern looks like in long term. This leads to many statistician using measurement error models to adjust for the difference.

As our society as a whole becomes more aware of our health and how our eating pattern may effect it, more and more studies have come to focus on such ideas. And more recently, studies have come to focus on just understanding what the distribution of a populations consumption pattern looks like, in hopes to answer questions such as how does our society as a general consume a nutrition of interest, are we over or under consuming a certain food group or nutrition, has our consumption pattern changed as time passes, and so on.

So far in studies that use measurement error models to help obtain a density curve that represents a populations consumption patterns, most studies require additional information or additional assumptions that are given without specifying a reason such as assuming a certain distribution for the error terms of the model. For our thesis, we wish to develop a method that can obtain an unbiased distribution of a populations long term consumption pattern without additional information and minimal assumptions.

In this thesis, we start with a simple classical error model that will work well for continuous data, this may be good with nutrition data such as protein, fat and fiber. We then move on to allowing replicates in our observed variable, in doing so, we can let go of most assumptions on the error term. We do this because most 24HR recalls collect multiple entries from the same subject, which can work as replicates. We then move on to a more complex error model that is designed for zero-inflated data. We are interested in such a model since data collection methods such as the 24HR recall also collects information on what food we eat in a day, since it is very rare that we will eat every type of food in a 24 hour period, the 24HR recall will contain a large amount of zero. We hope to develop a method that

can help with estimating a populations long term consumption pattern using data collected using this short term method that contains excess zero.