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.

Part II

Chapter 1: Introduction

Nowadays when we turn on the television or go onto the internet, we often see words such as “diet”, “calories”, “high protein”, “multivitamins” and so on. We are a society that is more and more interested in what we eat and how that effects our well being. Every now and then a new diet pops up and sometimes we see new findings on food that we think we already understand, but how much of these is actually helpful? How is our society doing compared to 5, 10, 15 years ago? What new things do we need to be careful of now? These are all questions that we as a society wish to understand, and these are also things that nutritionists wish to test out and prove.

In order to understand how our nutritional intake may effect our well being, one big step is to collect information on our food intake. This in itself is a big challenge. Collecting detailed, accurate and large quantity data will allow us to obtain significant results. But given the man power and financial resources, it is impossible for us to collect such data. The trick is to find the balance between man power, financial resource and accurate, large quantity data, the result is two data collection methods: food records and 24 hour recalls.

Food records and food diaries relies on test subjects keeping a food journal where they record all their food intake continuously for a significant period of time, commonly a few weeks or a few months. This method allows us to obtain a significantly large data, but compromises on the accuracy of the data, for example: a subject may have consumed a chocolate bar without much thought and then forgets to record it, or the subject may have consumed a bag of chips each night but feels too guilty to admit and therefore decides to under record the truth. There is also the concern that our food intake pattern may change due to the availability of each food, having recorded the data continuously for a period of time, the data will not be able to represent our food consumption for the whole year as the season and weather changes.

24 hour recall relies on test subjects recalling what they had for the past 24 hours, usually a test subject is asked to recall their intake every few months. A common pattern would be

once every three months. The data size for this type of collection method is compromised greatly, but the accuracy from each recall is improved. After all, it is much easier to recall exactly what we have consumed and how much we have consumed "yesterday". Also, since we are only asked to recall the past 24 hours, we are less worried about our "guilty pleasures" being exposed and will answer more truthfully. Another advantage would be that since we are collecting a recall every few months, this data will not have to worry about the changes in season. Although we do not need to worry about season changes, this does not mean that the data collected from a 24 hour recall can be used to represent our long term consumption pattern. Since we are collecting data from such a short period of time, there will be many different types of food that we will not be able to consume within this time period, which may be consumed on the many days where the test subject is not asked to recall their intakes. This means that a food may be part of this subjects diet but will not be recorded as part of their data.

As we have discussed the pros and cons of two of the most commonly used nutritional collection method, we see that both methods contain measurement error, either due to the recording error of the subject or due to the limitations of the method. We believe that it is important to correct these errors as best as we can in order to obtain unbiased results. Therefore for our thesis, we will be incorporating the use of measurement error models as part of our analysis on nutritional data.

As it is mentioned at the start of this chapter, there are a lot of questions we as a society wish to understand using information from our food intakes. For this thesis, we are less interested in how food intake may effect our other factors such as our weight, our health and our well being. The questions that this thesis wish to understand is how the population is doing in general for each type of food or nutrition, and given our current nutritional guidelines, are we meeting these guidelines. That is, we wish to develop a method that analysis a nutritional data and with the help of existing models and techniques, and we will be able to obtain a distribution of what the populations consumption pattern would look like.

Part III

Chapter 2: Literature review

There has been an increasing interest in the use of measurement error models in the field of food nutrition. Due to reasons such as time and cost in the data collection process, many people have come to rely on measurement error models to find the link between short term collection data and their long term true values. Nusser et al. (1996) looked at how to estimate a populations long term food intake semiparametrically, Nusser et al. (1995), Tooze et al. (2006) and Dodd et al. (2006) created and tested out different measurement error models which allows a more accurate understanding towards data collected that has a large amount of zero. Subar et al. (2006) combined covariates such as gender or age group into the error model to allow a more accurate understanding of each sub-populations nutritional intake. Tooze et al. (2002) explored the concept of replicated data for nutritional data collection. Kipnis et al. (2009) modeled the data to see how a populations food intake may affect certain health outcomes, such as how fish intake may affect blood mercury levels. Zhang et al. (2011) introduced multivariate measurement error model for dietary data. Subar et al. (2001) compared analysis outcomes from multiple sources of data collection.

From all these literature mentioned above, we see that some of the more popular methods for data extraction in the field of food nutrition is to either ask the subject to maintain a food diary which records their food and drink consumption for small periods of time, or to ask the subject on multiple days to recall their food and drink consumption for the past 24 hours. For both methods, the data collected can all be considered as a short term and is not the focus of our interest. In general, we tend to be more interested in how we consume nutrition on a long term basis since it is the long term consumption pattern that effects our body and health. Considering the cost and manpower it takes to collect nutritional data, it is both impossible and unpractical to accurately collect any subjects long term nutritional intake. We can see how analysing measurement error models have become so popular in the area of food nutrition. For the remainder of the chapter, we will have a look into measurement error models and some of the different methods that can be used

to analyse the measurement error model. Measurement error models have been a popular topic for many years. They have been explored by statisticians from as early as 1950, and are frequently studied by statisticians everywhere. Measurement error models are also called Error-in-variables models, which are regression models that account for the measurement error that occurs in variables. The idea behind these models is that there are times when obtaining a certain variable that we desire becomes very challenging or almost impossible, then we will find a similar and observable variable to help to estimate our variable of interest. A difference may occur between the observable variable and the unobservable variable, so we need to take into account this difference since if this is not done, any further analysis will become biased and unusable. An example that requires measurement error models is a questionnaire concerning food consumptions, it is impossible for the respondents to have exact memories of their food and drink consumptions for long periods of time, therefore most food questionnaires either asks the respondents to keep a diary journal for a specific period of time or only ask the respondents for their exact food intake for the past 24 hours. Both methods of extracting information from respondents allow error values to form and will create biased analysis results if the error is not accounted for. Here we will be exploring some popular methods in analysing a measurement error model.

1 Methods for analysing measurement error models

A measurement error model is often considered as two parts, the regression model and the error model. The regression model is a model that shows the relationship between the response variable and the unobserved variables. The error model is a model that shows the relationship between the observed variable and the unobserved variable. Let Y be the value of the response variables, T be the unobservable predictor, and W be the observable predictor. Two types of error should be considered in a measurement error model: e is the error of the regression model and U is the error term in the error model. The goal is to use information from W and sometimes from U to estimate T and therefore obtain a more accurate estimation in the regression model.

Let's start by using a simple classical error model to demonstrate what a measurement

error model looks like, and later we will discuss three different types of error models, and in the end we will determine which error model is more suited for our research:

$$Y = T\beta + \epsilon, \tag{1}$$

$$W = T + U. \tag{2}$$

Here equation 1 is a basic linear regression model, identifying the relationship between the response variable and the unobserved variables, where β would be a set of coefficients for the unobserved variables. A *naive analysis* is when we analyse this regression ignoring the presence of measurement error, where the response is modeled directly on the observed variable. That is we analyse the regression $Y = W\beta + \epsilon$ instead of $Y = T\beta + \epsilon$. In most cases, a naive analysis is biased for the true regression relationship between Y and T (Fuller (2009), Cook and Stefanski (1994)). Given the type of error model used, the naive analysis may give overestimated or underestimated results of the coefficient β.

Equation (2) is the error model, as an example we used the simplest form of a classical error model. Though generally the error model be separated into three groups: Classical error model, Berkson error model and mixed error model. We will give a brief explanation on the differences between these three types of error models.

Classical error models: The classical error model is a widely used model in the measurement error literature (Carroll et al. (2006)). The simplest classical error model takes the form

$$W = T + U. \tag{3}$$

This is when the observed variable has a larger variation than the unobserved variable. For the example modeling body mass index using daily calorific intake, since the daily average intake that each subject reports is very different from the actual long term average intake.

Berkson error models: Berkson (1950) introduced a model where the unobserved variable T has a larger variance than the observed variable W, as opposed to the classical error models where the variance of the observed variable is larger. The simplest Berkson's error model can be written as

$$T = W + U.$$

One example of where the Berkson model occurs in practice is where multiple individuals, having the same characteristics, are all assigned the same value of covariate. For example, the Berkson model could be used when all those of the same age, gender and living environment are assigned the same dose for pollution when modeling the relationship of children on the eventual development of lung disease and their long-term NO_2 intake. In this case, Y would be a measurement of lung disease, T would be the true long-term NO_2 level and W would be the NO_2 level collected from the bedroom and kitchen with stationary recording devices.

Carroll et al. (2006) documented that even though the classical error model and the Berkson error model may look quite similar, it is important to be able to know when to use which model. Carroll et al. (2006) gave a small guide to help differentiate these two models: if the error-prone variable is measured uniquely to each individual and can possibly be replicated, then we should use the classical error model, if all individuals in a group are assigned the same error-prone value, but the true variable is particular to an individual, then the measurement error is Berkson.

Mixed error models: When working with epidemiology problems, Reeves et al. (1998) considered working with a mixture of classical and Berkson error model. This model incorporates a latent variable L which acts as an intermediate between the observed variable W and the unobservable variable T. The simplest mixed error model can be written as:

$$log(T) = log(L) + U_b,$$
$$log(W) = log(L) + U_c.$$

Here U_b is the Berkson error, and U_c is the classical error. When $U_b = 0$ we obtain the classical error model, and likewise when $U_c = 0$ we obtain the Berkson error model. This model is a lot less common and the amount of papers on this problem are few, and almost all deal with radiation research (Mocanu and Oliver (1999), Mallick et al. (2002)).

Since we are focusing on analysing data from food questionnaire, we can say that W represents the data that we have collected for the questionnaire, an example of the type of data we collect are protein intake, alcohol intake and such, then T would be the true long term average daily intake of said variable, that is how much on average a person would consume a particular food or nutrition in a day, U would be the difference between what we

collected for the questionnaire W and what the true value T should be. Therefore in this situation, it is more reasonable to use a classical error model. For our research, we will start with the simplest form of classical error model, we will then move to a more complicated form of classical error model that is specific to data with excess zero. Now if Y is an outcome in which we believe may be influenced by our variable of interest, such as BMI, cholesterol level and so on. Then by using a regression model on the outcome Y and the true intake value T, we will be able to obtain an unbiased result on their relationship β.

Many methods have been developed to analyse these measurement error models, such as Bayesian (Castro et al. (2013), Mallick and Gelfand (1996), Carroll et al. (2004)), regression calibration (Carroll and Stefanski (1990), Armstrong (1985)), and SIMEX (Cook and Stefanski (1994), Stefanski and Cook (1995)). These methods have been extensively researched, used and compared with each other.

1.1 Bayesian

Bayesian methodology is an efficient tool to analyse measurement error models. This method can be used on almost any type of measurement error models, even some that contain highly nonlinear regressions and mixed or multiplicative error models (Berry et al. (2002); Carroll et al. (2004); Holmes and Mallick (2003)). This method requires an input of a prior distribution for all parameters of variable distributions that is used in the process of estimation, that is, we need some basic information on the distribution parameters of Y, T, W, ϵ, U and β, and sometimes knowledge of the distribution parameters of covariates and instrumental variable if such variables are part of the regression and error model. In this case, covariates are considered error-free variables that are mainly characteristics of each participant, and instrumental variables are error-free variables that are related to the true variable. With all these detailed information on the distributions of all variables, Bayesian approaches can give quite an accurate estimation and inference, but at the same time, it can be hard to compute and the computational running time for Bayesian analyses can be quite long (Berry et al. (2002)).

"RStan" is an existing software package in the statistical program "R" which can be

used to analyse measurement error models using Bayesian methods. The advantage of using such programs is that there is no need to manually compute the joint likelihood distribution since this step is programmed as part of the package, and therefore it removes the problem of having to work with complicated distributions by hand. But since this method needs to input information on the parameters of variables Y, W and T in the form of prior distributions, it can be a bit difficult when we need to identify these information on variables which we are unsure of such as the unobservable variable T and the error terms ϵ and U.

Bayesian methods for measurement error models have been frequently explored. Most studies focus on more commonly used regression models and error models. Castro et al. (2013) used Bayesian methodology on measurement error models with linear regression models and replicated data. They concentrate on cases where the error model 2 may be a classical error, Berkson's error or a classical/Berkson's mixed error. In this paper, it has been specified that the data is unpaired and the replicates in the data can be either equally or unequally replicated, also the variance of the error term U can be either homoscedastic or heteroscedastic and the value is not assumed to be known. The results show that the Bayesian method performs well in such cases, and also shows how flexible Bayesian method can be with measurement error models.

Mallick and Gelfand (1996) look at the Bayesian approach for semiparametric regressions where both the response variable and the independent variable may contain errors, hence we have one regression model with multiple additive error models. In this paper, it is assumed that for the latent variable T, which is usually unobservable, there is a portion that is observed without error. The author looked at the cases where the error model may be either classical or Berkson. Reasonable results were obtained in the simulations, though this largely relies on assuming the correct covariate link and calibration function.

Holmes and Mallick (2003) focused on demonstrating how to use a Bayesian approach on the case where we have a generalised nonlinear model as the regression model and either a simple classical error model or a simple Berkson's error model as the error model. For this paper, they only provide the response variable Y and the observed error prone variable W, also it is assumed that the link function for the generalised nonlinear model is unknown. Simulations show that the Bayesian approach works well on both when the error model is

classical and Berkson.

Mallick et al. (2002) use the Bayesian approach on a dose-response regression model where the error model is a Berkson and classical mixed error model. This paper was developed specifically for the Nevada test-site data. They compared the results with the cases where the error model is just a classical error model and just a Berkson error model. The results show that using the mixed error model for this particular data gives the smallest deviance information criterion and yields a much larger relative risk for a high dose case. Suggesting that a mixed error model is the way to go for the Nevada test-site data.

Sinha et al. (2010) developed a semiparametric Bayesian method for logistic regression with a classical error model where the unobserved variable distribution is nonparametric. This method was developed with epidemiology data in mind and shows that this method performs well for the NIH-AARP diet and health data. It is also mentioned in this paper that they discarded the use of regression calibration (a method that will be discussed in a later section) because of its poor performance when it come to semiparametric regression models.

Berry et al. (2002) looked at improving a nonparametric regression by modeling a smoothing spline from a Bayesian standpoint. For this paper, a classical error model was used, and for the regression model, the authors looked at a smoothing spline and a P-spline. Simulations were performed for both cases and both yielded satisfactory results. Although there was no discussion on how the priors were selected, it was mentioned that a number of different priors were used and the results compared, it shows minimal changes, indicating how robust the Bayesian approach can be.

Carroll et al. (2004) incorporated the Bayesian method into a set of regression that is both nonlinear and nonparametric with an instrumental variable available. For this paper, they looked at a regression model $Y = m(T, \beta) + \epsilon$ where $m(\cdot)$ is a polynomial function, a error model $W = T + U$ and an additional instrumental model $S = \alpha_0 + \alpha_1 T + v$, where S is an instrumental variable. In this case, we assume that only (Y, W, S) is observed and that all error terms (ϵ, U, v) have mean 0 and that variables T, U, ϵ and v are mutually uncorrelated. This paper concludes that using a Bayesian approach yields a root n consistent estimation on the measurement error variance.

1.2 Regression Calibration

Regression calibration is an easy to implement method that can be used on most general regressions. The principle behind regression calibration is to replace T with a calibration function, namely the expectation of T given W, and fit the regression model using the expectation instead of T.

One of the biggest challenges in implementing regression calibration is determining the calibration function. Since T is unobserved, it is difficult to find and validate the relationship between the unobserved T and the observed W. Carroll et al. (2006) proposed some options to overcome this issue, such as:

- If there is internal validation data, where both T and W are observed, we can regress T on W and the other covariates in the validation data. Though Carroll et al. (2006) argues that this is a missing data problem and should use missing data techniques instead of regression calibration.

- If there is an unbiased instrumental variable Ψ within the data, then we can use the information on Ψ for help to obtain a calibration function. An instrumental variable is a variable that is related to the case of interest and is uncorrelated with all variables apart from T. Using the example of the food questionnaire, the instrumental variable Ψ can be the intake assessed by other food diaries. In this case, the instrumental variable Ψ can be considered as a replicate of W, then variable Ψ can also be considered as an unbiased substitute to the variable T. Thus we can use the regression of Ψ on W as the calibration function.

One of the advantages of using the regression calibration method is that it is easy to program and provides a consistent estimate of the slope parameters for different types of regressions. However, it is known to have poor performance for the more complex regressions models such as generalised linear mix effect models (Wang et al. (1998)).

Some papers study how well regression calibration performs on more complex regression models, for example Carroll et al. (2006) performed regression calibration on some highly nonlinear regression models with an additive error model, and Küchenhoff and Carroll (1997)

used regression calibration on segmented linear and logistic regressions with simple classical error model. Both of these papers compared this method with an alternative method called SIMEX which will be discussed in more detail in later sections.

Rosner et al. (1989) explored using regression calibration to estimate parameters for probit and logistic regressions when the error model is a simple Berkson error model where the error term U is assumed to have a normal distribution. In this paper, a separate validation study is needed to assist in estimating the relationship between the observed variable W and the latent variable T. Based on the simulation studies, they believe that regression calibration performs reasonable for cases where the true odds ratio $exp(\beta)$ (β is the coefficients in regression model 1) is less than 3.

Carroll and Stefanski (1990) performed regression calibration on a quasi-likelihood regression model with simple Berkson error model. For this paper, they assume that the data can be available in one of the following five ways: primary data, where we only have the response Y and the observed W; internal validation, where we have the response Y, the observed W and a small portion of the truth T from the same data; internal reliability data, where we have the response Y and the observed W where W have replicates; external validation, where we have additional information on W and T from an external source; and lastly external reliability, where we have additional replicated W from an external source. The asymptotic theory was also developed for all five situations. The simulation yielded reasonable results when using regression calibration on regression models with Berkson error.

1.3 Simulation Extrapolation (SIMEX)

An alternative method to regression calibration for measurement error model analysis is simulation extrapolation (SIMEX). This method was first developed by Cook and Stefanski (1994). The method was designed to be able to fit a wide range of measurement error models without going through any complex coding and computational process. Three assumptions are made for this method:

1. The error model (2) has to be an additive model.

2. The variance of error U is homoscedastic.

3. The error U follows a normal distributions.

The idea behind SIMEX is that identifying and then correcting measurement error analytically is a difficult process. Instead, we can add on an additional set of successively larger error to the variables T in order to obtain a set of increasingly biased parameters. Using these biased parameter estimates, we can extrapolate back to where T has no measurement error and therefore we can estimate the unbiased parameters. SIMEX can be considered as a two-step process, the simulation step and the extrapolation step. And we will use a simple regression model (equation 1) with a simple classical error model (equation 2) to demonstrate how these two steps work.

Simulation step: Let $\lambda^* = (\lambda_1^*, \lambda_2^*, ..., \lambda_m^*)$ be a set of m error variances where $\lambda_1^* = 0$ and the values in λ^* is in ascending order, that is $\lambda_1^* < \lambda_2^* < ... < \lambda_m^*$. For each λ_i^*, a measurement of the observed variable W is generated by $W_i = W + \sqrt{\lambda_i^*}U$. Since W_i is a measurement of W, it should also be reasonable to consider W_i as a measurement of T. Therefore for each λ_i^*, a new β is estimated, we use $\hat{\beta}_i$ to denote the new estimate corresponding to each λ_i^*. These estimations are repeated a large number of times, and averaged for each value of i. The sets of average parameter estimates are then used in the extrapolation step.

Extrapolation step: Following the simulation of a set of mean parameter estimates for different values of λ_i^*, we fit a model between the biased estimates $\hat{\beta_{iT}}$ and the additional errors λ_i^*. Using lease squares estimation, the slope for the simple regression equation $\hat{\beta}_i$ consistently estimates $\beta\sigma_T^2/(\sigma_T^2+(1+\lambda_i^*)\sigma_U^2)$. Therefore to extrapolate back to where there is no measurement error (i.e. $(1+\lambda_i^*)\sigma_U^2 = 0$), we substitute $\lambda_i^* = -1$ into the final extrapolation function in order to find the unbiased estimate of β. The fitted equation between $\hat{\beta_{iT}}$ and λ_i^* is called the extrapolation function.

Cook and Stefanski (1994) described this method as a simulation based method that combines features of parametric bootstrap and method-of-moments inference. In this paper, they have only considered cases where the error model is a simple classical error model where the error is a normal distribution with the assumption that the error variance is known or can be reasonably well estimated from cases such as replicates. They also believed that the key to the success for this SIMEX method is the appropriate selection of the extrapolation

function, three extrapolation functions were proposed: a simple linear function ($\beta = a+b\lambda^*$), a quadratic function ($\beta = a + b\lambda^* + c\lambda^{*2}$) and a nonlinear function ($\beta = a + b/(c + \lambda^*)$). The authors argue that if the measurement error is normally distributed, then at least one of the three extrapolation function should be able to give a good estimate of the parameters. Simulations were performed on cases where the regression model (1) are either linear, logistic or probit, and the authors conclude that if given appropriate values for the parameters a, b and c, the nonlinear extrapolation function works best, and by considering the measurement error bias of these three functions, the authors conclude that the quadratic model and the nonlinear model gives smaller bias compared to the linear function.

One of the biggest advantages of SIMEX is that it is easy to compute. It requires minimal information on the unobservable variable T. To estimate β, SIMEX only needs the variables Y and W and the variance of U, or at least a well-estimated variance of U (Carroll et al. (2006)). One of the disadvantages of SIMEX is that the third assumption limits the distribution of U by only assuming that the error is a normal distribution. Therefore, for data sets with other distributions of error, such as a Laplace or gamma distribution, this method will not give accurate estimates of the parameters. This method also cannot be applied to Berkson's error models or mixed error models. Yet another disadvantage is that this method is sensitive to the variance of U, a poorly estimated variance can easily lead to overestimation or underestimation of the parameters.

The SIMEX method has since been extended. Stefanski and Cook (1995) introduced the idea of combining Jackknife technique with SIMEX. In the paper, they worked with a simple classical error model where they do not know the variance of the error term, and the variance is estimated using the Jackknife estimation. A simulation was performed to compare the performance of SIMEX with a known variance of U and SIMEX with an estimated variance of U using Jackknife, similar results were produced. This leads the authors to conclude that having Jackknife as part of the simulation process can both help estimate the variance of U for data where it is unreasonable to obtain σ_U^2.

Carroll et al. (1996) investigated the asymptotic distribution of the SIMEX method. For this paper, they looked at a simple regression model $E(Y|T) = \alpha+\beta T$ with a simple classical error model $W = T+U$, where the variance of the error term U is assumed known. The con-

clusion showed that the SIMEX estimator is typically asymptotically normally distributed. They also mentioned that the asymptotic distribution is derived with the assumption that the regression error term are independent and identically distributed, and that it is possible to modify the asymptotic distribution to allow the case where the regression error term is NON-IID, but seemed unnecessary.

Wang et al. (1998) explored how to use the SIMEX approach for cluster data where the regression model (1) is a generalised linear mixed model. In the paper, they discussed many different regression models such as linear, logistic, probit and log-linear where the within-cluster error in the error model can either be homogeneous or heterscedastic. Once again a simple classical error model was used where we assume to know the variance of the error term. Wang argued that although regression calibration has the ability to obtain good estimates of the fixed effects, it does not do so well for the random effects, and they believe that SIMEX may perform better. The results show that SIMEX has the potential to estimate parameters with minimal bias.

Küchenhoff and Carroll (1997) explored the idea of using both regression calibration and SIMEX for the case where the regression model is segmented with a simple classical error model. That is

$$E(Y|T) = H\{\alpha + \beta(T - \tau)\}, W = T + U \tag{4}$$

where $(T - \tau) = T - \tau$ if $T \geq \tau$ and 0 if otherwise, also $H(\cdot)$ is the link function in which they suggest it to be linear or logistic. For this method, we need recorded data on Y and W with the assumption that the distribution of U is known. A set of simulations was performed using both regression calibration and SIMEX, and the general conclusion is that regression calibration usually has more bias, but SIMEX produces more variance. Also it is suggested that for SIMEX, a non-linear extrapolant is to be avoided.

Eckert et al. (1997) relax the first assumption of SIMEX and explored how SIMEX performs with a multiplicative measurement error. In this paper, it was suggested that we transform a multiplicative error model into an additive one, an example is that we can transform the multiplicative error model $W = T * e^U$ into the additive error model $log(W) = log(T) + U$. This leads to the new error model $h(W) = h(T) + U$, where $h(\cdot)$ is

a transformation function. Two families of transformation was suggested: power transformation and spline transformation, and simulations were performed for both cases, and both cases yield reasonable results, though the authors did point out that there is no guarantee that one can find a single transformation that will achieve an additive error model with near normal error distribution.

Devanarayan and Stefanski (2002) relax the second and third assumptions and developed a new SIMEX method (Empirical SIMEX) that can analyse the parameters of a regression model where the measurement error model may have heteroscedastic error variances. In this paper, they used a simple classical error model where they allow the variance of the error term to be unknown, but replicates of the observed variable W is required. The idea is to use the replicates to generate the error increasing data. Though there is no mention of the minimum amount of replicates needed, in their simulations, they used a data with only 2 replicates, and they compared the results with the traditional SIMEX method where we know the error variance and it was concluded that they results are similar even when the assumption of homogeneity is reasonable.

2 Methods for estimating the density analysing only the error model

Recently there has been an increasing interest in analysing only the error model part 2 of a measurement error model, and obtaining an unbiased density estimation on the latent variable T. Using the food questionnaire example again, if our variable of interest is alcohol intake or vegetable intake, then T can be something like the usual average daily intake of alcohol or the usual average daily intake of vegetables for a population group, estimating the density of such an intake will allow us to understand what the general trend is like for said intake, and this may lead us to conclusions such as how much of the general population over consume alcohol or under consumes vegetables. If we perform density estimations on different groups of populations, we can also compare the densities and obtain answers for questions such as does males in general drink more than females or which country, in general,

consumes more vegetables as part of their diet?

For this section, we are going to explore two methods which can estimate the density of T using only the error model 2: Kernel density deconvolution and sieve maximum likelihood.

2.1 Kernel Density Deconvolution

Using the simplest form of a classical measurement error (2), where f_W, f_T and f_U are the densities corresponding to variables W, T and U respectively. The density of T can be expressed as $f_W = f_T * f_U$, where $*$ is the convolution operation. *Deconvolution* uses the density of observed variables f_W and f_U to estimate f_T. Stefanski and Carroll (1990) first developed this method where they called it deconvolving kernel density estimator. In the paper, they used a simple classical measurement error model (2), where only one set of W is used and we require the assumption that the density of U, f_U, is known. For the *kernel density deconvolution*, a kernel estimator $f_T(z;h) \approx (nh)^{-1}\sum_{j=1}^{n} K\{(z - W_j)/h\}$ is used as the deconvolution process, where h is the *bandwidth* and $K\{\cdot\}$ is a *kernel function* and can be expressed as $(2\pi)^{-1}\int e^{-ilz}\phi_K(l)/\phi_U(l/h)dl$, where ϕ_K and ϕ_U represents the characteristic function of the function or density of K and U. In this paper, they also discussed the asympotic theory, and performed a set of simulations where it was concluded that 23 out of the 25 simulations obtained reasonable results. It is believed in this paper that kernel density deconvolution will be a viable technique for data with large sample sizes.

Carroll and Hall (1988) investigated the rate of convergence for deconvolving a density using kernel density deconvolution. Once again a simple classical error model was used, and they require the assumption that the density of the error term U is known. In this paper, they explored the optimal convergence rate for cases where the known density of U is normal and for more general errors such as gamma and double exponential. The conclusion was if the density of U is a normal distribution and that the density of T has k bounded derivatives, the optimal convergence rate is $(logn)^{-k/2}$, where n is the sample size. Whereas for when the known density of U is a gamma distribution with a shape parameter of α, the optimal convergence rate is $n^{-k/(2k+2\alpha+1)}$, and the rate of convergence for when the density of U is a double exponential is $n^{-k/(2k+5)}$. Concluding that it is difficult to converge efficiently when

the error term U has a normal distribution. This paper also briefly mentioned a multiplicative error model $W = TU$ and quickly concluded that deconvolution will be difficult for such a model.

There are two important decisions to make when performing this method. One is the choice of bandwidth h and the other is the choice of kernel function K.

Delaigle and Gijbels (2004b) emphasized that the choice of the bandwidth can strongly influence the shape of the estimated f_T. In this paper, the model used was once again a simple classical error model where the density of the error term U is known. It is also suggested that when f_U is only known up to some parameters, we can estimate f_U through repeated measures. They investigated five different selection techniques: NR bandwidth selector, PI bandwidth selector, SEQ bandwidth selector, CV bandwidth selector and BT bandwidth selector. Simulations were conducted to test which bandwidth selector will give optimal results. In the simulations they considered two types of error, where U is a normal distribution and where U is a Laplace distribution, they also worked with three different sample sizes 50, 100 and 250. The final conclusion is that even though there is no one technique that is the optimal bandwidth selector, the plug-in (PI) bandwidth selector and bootstrap (BT) bandwidth selector are two of the best techniques, and Delaigle and Gijbels (2004a) support this statement. PI bandwidth selector finds the bandwidth by minimising the squared asymptotic bias, whereas BT bandwidth selector uses PI bandwidth as a pilot bandwidth and substitutes this pilot bandwidth into a modified mean integrated square error formula.

When choosing kernel functions, Delaigle and Gijbels (2004a) suggest that it would be best to choose among densities whose characteristic function have compact and symmetric support. Functions such as a Fourier transformation, normal density, and some sine and cosine functions have all once been used as the kernel functions for kernel density deconvolution. The paper suggested that a good kernel function for estimating the density of T is

$$K(t) = \frac{48t(t^2 - 15)cos(t) - 144(2t^2 - 5)sin(t)}{\pi t^7}. \tag{5}$$

A software package exists for kernel density deconvolution on measurement error models in

the statistical program "R". This software package is called "fDKDE", it includes commands for calculating bandwidths and for calculating the density of T when both the variance and the distribution type for U is known. This software was originally a Matlab code developed by Aurore Delaigle and has later been translated into the R language. The R version of this code has yet been uploaded into CRAN. "fDKDE" contains two bandwidth selectors: the plug-in (PI) bandwidth selector and the cross-validation (CV) bandwidth selector. The PI bandwidth is based on Delaigle and Gijbels (2004b) and the CV bandwidth is based on the method from Stefanski and Carroll (1990). For the kernel function, this software uses a second-order kernel $\phi_K(t) = (1-t^2)^3$ which corresponds to the kernel function in Equation 5. Though through experimentation, we have concluded that this program does not allow the option that W may have replicates.

In the work of Wand (1998), they focused on how using a limited sample may effect the performance of kernel density deconvolution. For this paper, a simple classical error model is used where the density of U is known. MISE was used as an indicator on how well kernel density deconvolution performs in varies scenarios. They chose to compare two kernel functions ($K_1 = (48t(t^2-15)cos(t)-144(2t^2-5)sin(t))/(\pi t^7)$ and $K_2 = 3/(8\pi)(sin(t/4)/(t/4))^4$), along with two types of error term (normal and Laplace), two types of sample ($n = 100, 1000$), and five different error percentages ($p = 10\%, 20\%, 30\%, 40\%, 50\%$, where $p = Var(U)/Var(W)$). The conclusion was that both kernel functions perform similarly in most scenarios, though K_1 performs better for a normal error with a higher p. The MISE are all relatively low for low levels of error, but increases dramatically as the error percentage increases, also in increase is MISE is more pronounced for the normal error case that it is for the Laplace case.

As mentioned previously in Carroll and Hall (1988), Delaigle and Gijbels (2004b), Wand (1998) and many other papers, when the error term has a supersmooth distribution such as a normal distribution the rate of convergence is very low. But Fan (1992) argues that having a normal distribution for the error term U in practice is more common than any other distributions, therefore in their paper, they concentrated on answering the question how large can the error term be to be feasible when using kernel density deconvolution. For this paper, a simple classical error model is used with the assumption that the distribution of the error term is known. In their simulations, six different levels of error percentage was

used (including one that is error-free), and it is determined that performing kernel density deconvolution on models with large error terms may be difficult and that some information on the latent variable T may be required. They believe that for some special cases, an error percentage up to 80% may be feasible, but can not conclude what percentage would work in general.

Delaigle and Gijbels (2007) mentioned that although kernel density deconvolution works well in theory, it is less successful in practice depending on the techiques used on the calculating the integrals and optimisation process. They pointed out the places in the estimator that may have computational problems along with the different common types of computational problems that may occur, and some guidelines on how to solve these problems. In this paper, they mentioned that although many papers such as Delaigle and Gijbels (2004a), Delaigle and Gijbels (2004b), Wand (1998) and Fan (1992) use $K(x) = (48x(x^2 - 15)cos(x) - 144(2x^2 - 5)sin(x))/(\pi x^7)$ as the kernel function, there is no set kernel function and depending on the choice of kernel function, we may end up with an integral with no closed form. In this case, they have suggested to use numeric approximation methods that is devoted to approximating Fourier transformations.

In Achilleos and Delaigle (2012), they discuss the importance of choosing the bandwidth when performing kernel density deconvolution. For this paper, they focused on a simple classical error model where they assume to know the density of the error term. It is mentioned in this paper that there are already many existing methods in bandwidth calculation, but all construct a global bandwidth, that is the bandwidth is the same on all points, the goal of this paper is to develop local bandwidth selectors where the bandwidth are no longer constant at all points. Though simulations, they have illustrated that local bandwidth selectors bring significant improvement over the global bandwidths when the estimated density has local features. They also proposed that no one bandwidth selector out stands the rest, their performance depends on each individual case.

2.2 Sieve Maximum Likelihood

The method of sieve was first introduced by Grenander (1981) and consists of two steps: a step which performs optimisation within a subset of the parameter space, then another step to allow this subset to "grow" with the sample size. Geman and Hwang (1982) recognises that by combining the method of sieve with the maximum likelihood, it is possible to perform estimation where a classical maximum likelihood fails to do so, such as performing maximum likelihood over infinite dimensional space. This is later commonly mentioned as *sieve maximum likelihood* (SML). The general idea for SML is that each likelihood, $\mathcal{L}(\theta)$, is broken down into M sections: $\mathcal{L}(\theta) = \big(\mathcal{L}_1(\theta), \mathcal{L}_2(\theta), ..., \mathcal{L}_M(\theta)\big)$, where each section contains a smoothing parameter π_i where $i = 1, 2, ..., M$. The first step from the method of sieve optimises the smoothing parameters individually. Then we allow the likelihood function to "grow" by adding every optimised likelihood together

$$\mathcal{L}(\theta) = \textstyle\sum_{i=1}^{M} \pi_i \mathcal{L}_i(\theta).$$

Carroll et al. (2010), Shen (1997) and Geman and Hwang (1982) have suggested that SML works well when estimating densities of semiparametrically and nonparametrically. Shen (1997) developed the general theory on asymptotic normality and the efficiency for semiparametric and nonparametric sieve maximum likelihood and claims that the results depend on two aspects: the smoothness of the likelihood function and the size of the parameter space. The authors believe that when the parameter space is large, a classical method of likelihood may meet difficulties whereas the SML method may overcome these difficulties, also a smoother likelihood function may result in a slow rate of convergence.

Chen and Pouzo (2009) explored the use of sieve estimators on conditional moment models that contain unknown functions depending on endogenous variables. Focusing specifically on models where the residual function is non-smooth. In addition, they also obtained the asymptotic theory of normality and rate of convergence for the model.

Carroll et al. (2010) introduced the SML method into measurement error models. In this paper, a specific type of data was used: a data with two samples, a primary sample and an auxiliary sample. The assumption is that these two samples are correlated and have different joint distributions and that their true variable is of the same distribution. Also for

both samples, they require a variable that contains a nonclassical measurement error and a covariate variable that is discrete. The authors demonstrated to us how we can obtain nonparametric estimation without the knowledge of the measurement error distribution or the use of additional variables, using only the assumption that the regression function is the same between the two samples. Through simulations, the authors compared the results estimated using this two sample SML estimation method with four other different estimation methods: one where the measurement error is ignored, one using a parametric estimator where all parameters is correctly specified, one using a parametric estimator where the distribution of the measurement error is misspecified and lastly one using a parametric estimator where both the measurement error distribution and the latent variable model is misspecified. The results showed that using two sample SML obtained satisfying results that are just as good as the one from a parametric estimation with all parameters correctly specified.

There are many advantages to using SML to estimate the density of T. Geman and Hwang (1982) mentioned that the sieve estimators change only slightly as the sample size changes, showing that this method is able to work quite well with small sample sizes and much as large sample size. Chen and Pouzo (2009) show that, with the same number of sieve terms, any initial set of smoothing parameters can converge to an optimal set of parameters. This allows flexibility when choosing the initial parameters to perform the estimation of sieve maximum likelihood. Carroll et al. (2010) argue that this flexibility of SML is one of the characteristics that give an advantage over the kernel density deconvolution.

Shen (1997) suggested using orthogonal polynomials such as Hermite polynomials as smoothing parameters in the sieve maximum likelihood, though no further research was performed. For this thesis, we will be exploring this idea on a simple classical error model both semiparametrically and nonparametrically, then extending this idea to a more complicated error model which allows data with excess amounts of zero. For the entirety of this thesis, we will be operating with only the information on the observed variable W and occasionally the information on error term U with no additional information needed.

3 Nutritional data

As a society have become more and more interested in what we eat and how we are doing in comparison to others around us. Given that it is very unpractical and financially draining to follow subjects their whole life and record every food and beverage they consumed, methods were developed to collect data that is in someways considered accurate and is also a good balance financial-wise.

One of such method is the food frequency questionnaire (FFQ). Such a questionnaire contains a checklist of food and beverages with a frequency response. There are some obvious advantages to such a collection method, one being that this is someways shows a subjects eating pattern, it is also less expensive since they are usually self-administrative forms. But there are also some big disadvantages, since this method of collection largely relies on a subject's memory. When asked how much water you usually drink in a day, the answer is most likely going to be a guess work. We also may forget that we had a chocolate bar as an afternoon snack.

Some more commonly used food frequency questionnaires are: Harvard FFQ that is developed by Walter Willett, M.D. and his colleagues (Chan (Chan)), Diet History Questionnaire that is directed by Fran Thompson and Amy Subar from the National Cancer Institute (Institute (a)) and also the Block FFQ that is also developed in the National Cancer Institute this time directed by Gladys Block (Institute (c)).

Subar et al. (2001) compared these three food frequency questionnaires by using a study from Eating at America's Table Study (EATS), first by analysing the subjects as is, then they analysed the subjects after some energy adjustments. They also seperated the data by gender. Their results show that both DHQ and Block FFQ obtain better results when the data is not adjusted with energy, but all three perform similarly when energy adjustment in included.

Block et al. (2006) looked at how food frequency questionnaire works on a specific group of the population such as the Hispanics. Where the conclusion yielded favorable results. Hernández-Avila et al. (1998) also did something similar but on women in Mexico city, results were also favorable, but did mention that results were less favorable when the data

were obtained outside the city or with a different gender.

Salvini et al. (1989) also looked at the food frequency questionnaire of a nurses' health study, focusing on the reproducibility and validity of each food and beverage intakes, concluding that the difference in the degree of validity for specific foods may help improve the questionnaire design.

Another common nutrition collection method is the 24HR recall (Institute (b)). This is another method developed by the National Cancer Institution. This collection method relies on subjects recalling their intake of the past 24 hours. Given that each subject is to recall from memory that has just happen not too long ago, 24HR recall is more likely to collect information that is more specific, for example when recalling that a subject has salad for lunch, the subject is more likely to correctly recall what type of lettuce was in the salad and what dressing was used. This allows an accurate data for each subject's short term intake. But there are also some obvious cons, such as even though 24HR recall gives accurate results in short term intake, we can not use it as a direct representation of a populations long-term intake. It is also time consuming, since the consumption is recalled in detail.

Subar et al. (2006) compared this short term method to food frequency questionnaire, and saw that the probability of consumption in 24HR recalls is strongly correlated with what was reported in food frequency questionnaire.

Nusser et al. (1995) developed a method that can adjust for measurement error and non-normality, this model was designed for estimating dietary intake distributions for 24hr recalls. Though the method does require at least two positive intakes for each subject.

Tooze et al. (2006) also look at analysing data received from 24hr recalls and how these data may help analyse the relationship between our eating pattern and various health outcomes. This paper focused on episodically consumed food, and introduced a new two-part method to help with excess zero problem. It assumed that all variables are normally distributed and estimated the distribution of consumption for various foods. As this method allows with-in and between person variability and the addition of covariates, it allows more flexibility than all other existing methods.

Part IV

Chapter 3: Semiparametric Density Deconvolution for Continuous Data without Additional Information

1 Introduction

In the previous chapter, we looked at several popular methods for analysing measurement error models. Some methods allow us to obtain a better understanding between the latent variable T and an outcome of interest Y, that is given we have a observed variable W that is in someways related to the latent variable T with error (e.g. $W = T + U$) and we use this relation to help obtain an unbiased estimate of coefficient β in the regression $Y = m(T, \beta) + \epsilon$, but we also see that there is an increased interest in estimating just the density of a latent variable T either semiparametrically or nonparametrically, after all looking at just the distribution of T itself may also yield many useful information. The idea is to use the limited amount of subjects that are in a particular data set to understand what the distribution for the general population would be like.

The concept of estimating the density of unknown variable T is useful in the area of public health. For example, for the continuous data collected from a health questionnaire such as the 24HR recall or food frequency questionnaire, where we wish to estimate the distribution of the population long-term nutritional intake using data collected from short-term food intakes. Having an accurate estimate on the distribution will provide information for questions such as what the percentage of people who over or under consume certain nutritional components is, or, in general, how little or how much a normal human would consume the nutrition of interest. The estimated distribution may also help us understand whether the existing health guidelines are accurate enough for the population to maintain a healthy lifestyle.

For this chapter, we will be looking at how to estimate the density of the latent variable in a simple classical error model utilising Hermite polynomials.

Suppose that we wish to determine the density of variable T, but we only have information on an observable variable W, where the relationship between the observed variable and the variable of interest can be expressed as a simple classical error model.

$$W = T + U,$$

here variable U is the error term assumed to have mean zero and is independent of latent variable T.

Now if we assume that we have an infinite number of subjects, then $\overline{W} = \overline{T} + \overline{U}$. Given that U has mean zero, then in the ideal yet impossible case of infinite subjects, we can say that the distribution of W will be very close representation to that of the distribution of T, but unfortunately, it is very unpractical and very expensive to obtain such a large amount of participants. Therefore we can not obtain an unbiased estimate of the distribution of T without taking into account the observation error U.

Many methods have been developed to analyse T from a simple classical error model, but most of these methods require additional information or assumptions in order to perform the analysis (Carroll and Hall (1988); Delaigle and Gijbels (2004b); Wand (1998)). Inspired by Schennach and Hu (2013), I wish to develop a method that can estimate the true density of T semiparametrically without any additional information beyond measurement W and minimal assumptions. This method will be optimising all parameters of interest using maximum likelihood where all unknown densities will be represented in a form that contains a set of orthogonal polynomials, we will be using Hermite polynomials. This method was only mentioned by Schennach and Hu (2013) as a theory. For this chapter, we will be taking a deeper look into this method, develop the methodology and see how well this method works in comparison to another density deconvolution method - kernel density deconvolution (Wand (1998)).

We will focus on the case where we observe only one measurement of W per individual, since there is such a large limitation on the amount of information given to us, for this chapter we will be making assumptions on the type of distribution that the error term U

will have. We will be relaxing this assumption in future chapters.

Section 2 starts with a description of Hermite polynomials following with a discussion on how this set of polynomials will aid in estimating unknown densities. Then in section 3 we will look at the methodology on estimating the density of T with a selected distribution for U. Section 4 will show simulation results on how this method works and compare the results with kernel density deconvolution. Section 5 will be describing computational issues and some concluding remarks.

2 Hermite Polynomials

Hermite polynomials are a set of orthogonal polynomials with a recursive property. They were first defined by Pierre-Simon Laplace in the early 1800's (Laplace (1820)), but it is more recognised from Hermite's work in 1864 (Hermite (1864)). For our work, we will be defining $H_{k-1}(\cdot)$ as the k^{th} term in the Hermite polynomials. The first two terms of this polynomial are $H_0(x) = 1$ and $H_1(x) = 2x$, and for the rest of the Hermite polynomials, each term will have one power higher than the previous term and can be calculated using the formula $H_{k+1}(x) = 2xH_k(x) - 2kH_{k-1}(x)$. Here we will display the first five Hermite polynomials:

$$\begin{aligned} H_0(x) &= 1, \\ H_1(x) &= 2x, \\ H_2(x) &= 4x^2 - 2, \\ H_3(x) &= 8x^3 - 12x, \\ H_4(x) &= 16x^4 - 48x^2 + 12. \end{aligned}$$

Figure 1 shows in a graphical manner what the first five terms of the Hermite polynomials look like, where k determines which term of the Hermite polynomial each curve represents.

We then rescale the Hermite polynomials to allow an orthonormal property, this transition from orthogonal to orthonormal will be useful to us when we apply the set Hermite

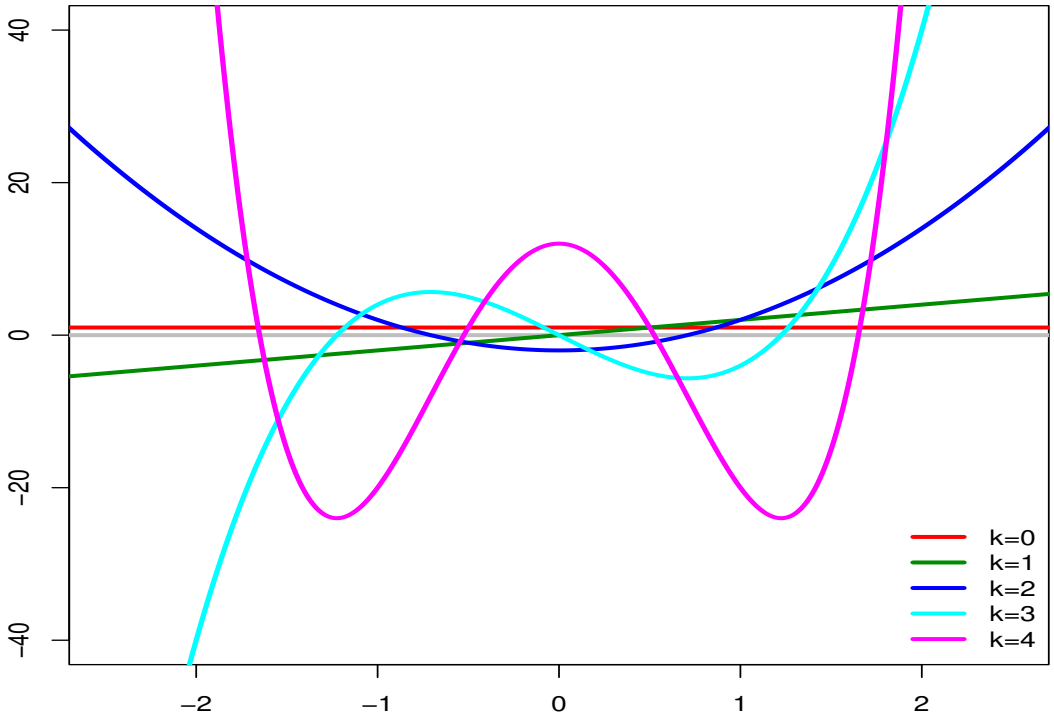


Figure 1: The first 5 terms of Hermite polynomials.

polynomials as part of a formula for unknown densities, we will explain how this helps a little bit later. Let $p_{k-1}(x)$ be the k^{th} term of the rescaled Hermite polynomials, defined as

$$p_k(x) = (\sqrt{\pi} k! 2^k)^{-1/2} H_k(x) e^{-x^2/2}, \tag{6}$$

With this rescaling, we see the first term of the rescaled Hermite polynomial is now a standard normal distribution, and the first five terms of the rescaled Hermite polynomial can be shown through figure 2.

Then we can say that this rescaled Hermite polynomial series has the following two orthonormal properties: $\int_{-\infty}^{\infty} p_k(x) p_j(x) dx = 0$ for $j \neq k$ and $\int_{-\infty}^{\infty} p_k^2(x) dx = 1$. These two properties will be crucial when we apply constraints on any unknown densities that will be represented as functions of $p_k(x)$.

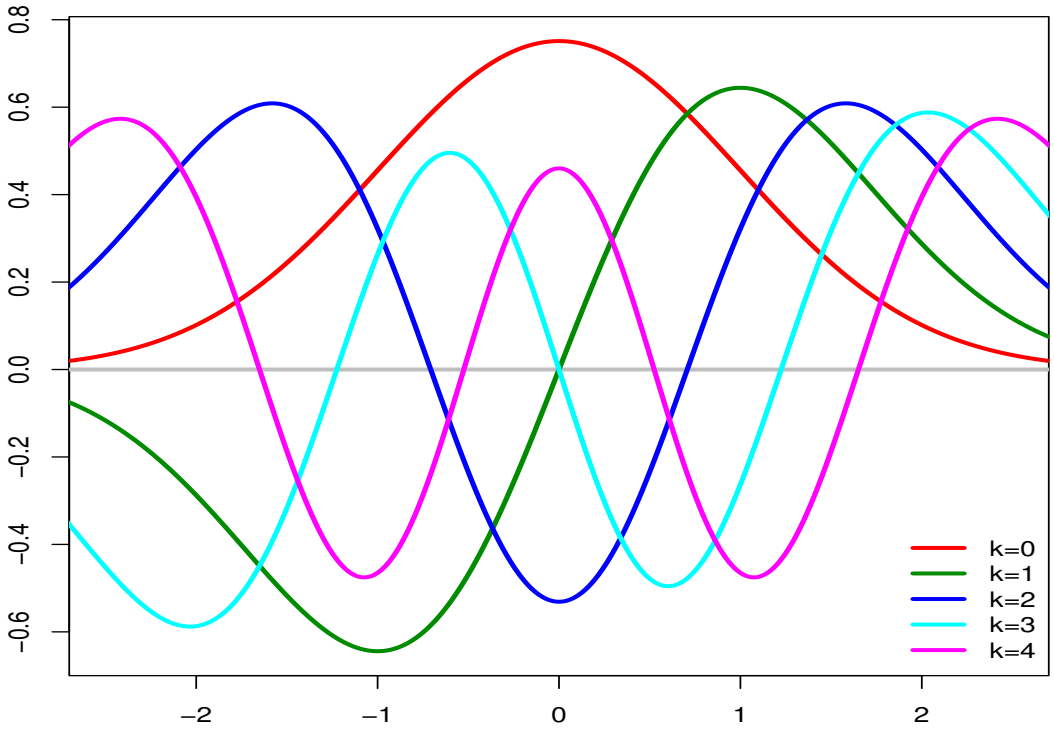


Figure 2: The first 5 terms of rescaled Hermite polynomials such that the polynomials are now orthonormal.

2.1 How to express an unknown density using Hermite polynomial

For any density where we do not assume its distribution shape or any of its parameters, we can approximate this density function $f(x)$ as

$$f(x) = \{\sum_{k=0}^{K} \theta_k p_k(x)\}^2, \tag{7}$$

here θ_k is the corresponding coefficient for $p_k(x)$, and it will be these coefficients that determine the shape of density curve $f(x)$. To determine how many polynomials we will be using to estimate the density $f(x)$, we compare the BIC value of $f(x)$ as the value of K changes, and take the K with minimal BIC value. This process in explain in more detail with example in a future section.

For any density we have the properties that the density has to be non-negative and integrates to 1. The non-negativity problem is solved naturally from the squared formula in 7. As for the other functional restriction that the density needs to integrate to 1, we apply the orthonormal properties $\int p_k^2(x)dx = 1$ and $\int p_k(x)p_j(x)dx = 0$ for $j \neq k$ that was

mentioned previously:

$$\begin{aligned}
\int f(x)dx &= \int \Big\{ \sum_{k=0}^{K} \theta_k p_k(x) \Big\}^2 dx, \\
&= \int \Big[\sum_{k=0}^{K} \theta_k^2 p_k(x)^2 + 2 \sum_{k=0}^{K-1} \sum_{j=k+1}^{K} \theta_k \theta_j p_k(x) p_j(x) \Big] dx, \\
&= \sum_{k=0}^{K} \theta_k^2 \int p_k^2(x)dx + 2 \sum_{k=0}^{K-1} \sum_{j=k+1}^{K} \theta_k \theta_j \int p_k(x) p_j(x)dx, \\
&= \sum_{k=0}^{K} \theta_k^2.
\end{aligned}$$

Therefore this leads to a simple coefficient constraint:

$$\int f(x)dx = 1 \rightarrow \sum_{k=0}^{K} \theta_k^2 = 1. \tag{8}$$

Figure 3 shows a few examples of how we can use Hermite polynomials to obtain density curves. Four curves were produced where for each curve we used only the first three terms of the Hermite polynomial. Given that the highest power of these three rescaled Hermite polynomial is the power of 4, we can only produce some simple density curves. As we increase the number terms used in function 7, we can produce density curves that are more complex with more peaks. In this figure, the coefficient used for producing each density curve is listed in the legend of the figure, these coefficients are not chosen for any specific reason other than being able to produce these specific curves that we desire. In the figure, we have shown that using Hermite polynomials as a media to obtain density curves, we can produce curves that are symmetric, left-skewed, right-skewed and even curves with multiple peaks, as long as the coefficient for each density follows the constraint in equation 8.

Now in this chapter, we are only exploring how to obtain the density of T when we give assumptions to the density of the error term U, but we did mention that we wish to relax this assumption in the future. When this happens, any error term U that will not have assumptions will need to also be represented as the function 7. In that case, one more constraint is required for the density of U, and that is the density will have mean 0 ($\int xf(x)dx = 0$).

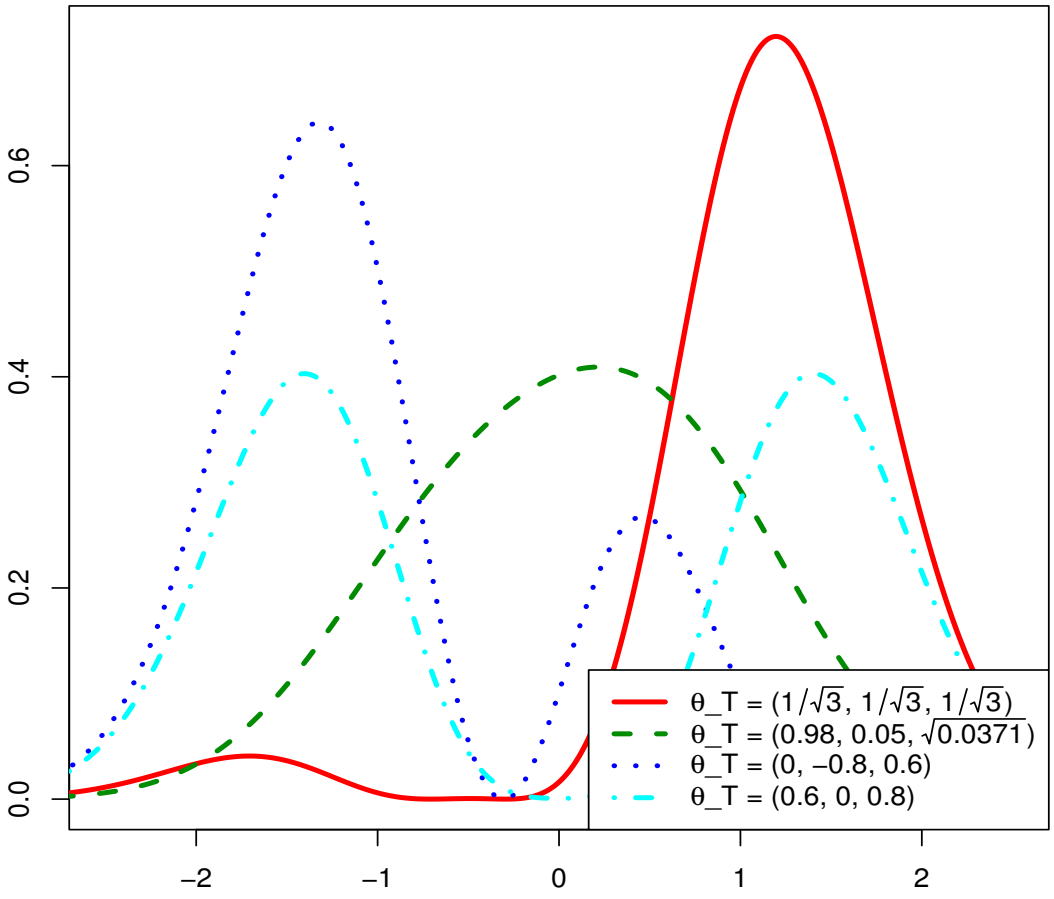


Figure 3: A set of examples showing how changing the coefficient for each polynomial effects the shape of the density curve. All densities are calculated with using only the first three polynomials of the rescaled Hermite ($p_k(x)$, $k = 0, 1, 2$), and the legend list the coefficient for each polynomial for each density curve.

Using the recursive property of Hermite polynomials $H_{k+1}(x) = 2xH_k(x) - 2kH_{k-1}(x)$, and formula 6, we can derive:

$$\sqrt{2(k+1)}p_{k+1}(x) = 2xp_k(x) - 2\sqrt{k/2}p_{k-1}(x). \tag{9}$$

For now, let us look at four different scenarios:

1. $\int \sqrt{2(k+1)}p_{k+1}(x)^2dx$

   Using the orthonormal properties of $p_k(x)$, we get $\int \sqrt{2(k+1)}p_{k+1}(x)^2dx = \sqrt{2(k+1)}$.

   Using equation 9, we get

$$\begin{aligned}\int \sqrt{2(k+1)}p_{k+1}(x)^2dx &= \int p_{k+1}(x)\Big[2xp_k(x) - 2\sqrt{k/2}p_{k-1}(x)\Big]dx,\\ &= 2\int xp_{k+1}(x)p_k(x)dx - 2\sqrt{k/2}\int p_{k+1}(x)p_{k-1}(x)dx,\\ &= 2\int xp_{k+1}(x)p_k(x)dx.\end{aligned}$$

   By combining the two answers, we therefore get $2\int xp_{k+1}(x)p_k(x)dx = \sqrt{2(k+1)} \rightarrow \int xp_{k+1}(x)p_k(x)dx = \sqrt{(k+1)/2}$.

2. $\int \sqrt{2(k+1)}p_{k+1}(x)p_k(x)dx$

   Using the orthonormal properties of $p_k(x)$, we get $\int \sqrt{2(k+1)}p_{k+1}(x)p_k(x)dx = 0$.

   Using equation 9, we get

$$\begin{aligned}\int \sqrt{2(k+1)}p_{k+1}(x)p_k(x)dx &= \int p_k(x)\Big[2xp_k(x) - 2\sqrt{k/2}p_{k-1}(x)\Big]dx,\\ &= 2\int xp_k(x)^2dx - 2\sqrt{k/2}\int p_k(x)p_{k-1}(x)dx,\\ &= 2\int xp_k(x)^2dx.\end{aligned}$$

   By combining the two answers, we therefore get $\int xp_k(x)^2dx = 0$.

3. $\int \sqrt{2(k+1)}p_{k+1}(x)p_{k-1}(x)dx$

   Using the orthonormal properties of $p_k(x)$, we get $\int \sqrt{2(k+1)}p_{k+1}(x)p_{k-1}(x)dx = 0$.

Using equation 9, we get

$$\begin{aligned}\int \sqrt{2(k+1)}p_{k+1}(x)p_{k-1}(x)dx &= \int p_{k-1}(x)\Big[2xp_k(x) - 2\sqrt{k/2}p_{k-1}(x)\Big]dx,\\ &= 2\int xp_k(x)p_{k-1}(x)dx - 2\sqrt{k/2}\int p_{k-1}^2(x)dx,\\ &= 2\int xp_k(x)p_{k-1}(x)dx - 2\sqrt{k/2}.\end{aligned}$$

By combining the two answers, we therefore get $\int xp_k(x)p_{k-1}(x)dx = \sqrt{k/2}$.

4. $\int \sqrt{2(k+1)}p_{k+1}(x)p_{k-2}(x)dx$

Using the orthonormal properties of $p_k(x)$, we get $\int \sqrt{2(k+1)}p_{k+1}(x)p_{k-2}(x)dx = 0$.

Using equation 9, we get

$$\begin{aligned}\int \sqrt{2(k+1)}p_{k+1}(x)p_{k-2}(x)dx &= \int p_{k-2}(x)\Big[2xp_k(x) - 2\sqrt{k/2}p_{k-1}(x)\Big]dx,\\ &= 2\int xp_{k-2}(x)p_k(x)dx - 2\sqrt{k/2}\int p_{k-2}(x)p_{k-1}(x)dx,\\ &= 2\int xp_{k-2}(x)p_k(x)dx.\end{aligned}$$

By combining the two answers, we therefore get $\int xp_{k-2}(x)p_k(x)dx = 0$.

Now, from these four scenarios, we can see that $\int xp_k(x)p_j(x)dx$ only yeilds results when k and j are right next to each other, that is $k = j+1$. Also $\int xp_k(x)p_j(x)dx = \sqrt{j/2}$ when $j = k+1$ and becomes 0 for all other cases.

For the density mean:

$$\begin{aligned}\int xf(x)dx &= \int x\Big\{\sum_{k=0}^{K}\theta_k p_k(x)\Big\}^2 dx,\\ &= \int x\Big[\sum_{k=0}^{K}\theta_k^2 p_k(x)^2 + 2\sum_{k=0}^{K-1}\sum_{j=k+1}^{K}\theta_k\theta_j p_k(x)p_j(x)\Big]dx,\\ &= \sum_{k=0}^{K}\theta_k^2\int xp_k(x)^2dx + 2\sum_{k=0}^{K-1}\sum_{j=k+1}^{K}\theta_k\theta_j\int xp_k(x)p_j(x)dx,\\ &= 2\sum_{k=0}^{K-1}\theta_k\theta_{k+1}\sqrt{(k+1)/2},\\ &= \sqrt{2(k+1)}\sum_{k=0}^{K-1}\theta_k\theta_{k+1}.\end{aligned}$$

When there is a need for a mean 0 restriction on an unknown density, we can then use the constraint:

$$\int xf(x)dx = 0 \rightarrow \sqrt{2(k+1)}\sum_{k=0}^{K-1}\theta_k\theta_{k+1} = 0. \tag{10}$$

3 Methodology

Assume that we only have one observation per individual of variable W and this variable is the only variable that is observed, also we assume that the distribution of the error term U is known. We express the error model as

$$W_i = T_i + U_i, \tag{11}$$

here i represents the i^{th} subject, where $i = 1, ..., n$.

Let $f_W(\cdot)$, $f_T(\cdot)$ and $f_U(\cdot)$ be the density functions corresponding to the variables W, T and U. The density of W for each individual subject can be expressed via the integral equation

$$f_W(w_i) = \int f_T(t_i)f_U(w_i - t_i)dt_i. \tag{12}$$

Following Shen (1997), we will approximate any unknown density, in this case $f_T(\cdot)$ as a sum of basis functions using the first $K+1$ terms of the series of Hermite polynomials where we have expressed in the 7. Let $\theta_T = (\theta_{1T}, ..., \theta_{KT})$ be the set of coefficients for density $f_T(\cdot)$.

For any given $K+1$ number of polynomials, the coefficients θ_T can be estimated by maximizing the log of the following likelihood for each subject i

$$\mathcal{L}_i(f_T|W_i, \theta_{kT}) = \int \left\{\sum_{k=0}^{K}\theta_{kT}p_k(t_i)\right\}^2 f_U(w_i - t_i)dt_i, \tag{13}$$

subject to constraint 8 that the unknown density $f_T(\cdot)$ have to integrate to 1. Of course, K, the number of basis functions, is also a variable that needs to be estimated.

As calculating equation 13 analytically can be challenging, we need to find a method that can make the computation process easier. To approximate the likelihood (13), we use Gauss-Chebyshev quadrature estimation (Gauss (1815)). The likelihood is now simultaneously

approximated by a finite sieve space (Shen (1997)). Equation (13) can be rewritten as

$$\mathcal{L}_i(f_T|W_i,\theta_{kT}) \approx \frac{(W_L - W_S)}{2}\sum_{m=1}^{M}\frac{\pi}{M}\sqrt{1-s_m^2}\left\{\sum_{k=0}^{K}\theta_{kT}p_k(l_m)\right\}^2 f_U(W_i - l_m), \quad (14)$$

here s_m is the Gauss-Chebyshev approximation with range [-1,1]: $s_m = \cos\{(2m-1)\pi/(2M)\}$, l_m is the scaled Gauss-Chebyshev approximation of s_m with range $[W_S, W_L]$: $l_m = s_m(W_L - W_S)/2 + (W_L + W_S)/2$. Here $[W_S, W_L]$ is the smallest and largest number in the observed variable W respectively, and M is the number of nodes for Gauss-Chebyshev approximation. The final objective is to obtain an estimate of the density of T, by optimising the problem to jointly find the set of coefficient parameters such that $\sum_{i=1}^{n}\log\{\mathcal{L}_i(f_T|W_i,\theta_{kT})\}$ is maximised, subject to the constraint $\sum_{k=0}^{K}\theta_k^2 = 1$. To estimate the density coefficients θ_k we will use an optimisation program that allows non-linear constraints. In R, we will be using an existing package "NLopt" which is designed specifically for calculating and optimising results with non-linear constraints.

4 Simulations

In the previous section, it was mentioned that K is also a variable that needs to be estimated. In this section, we start with showing how K is chosen, then we consider scenarios for the simulation study where each scenario is compared with the kernel density deconvolution (KDD) method.

KDD was developed and investigated by Carroll and Hall (1988) and Stefanski and Carroll (1990). It uses characteristic functions to analysis nonparametric measurement error models, and is considered to obtain good estimates for the distribution of T (Fan, 1992; Wand, 1998). The package used for KDD is developed in the paper Achilleos and Delaigle (2012).

For this simulation, we will specify the same information into "fDKDE" and our software package for Hermite polynomial deconvolution (HePD).

4.1 HePD: choosing the number of smoothing parameters

For the HePD method, it requires specification on the number of smoothing parameters (K) used. We will use the Bayesian information criterion (BIC) to determine the optimal number

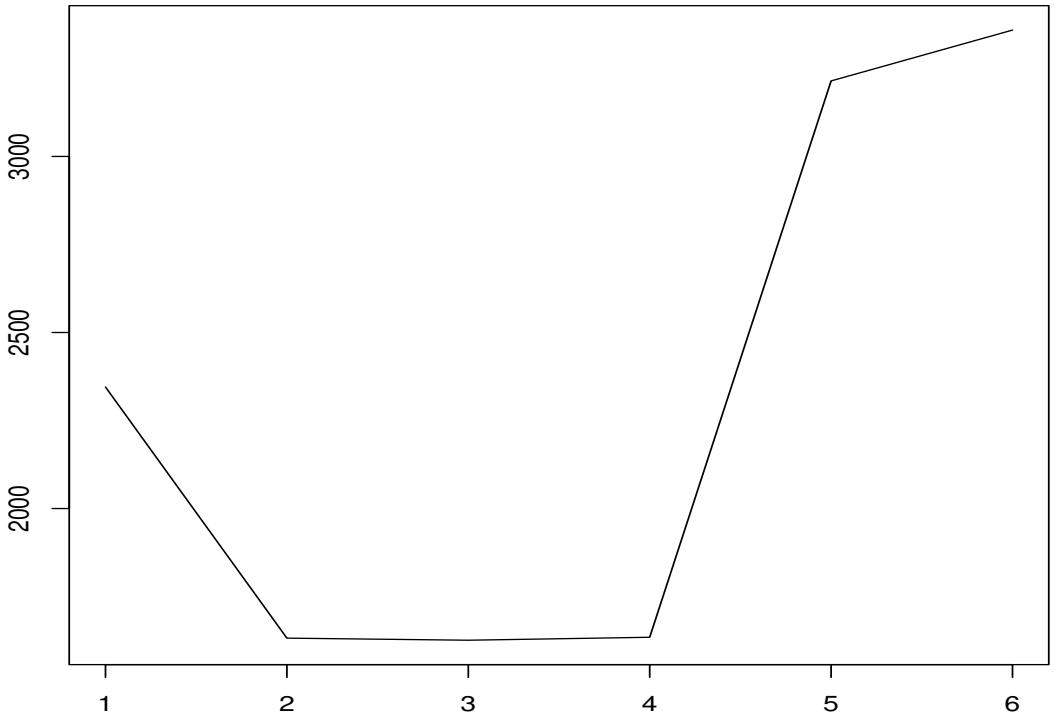


Figure 4: With $T \sim N(0,1))$ and $U \sim Lap(0, 1/\sqrt{3})$, we see the change in BIC value as we estimate the density using increasing number of smoothing parameters (from $K = 1$ to $K = 6$). In this estimation, the sample size for each estimation is 500 with only one set of W and the assumption that U has the correct distribution type.

of smoothing parameters for every simulation. In this section, all simulations have a sample size of 500. Figure 4 shows the BIC value for just one simulation of this sample size given the different number of smoothing parameters, BIC determines that the optimal number of smoothing parameters would be around 3 to 5 ($K = 2$ to 4).

4.2 Generating observed variable W for simulation

As mention previously, observed variable W is the only variable that we will have access to when it comes real applications. But for the following simulations, W will not be the only variable that we know, in fact we will be using information from T and U to obtain this simulated W variable.

At the start of each simulation, we assign a density to T and U where we known the distribution type and all the parameters needed. We then generate a set of n values from

the density of T and U respectively. Following the function $W = T + U$, we add the set of generated T values with the set of generated U values to create the simulated observed variable W with n inputs.

The reason we choose to generate the variable W this way is because the final result we aim to obtain through the analysis is an estimation of the density f_T, and having an pre-existing knowledge of what the true density looks like allows us for a more straight forward comparison. Keep in mind that we do not use this pre-existing knowledge of the true f_T through out the analysis process, we will still be treating the density of T as an unknown density. This true density of T will only be used in the results as an comparison.

4.3 Assuming known parameters for U

For this subsection, two types of distribution is considered for T:

- a standard normal distribution: $T \sim \text{Normal}(0, 1)$
- a gamma distribution: $T \sim \Gamma(9, 1/3)$, where the shape parameter for this gamma distribution is 9 and the scale parameter is $1/3$

Also two types of distribution is considered for U:

- a Laplace distribution: $U \sim \text{Laplace}(0, 1/\sqrt{3})$
- a normal distribution: $U \sim \text{Normal}(0, 1/\sqrt{3})$

For both cases of U above, we have the mean and standard deviation for each distribution respectively. We will be looking at all four combinations using the previously mentioned distribution types for T and U.

For each combination, we perform 350 simulations. And for each simulation, the sample size is $n = 500$. For estimating the density of T, the number of smoothing parameters is determined using BIC, we concluded that in most cases $K = 2$ gives the lowest BIC results, this indicates that using three basis functions for most examples will give optimal estimations.

Figure 5, 8, 11 and 14 will be showing the estimated densities of T for the four examples. Three lines will be shown in each plot, where the black solid line is the true density of T which we used in the simulation to generate observed variable W, red dashed line is the average estimated density curve of T from the 350 simulations using HePD method, and blue dashed line is the average estimated density curve of T using KDD from the 350 simulations. For both HePD method and KDD method, the information given in order to estimate the density of T are: the observed variable W, the correct distribution shape for U and the true standard deviation for error term U.

Figure 5 looks at the case where the true T is a normal distribution and true U has a Laplace distribution. Given that we known the true information of error term U. The average density curve estimated using method KDD has a larger variance compared to the true density of T, but captures the correct center and also the correct shape of T. For the method HePD, the average density curve also has a larger variance compared to the true density of T, it also captures the correct center, compared to KDD, the peak of the curve is closer to the true peak, but the curve is tri-modal. We also compared the mean squared error (MSE) and the mean absolute error (MAE) for both methods to the true density of T, this is calculated by comparing the estimated density value at each grid point to its corresponding true density value. Figure 6 looks at a boxplot of the mean squared error values and figure 7 looks at a boxplot of the mean absolute error values. For both figures, the boxplot on the left is the values from method KDD and the boxplot on the right is the values from method HePD. We see that in general KDD has slightly smaller error values, and it is more stable between simulations than HePD.

Figure 8 looks at the case where both the true T and the true U are normally distributed. Once again, both KDD and HePD methods on average have estimated a larger variance compared to the true T. Similar to the previous example, both methods have estimated the correct center, HePD is able to estimate the peak of the curve better, but not so much on the shape since the red dashed curve still has a slight tri-modal shape. Once again we look at the MSE and MAE values from both methods. Figure 9 looks at the MSE values and figure 10 are the MAE values. In this example the median MSE and MAE value for HePD is slightly smaller but KDD is more stable between simulations.

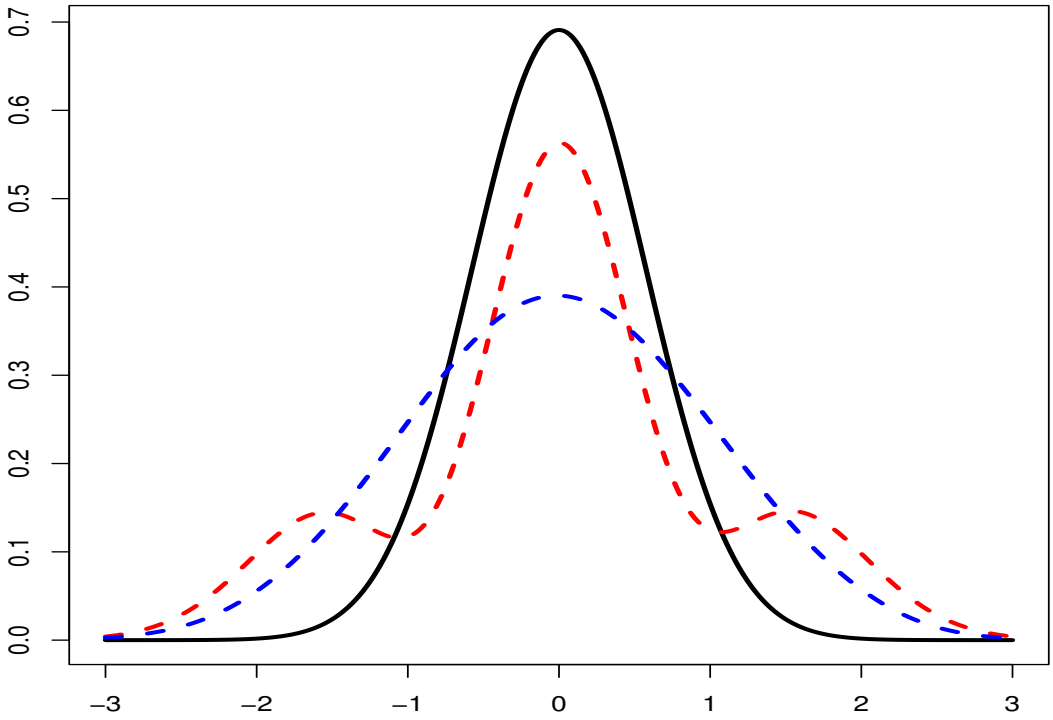


Figure 5: Comparing the average density estimate of T using method KDD(blue line) and the average density estimate of T using Hermite polynomials(red line) of 350 simulations to the true density of T(black line), where T has a Normal distribution($T \sim N(0,1)$) and U has a Laplace distribution ($U \sim Lap(0, 1/\sqrt{3})$).

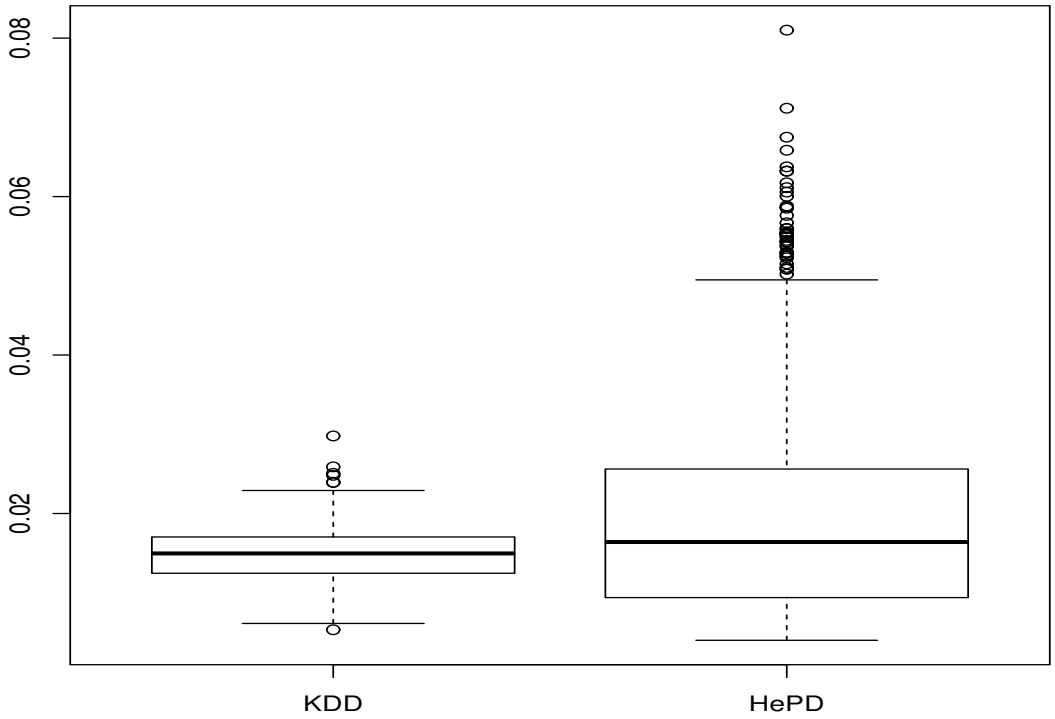


Figure 6: Comparing the mean squared error between method KDD and method HePD, where T has a Normal distribution($T \sim N(0,1)$) and U has a Laplace distribution ($U \sim Lap(0, 1/\sqrt{3})$).

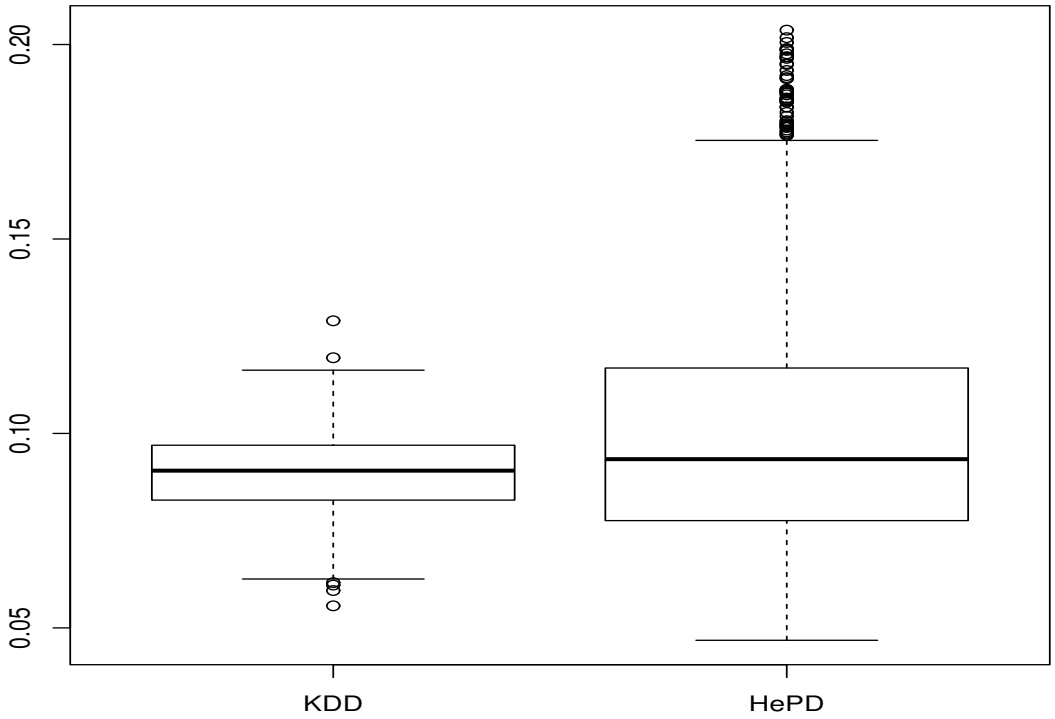


Figure 7: Comparing the mean absolute error between method KDD and method HePD, where T has a Normal distribution($T \sim N(0,1)$) and U has a Laplace distribution ($U \sim Lap(0, 1/\sqrt{3})$).

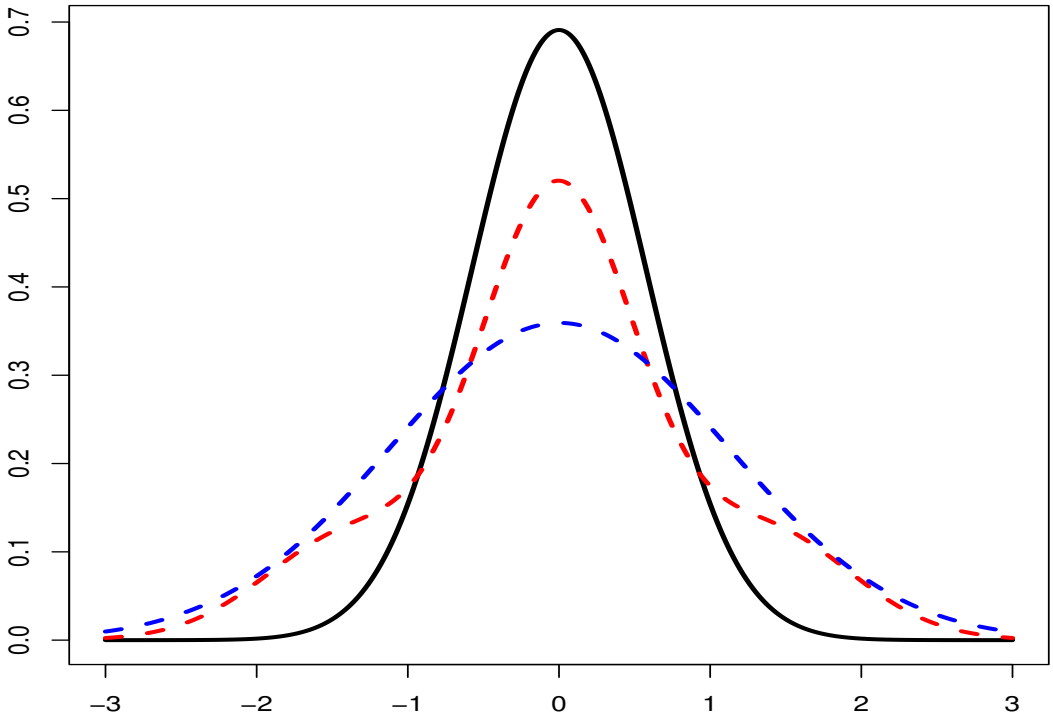


Figure 8: Comparing the average density estimate of T using method KDD(blue line) and the average density estimate of T using Hermite polynomials(red line) of 350 simulations to the true density of T(black line), where T has a Normal distribution($T \sim N(0,1)$) and U has a Normal distribution ($U \sim N(0, 1/\sqrt{3})$).

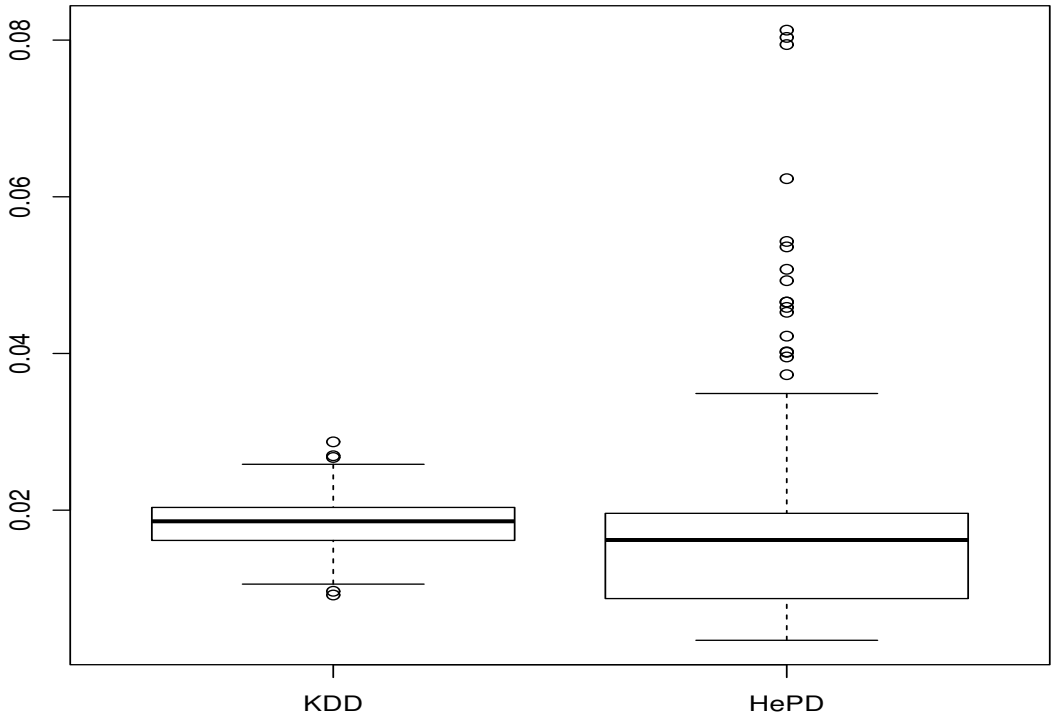


Figure 9: Comparing the mean squared error between method KDD and method HePD, where T has a Normal distribution($T \sim N(0,1)$) and U has a Normal distribution ($U \sim N(0, 1/\sqrt{3})$).

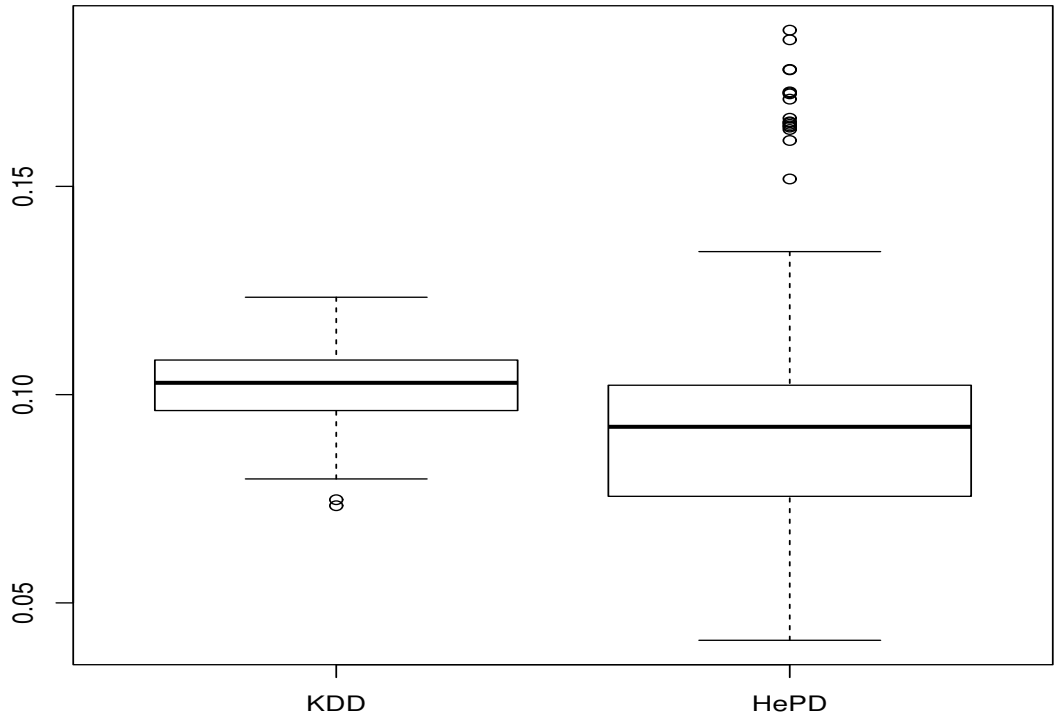


Figure 10: Comparing the mean absolute error between method KDD and method HePD, where T has a Normal distribution($T \sim N(0,1)$) and U has a Normal distribution ($U \sim N(0, 1/\sqrt{3})$).

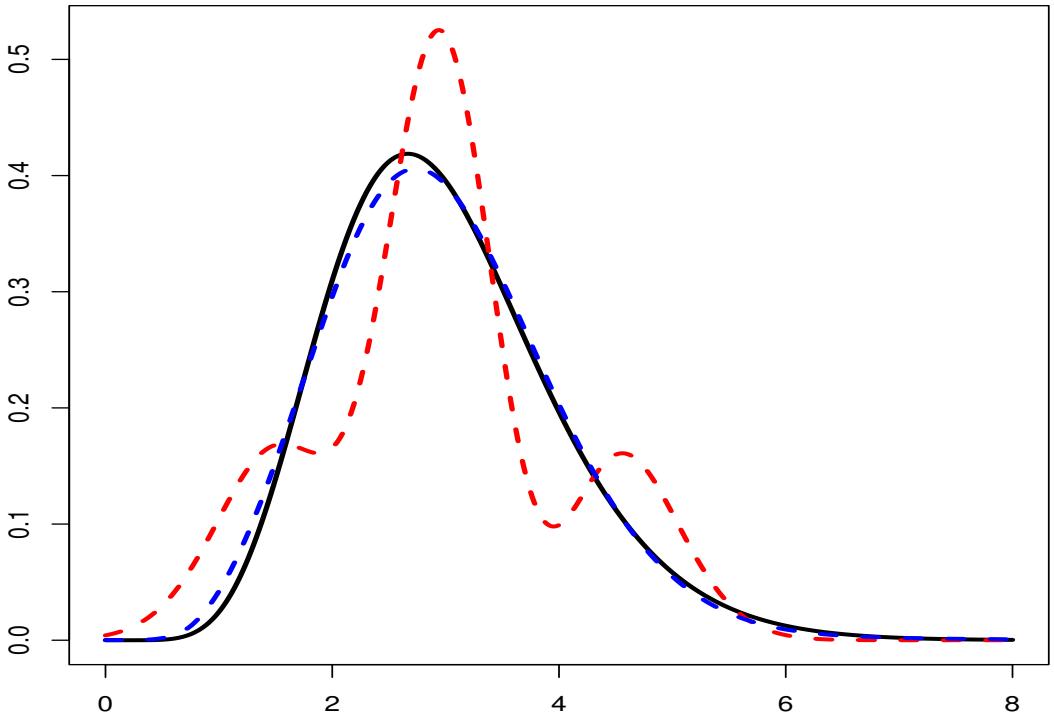


Figure 11: Comparing the average density estimate of T using method KDD(blue line) and the average density estimate of T using Hermite polynomials(red line) of 350 simulations to the true density of T(black line), where T has a Gamma distribution($T \sim \Gamma(9, 1/3)$) and U has a Laplace distribution($U \sim Lap(0, 1/\sqrt{3})$).

Figure 11 looks at the case where T has a Gamma distribution and U has a Laplace distribution. We can see that the averaged density curve estimated using KDD method is almost a perfect match to the true density curve, where as the HePD method on average estimated a curve that is tri-modal with a slightly higher peak. Figure 12 and 13 also confirms that KDD has a lower MSE and MAE value and is also more stable.

Figure 14 looks at where T has a Gamma distribution and U has a normal distribution. In this case, the averaged density curve estimated from both KDD and HePD method are quite close to the true density curve. Figure 15 and 16 also shows that the average MSE and MAE value for both methods are very similar, though once again KDD gives more stable result.

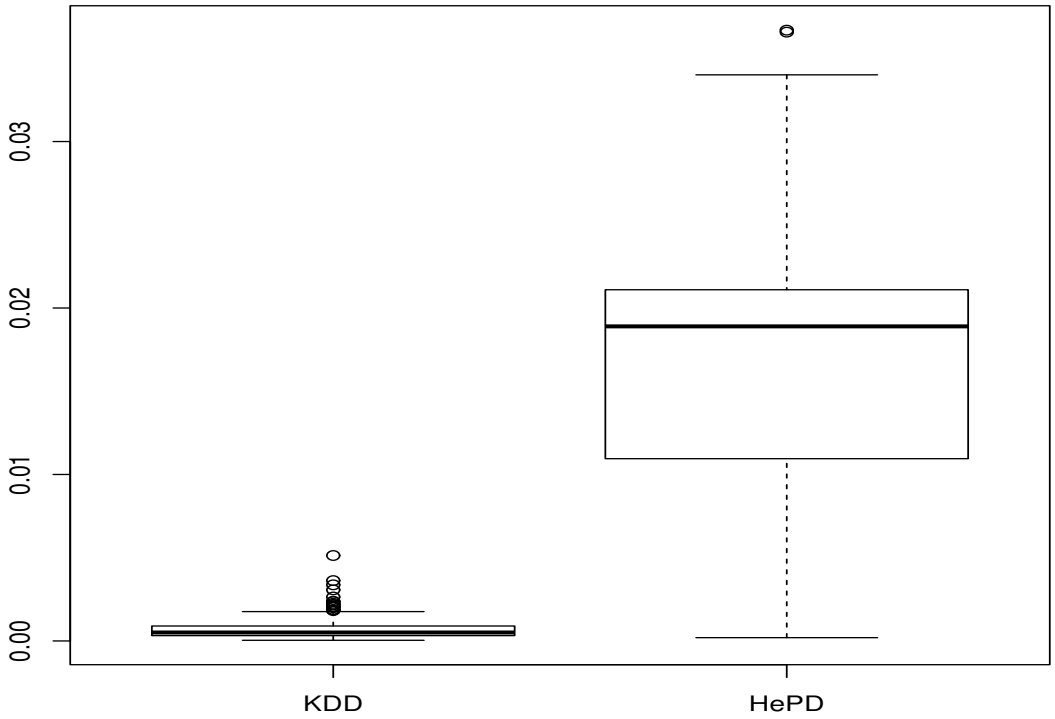


Figure 12: Comparing the mean squared error between method KDD and method HePD, where T has a Gamma distribution($T \sim \Gamma(9, 1/3)$) and U has a Laplace distribution ($U \sim Lap(0, 1/\sqrt{3})$).

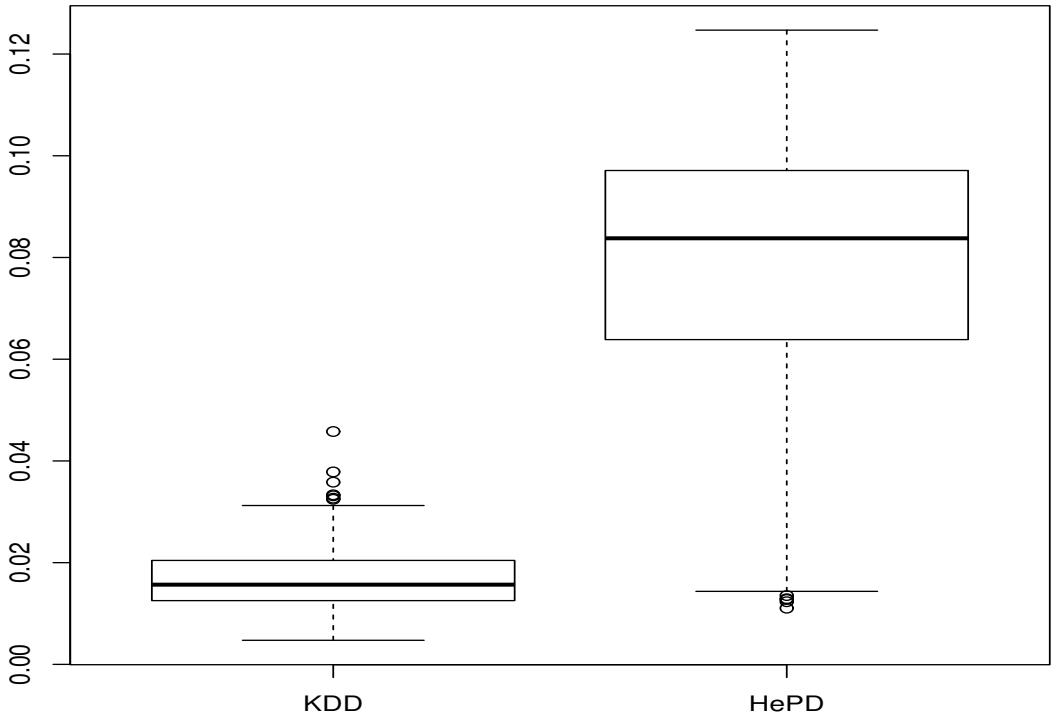


Figure 13: Comparing the mean absolute error between method KDD and method HePD, where T has a Gamma distribution($T \sim \Gamma(9, 1/3)$) and U has a Laplace distribution ($U \sim Lap(0, 1/\sqrt{3})$).

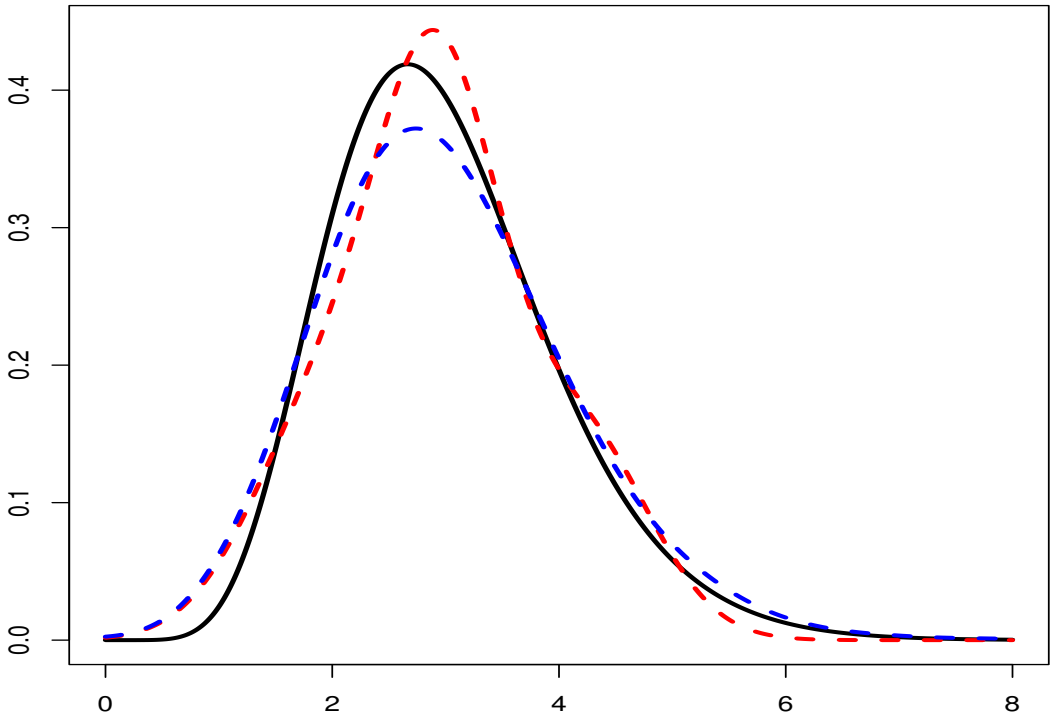


Figure 14: Comparing the average density estimate of T using method KDD(blue line) and the average density estimate of T using Hermite polynomials(red line) of 350 simulations to the true density of T(black line), where T has a Gamma distribution($T \sim \Gamma(9, 1/3)$) and U has a Normal distribution($U \sim N(0, 1/\sqrt{3})$).

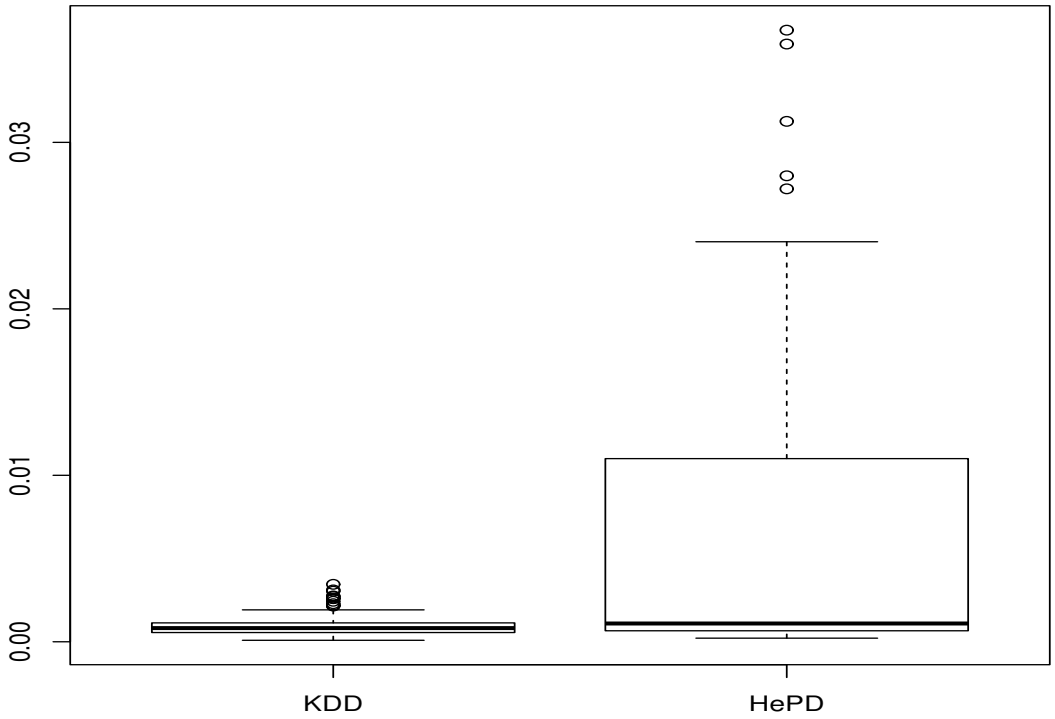


Figure 15: Comparing the mean squared error between method KDD and method HePD, where T has a Gamma distribution($T \sim \Gamma(9, 1/3)$) and U has a Normal distribution($U \sim N(0, 1/\sqrt{3})$).

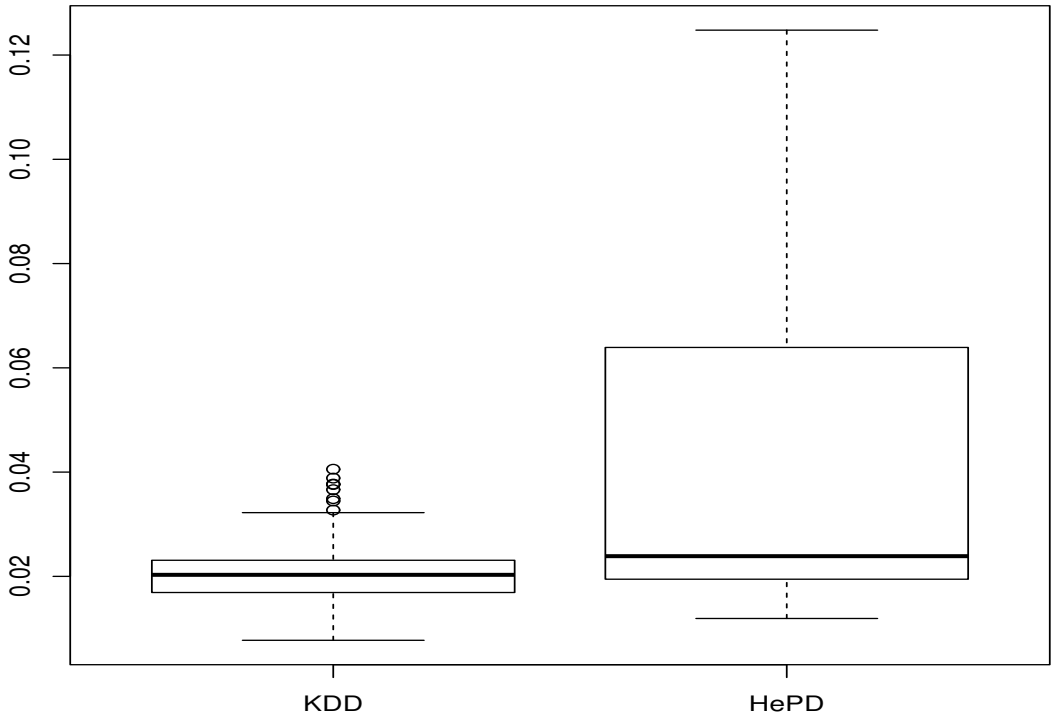


Figure 16: Comparing the mean absolute error between method KDD and method HePD, where T has a Gamma distribution($T \sim \Gamma(9, 1/3)$) and U has a Normal distribution($U \sim N(0, 1/\sqrt{3})$).

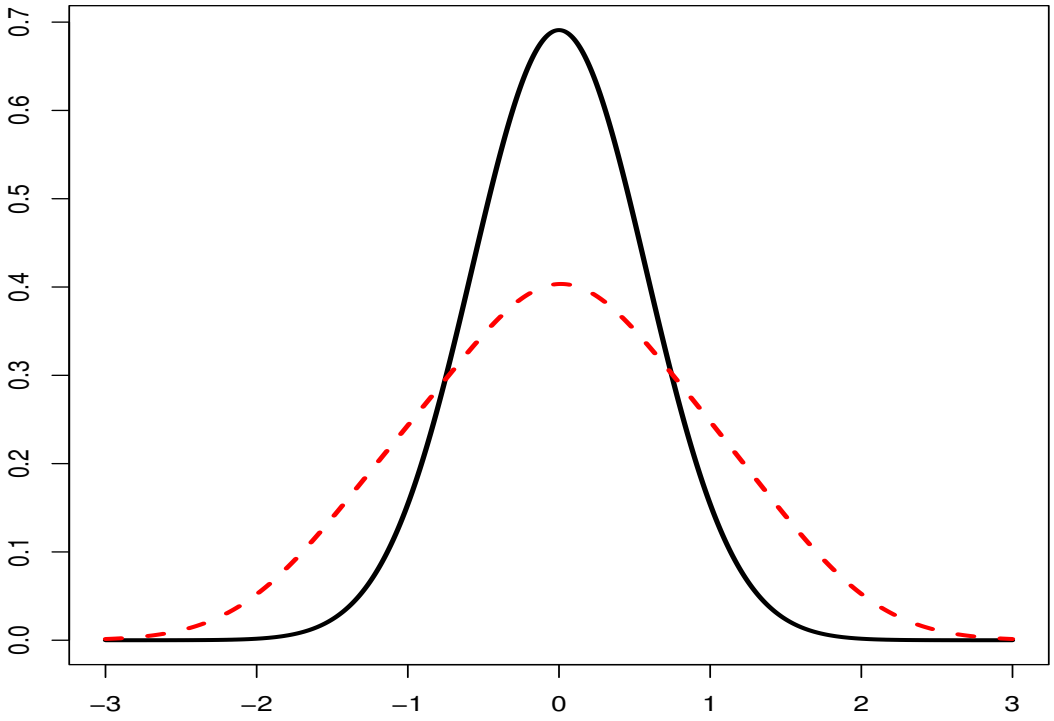


Figure 17: Comparing the estimated density of T (red dashed) with the true density curve(black solid, normal distribution), where U is assumed as a Laplace distribution with an estimated standard deviation.

4.4 Estimating both the density of T and the parameters of U

In the previous subsection, we looked at the cases where we only estimate the density of T by assuming the correct information on both the distribution type for U and also its standard deviation. For this subsection, we will extend the simulations by estimating both the smoothing parameters for density of T and the standard deviation for U. We will still be looking at the same four examples as the previous subsection. For this case, there is no existing R code for method KDD, therefore we will only be looking at how well method HePD estimates the density curves and comparing them to the true density of T.

Figure 17 looks at the case where T has a normal distribution and U has a Laplace distribution. The red dashed curve is the averaged density estimation curve for T, and it seems to have a larger variance compared to the true density curve, but it has captured the correct center and shape of the density.

Figure 18 looks at the case where T is still normally distributed, but U this time is also

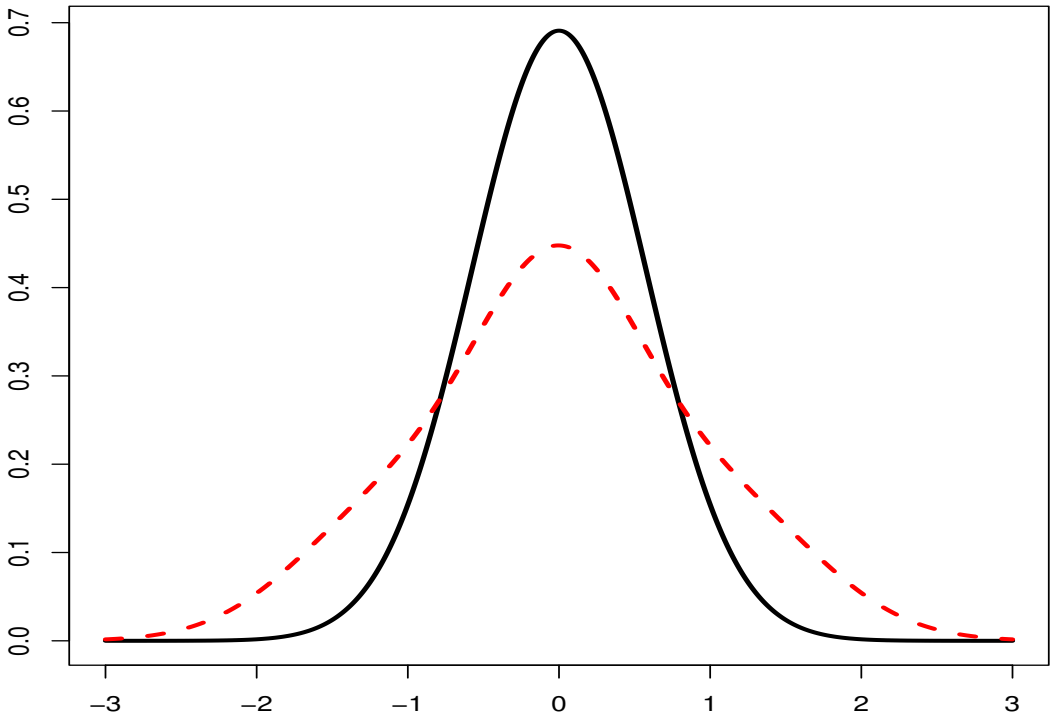


Figure 18: Comparing the estimated density of T (red dashed) with the true density curve(black solid, normal distribution), where U is assumed as a Normal distribution with an estimated standard deviation.

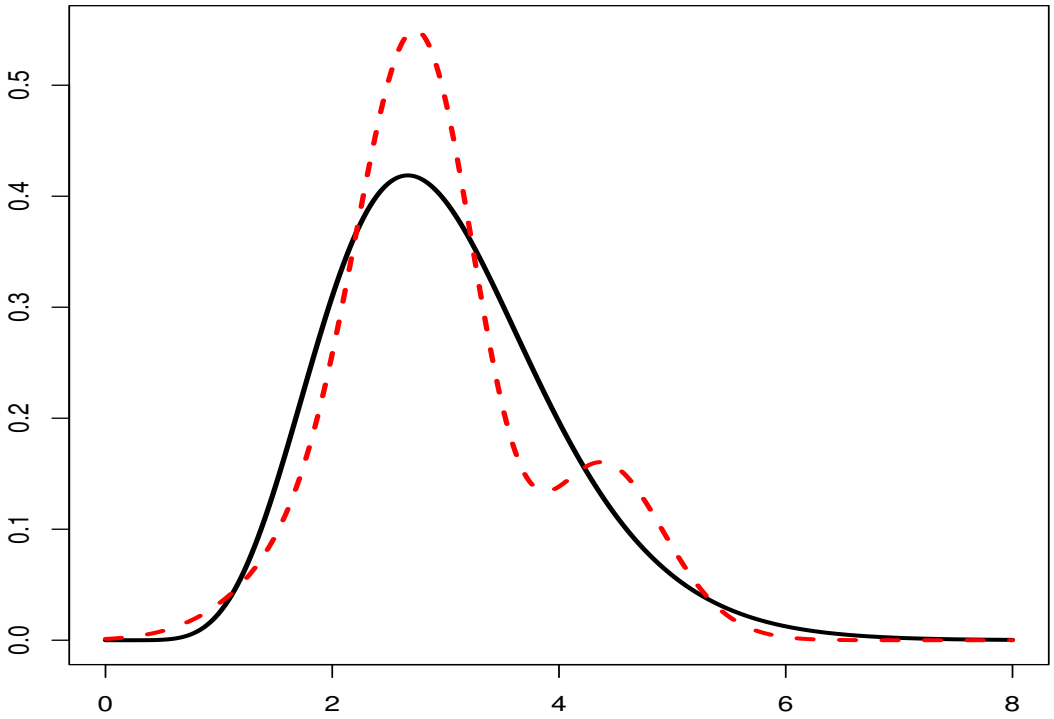


Figure 19: Comparing the estimated density of T (red dashed) with the true density curve(black solid, gamma distribution), where U is assumed as a Laplace distribution with an estimated standard deviation.

normally distributed. It seems that on average, the estimated density has a slight larger variance than the true density, but correct center and shape.

Figure 19 looks at the case where the true T has a gamma distribution and the true U has a Laplace distribution. In this example, it seems that the averaged density estimation is skewed towards to correct direction, but has a much smaller variance. The averaged estimated density curve is also bi-modal. The red dashed curve is an average of all 350 density estimations, but this does not mean that all the density estimations are bi-modal, it is more likely that in the optimisation process, the estimated smoothing coefficients ended up into two groups, one which formed curves that closely resemble the larger peak of the red curve, and a smaller group which formed curves which contains the smaller peak of the red curve.

Figure 20 estimates the density of T where the true T has a gamma distribution and U has a normal distribution. Similar to figure 19 the averaged estimated density of T is skewed

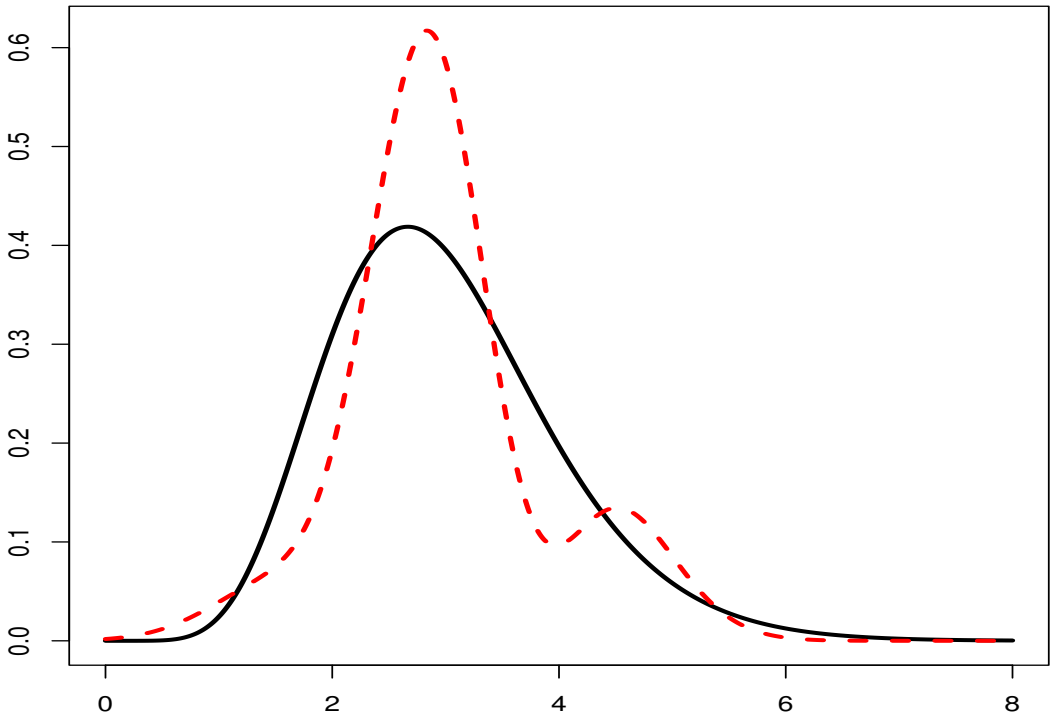


Figure 20: Comparing the estimated density of T (red dashed) with the true density curve(black solid, gamma distribution), where U is assumed as a Normal distribution with an estimated standard deviation.

in the correct direction, but has a smaller variance and is bi-modal. Once again, we do not believe the bi-modal curve occurs in most of the individually estimated densities, instead is a result of two groups with different optimised results.

5 Discussion

In section 4.3, we see that the estimated curve is not an accurate representation of the true distribution, one parameter we can change is the number of smoothing parameters K, so far through the calculation of BIC, it is determined that 2-3 smoothing parameters is the best, though we can increase the number of smoothing parameters if necessary. But we do need to keep in mind that an increase in K also increases the number of "peaks" in the plot, and the estimated density will become more wave like instead of a smooth curve. We also see that when there is only one set of observed variable W, by assuming the correct distribution error type and standard deviation for U, the KDD method seems to have more stable results than the proposed Hermite density deconvolution. But when comparing the averaged density curves, both HePD and KDD may obtain better results depending on the distribution of T. But we are also interested in how well we can estimate the density of the latent variable T if the problem is nonparametric, that is if we do not assume a distribution for the error term U. For a problem like this, it will be difficult to analyse the problem with only one replicate of W, which is why in the next chapter we will analyse how well we estimate the density of T nonparametrically for a continuous data using a simple classical error with multiple replicates of W and no additional information.

6 Computation details and errors

There are many optimisation packages in R, most packages work with non-constraint optimisation, and a few only allow linear constraints to be made for the optimisation. The R package we used to obtain the optimal smoothing parameters θ_{K_T} and θ_{K_U}(which will appear in the next chapter) is called "nloptr". This package allows constraint optimisation where the constraints can be either equality or inequality, it also allows the constraints to be either

linear or nonlinear. “nloptr” also allows the user the option to find local or global maximum or minimum.

Commands “constroptim.nl” and “auglag” from the R package “ALABAMA” can be used to find the optimal smoothing parameters. This is another R package which allows both equality and nonequality constraints and also nonlinear constraints. Unfortunately, for HePD, “ALABAMA” does not have the option to find the global maximum, and this resulted in inconsistent optimal parameter values.

Part V

Chapter 4: Semiparametric and Nonparametric Density Deconvolution for Continuous Data with Replicates

1 Introduction

In the previous section, we looked at how well we can estimate the density of a latent variable T when we have data on W and some assumption on the error term U. The results we obtained show that depending on the type of distribution used, some situations, KDD provide more favorable results, and in other situations, HePD provide more favorable results. And therefore we feel comfortable to use HePD as our method for further analysis.

Not all occasions will allow us insight into what the correct distribution type for U would be. This leads us to believe that there is a need to develop a methodology for estimating the density of T when both T and U are latent variables. For this chapter, we will be exploring the method by extending my research from the previous chapter, that is, we will still be using the HePD method. We do this by representing all unknown densities with the help of Hermite polynomials and we will be estimating all parameters with the use of maximising a log-likelihood function.

For this chapter, since both latent variables have no assumptions apart from allowing the mean of the error term to be zero, having only one replicate of observed variable W will not give us any result. We will need to have multiple replicates (≥ 2) in the observed variable W.

Unlike the previous chapter, we will not be comparing this method to any other existing methods. The reason is that there is no existing paper and program which will match this exact problem. In Hall and Ma (2007), let us suppose that we have replicates $W_{ij} = T_i + U_{ij}$ for $j = 1, 2$. If we define $\Delta W_i = (W_{i1} + W_{i2})/2$ and $\nabla W_i = (W_{i1} - W_{i2})/2$, and if U_{ij} is

symmetric, then $\Delta W_i = T_i + U_{i*}$, and ∇W_i has the same distribution as U_{i*}. They then estimate the distribution of U_{i*} by ordinary kernel density estimation of ∇W_i. They would then estimate the density of U_{i*} and then plug it into the ordinary density deconvolution. The problem is that in this case, there are two bandwidths: one for estimating the density of U_{i*}, and the other for the deconvolution kernel density estimate. No literature exists for how to choose the 2 bandwidths. No programs exist so far for this problem. Another paper Delaigle and Hall (2016) assume $W_i = T_i + U_i$, and that U_i is symmetric and the T_i is not symmetric. They do not need replicates, but they do need symmetric errors and non-symmetric latent variable T_i. Once again no programs are currently available for this problem. Delaigle and Hall (2016) also looked at the case where replicates are required, but again U_i has to have a symmetric density function as they stated that the characteristic function of U_i must be real and not vanish on the real line. Hu and Schennach (2008) show that identifiability is achieved if there are 3 or more replicates, even if the distribution of U is neither symmetric not homoscedastic. Sarkar et al. (2014) say that they believe that only 2 replicates are really needed. The understanding is that there are a number of papers in the deconvoluting kernel literature, they do not really give much advice on the selection of the bandwidth, require U to be symmetric, and no software exists for them. So, the problem in this chapter is unique in that (a) we do not assume that U is symmetric; and (b) that we are actually studying the properties of the density estimate of T.

For the next few sections, we will introduce the model again, this time the model will be specified to contain replicates for W, we will also look at what the likelihood for this new model would look like. We will then look at some simulations, and lastly an example using real-life applications.

2 methodology

For this chapter, we will express the simple classical error model as

$$W_{ij} = T_i + U_{ij}, \tag{15}$$

Staying consistent with the previous chapter, W is the observed variables, T is the latent variable and U is the error term corresponding to each replicate of W. Here $i = 1, ..., n$ indicates the number of subjects and $j = 1, ..., J$ represents the number of replicates, where $J \geq 2$. For this chapter, we will assume that each subject has the same number of replicates, the idea of having a different amount of replicates between subjects will be a topic for future discussion. Also we assume that each replicate is independent of each other, this is mainly due to the fact that there is a long period of space (3 months) between each repeated collection of recalls.

Once again we let $f_W(\cdot)$, $f_T(\cdot)$ and $f_U(\cdot)$ be the density functions corresponding to the variables W, T and U. The density of W for each individual subject can be expressed via the integral equation

$$f_W(w_i) = \int f_T(t_i) \prod_{j=1}^{J} f_U(w_{ij} - t_i) dt_i. \tag{16}$$

Since both variable T and variable U is unknown, we will approximate both densities using basis functions with Hermite polynomials. Here $f_T(t) = \{\sum_{k=0}^{K} \theta_{kT} p_k(t)\}^2$ and $f_U(u) = \{\sum_{k=0}^{K} \theta_{kU} p_k(u)\}^2$, where $p_k(\cdot)$ is the scaled Hermite polynomial mentioned in 2, also θ_{kT} and θ_{kU} are the corresponding coefficients for the k^{th} term of the scaled polynomials. In this chapter, we will assume that for each replicate of U, the distribution type will be the same. For any given K amount of polynomials, the coefficients θ_T and θ_U can be estimated by maximizing the log of the following likelihood for each person i

$$\mathcal{L}_i(f_T | W_{i1}, ..., W_{iJ}, \theta_{kT}) = \int \left\{ \sum_{k=0}^{K} \theta_{kT} p_k(t_i) \right\}^2 \prod_{j=1}^{J} f_U(w_{ij} - t_i) dt_i, \tag{17}$$

subject to the constraint that all unknown densities have to integrate to 1 and that the error term U has a mean of 0, which is expressed by these following three functions

$$\sum_{k=0}^{K} \theta_{kT}^2 = 1, \tag{18}$$

$$\sum_{k=0}^{K} \theta_{kU}^2 = 1, \tag{19}$$

$$\textstyle\sum_{k=0}^{K-1} \sqrt{2(k+1)} \theta_{kU} \theta_{(k+1)U} = 0. \tag{20}$$

Also the number of basis functions K will be estimated using BIC from section 4.1.

3 Simulations

For this section, we will be using the same information as the previous chapter. Therefore variable T will have two types of distribution:

- a standard normal distribution: $T \sim \text{Normal}(0, 1)$
- a gamma distribution: $T \sim \Gamma(9, 1/3)$, where the shape parameter for this gamma distribution is 9 and the scale parameter is $1/3$

And U will also have two types of distribution:

- a Laplace distribution: $U \sim \text{Laplace}(0, 1/\sqrt{3})$
- a normal distribution: $U \sim \text{Normal}(0, 1/\sqrt{3})$

For both cases of U above, we have the mean and standard deviation for each distribution. For this section, we will follow the previous chapter and look at all four combinations of T and U.

Similar to the simulations from the previous chapter. We will look at examples that use different combinations of T and U. For each example, we will perform 350 simulations, where each simulation will have a sample size of $n = 500$ and each subject will have $J = 4$ replicates. Once again, this pre-existing knowledge of the true T and U will not be using it the process of analysis, and will only be available at the result section as a comparison.

For all of the following examples, we will be comparing the estimated density curve of T averaged over all simulations to the true density of T. For all the following figures, the black solid curve will be representing the true density of T, and the red dashed curve will be representing the estimated density of T averaged over all simulations.

3.1 Estimate T when we know both the distribution type and standard deviation of U

We start off with the situation where the data we generate has 4 replicates and that we know both the type of distribution for U and its standard deviation. In this situation, the

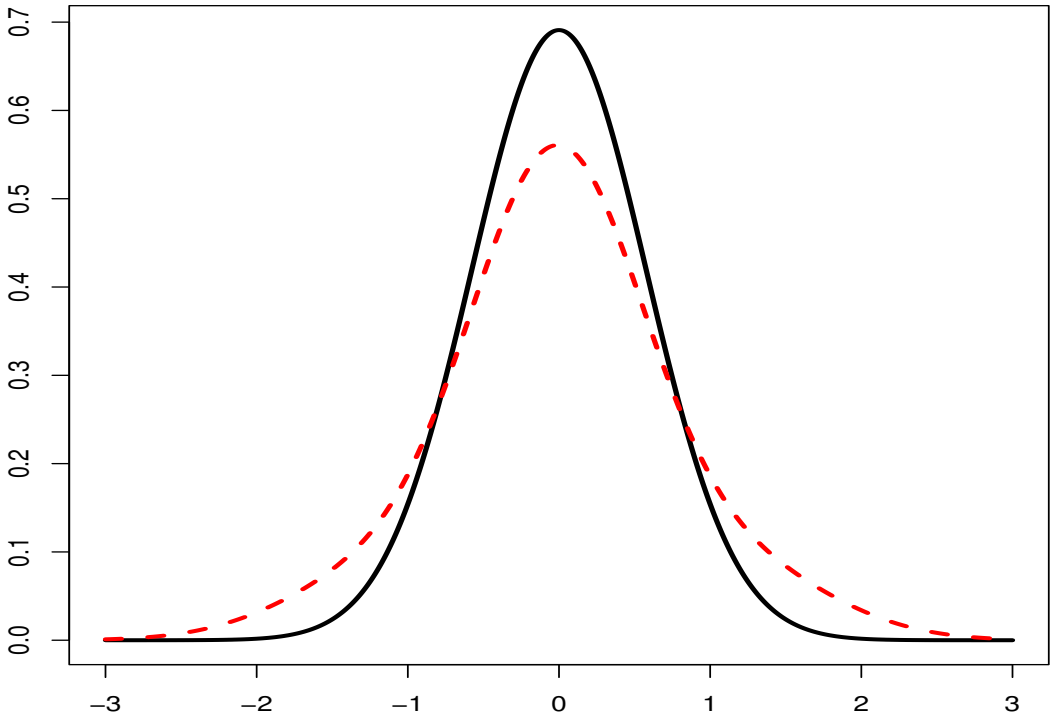


Figure 21: Comparing the estimated density of T (red dashed) with the true density curve(black solid, normal distribution), where U is assumed as a Laplace distribution and we also know the standard deviation of U.

only parameters that are being estimated are the smoothing coefficients that are used for the density estimation of T.

Figure 21 looks at the combination where the true density of T is normally distributed and U has a Laplace distribution with a standard deviation of $1/3$. Here the black solid curve is the true density curve and the red dashed curve is the estimated density curve averaged over 350 simulations. We can see that on average, the density estimated is centered correctly, but has a much smaller variance compared to the true density curve.

Figure 22 looks at the combination of both T and U having normal distributions. Similar to figure 21 the average of all estimated density curves has the correct centering, but a much smaller variance.

Figures 23 and 24 look at the density estimation where T has a gamma distribution, where the first figure has a Laplace distribution for U and the latter figure uses a normal distribution for U. For both averaged density estimations, the estimated curves has captured

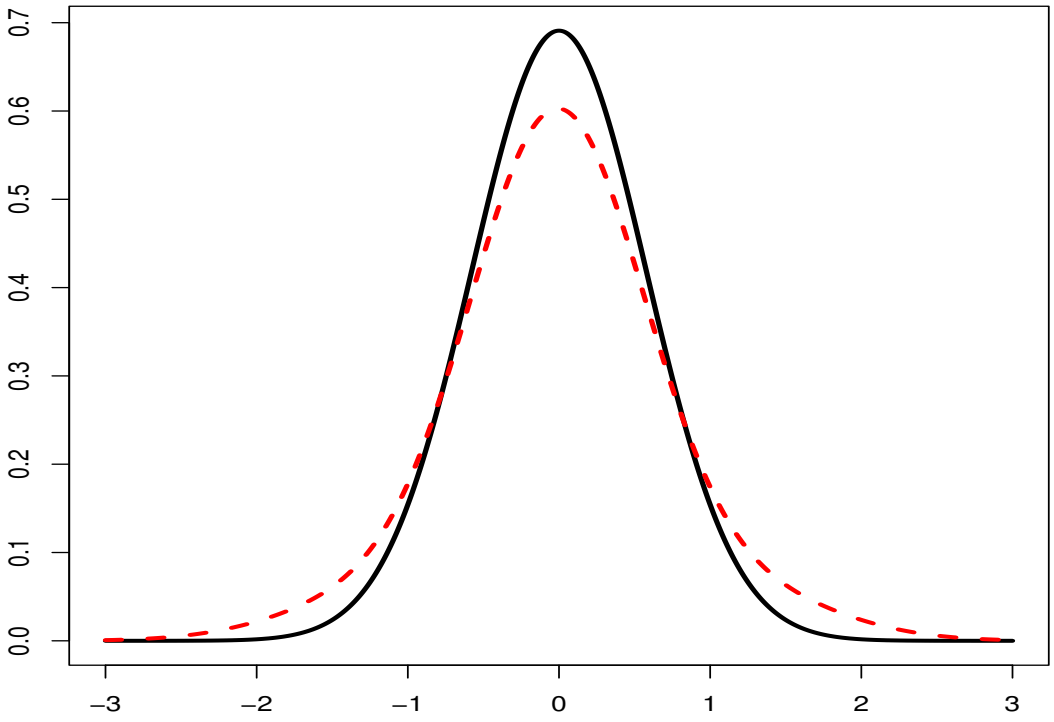


Figure 22: Comparing the estimated density of T (red dashed) with the true density curve(black solid, normal distribution), where U is assumed as a Normal distribution and with a known standard deviation for U.

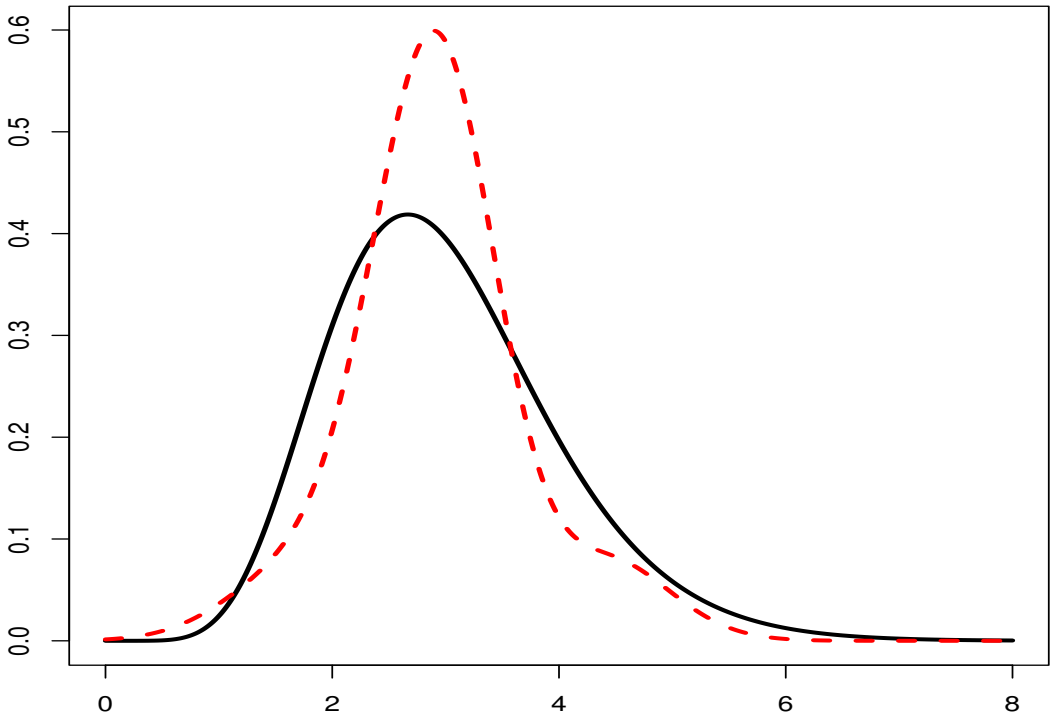


Figure 23: Comparing the estimated density of T (red dashed) with the true density curve(black solid, gamma distribution), where U is assumed as a Laplace distribution with an known standard deviation for U.

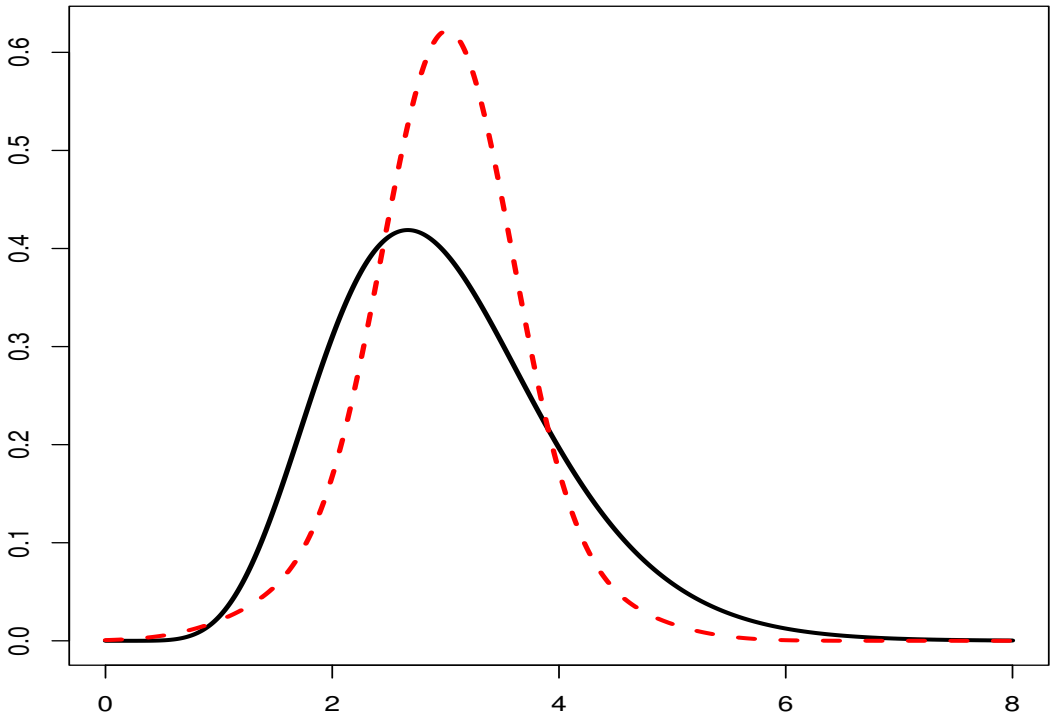


Figure 24: Comparing the estimated density of T (red dashed) with the true density curve(black solid, gamma distribution), where U is assumed as a Normal distribution with an known standard deviation for U.

the fact that the true density is skewed, and the mode of the curve is around the correct place, but once again, the estimated curves have a much smaller variance than the true curves. Figure 23 shows a smaller "bump" next to the mode of the averaged estimated curve, this is due to the averaging of all estimated curves from the 350 simulations, it is not an indication that the estimated density curve is bi-modal.

3.2 Estimating the density of T when we only know the distribution type for U

In this subsection, we will be continuing with the same examples as the previous subsection. This time we will relax the problem a bit by giving information on the distribution type for U, but not on the standard deviation. Therefore in the likelihood function, we will be estimating both the smoothing coefficient for T and the standard deviation for U. Since we are only interested in how well the density of T is estimated, therefore in the results we will only present the comparison the average of all estimated densities of T with the real density curve of T.

Figure 25 and 26 shows the density estimation where the true density of T is normally distributed, where U is a Laplace distribution in figure 25 and U is a normal distribution in the latter figure. Similar to the cases where we know the standard deviation value of U, the averaged density estimation of T is centered correctly, but the average of all estimated density curves has a heavier tail.

Figure 27 and 28 shows the density estimation for T where the true T has a gamma distribution. U has a Laplace distribution in figure 27 and U is normally distributed in figure 28. For both cases, the average of the estimated density curves is skewed in the same direction as the true density, but the estimated mode is shifted to the right and the variance is much smaller than the true curve.

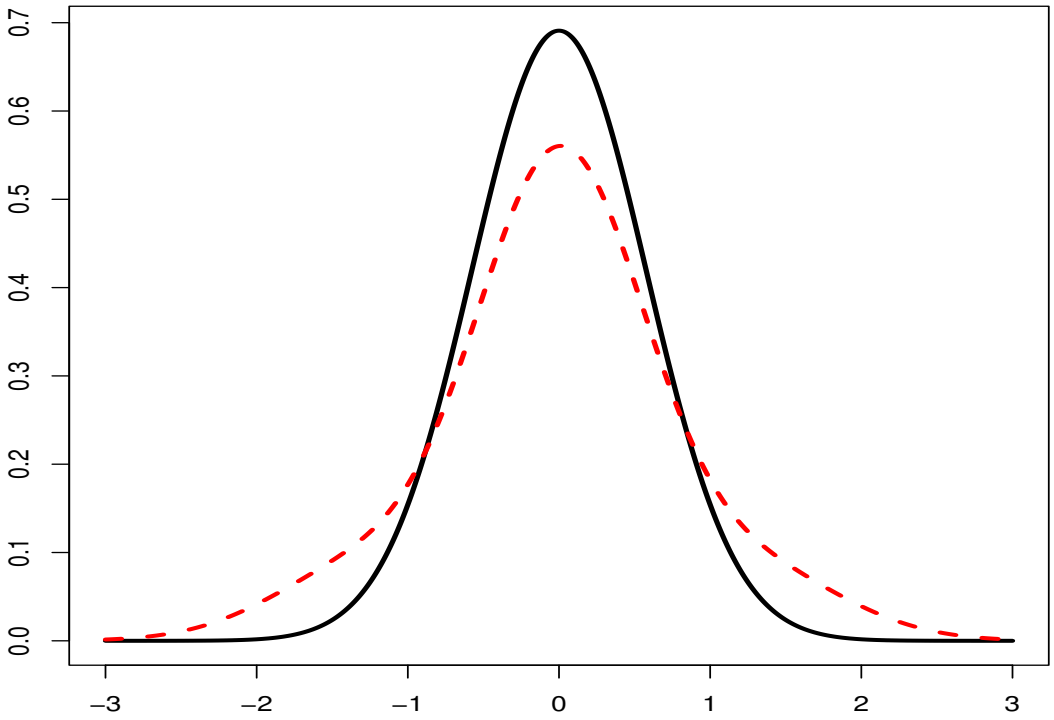


Figure 25: Comparing the estimated density of T (red dashed) with the true density curve(black solid, normal distribution), where U is assumed as a Laplace distribution where the standard deviation of U needs estimation.

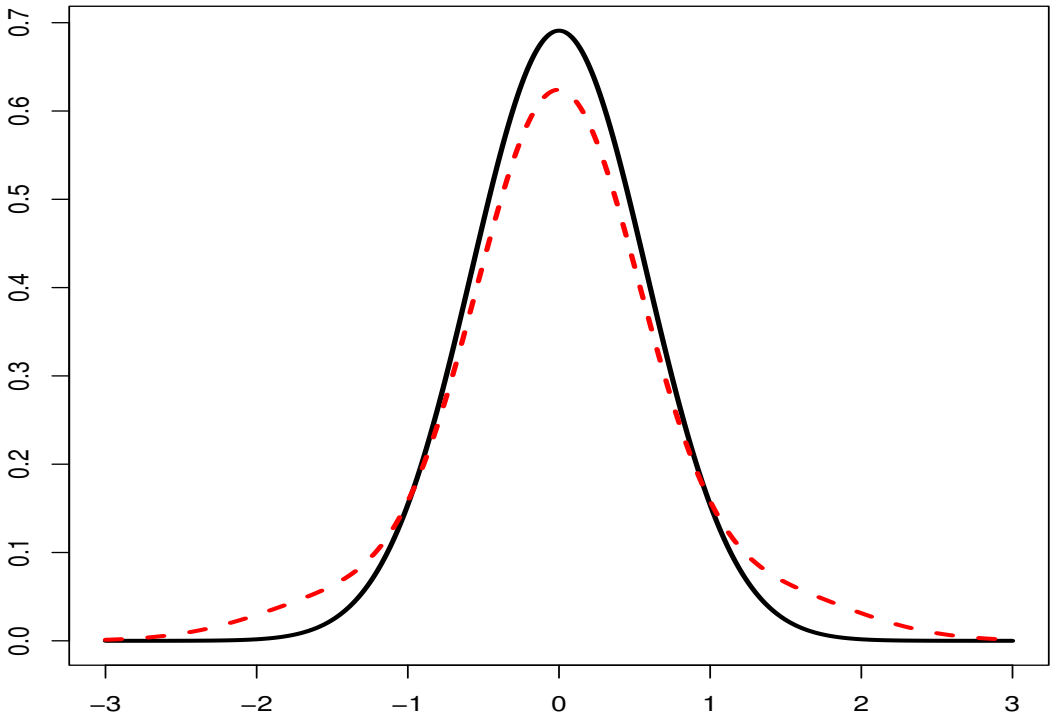


Figure 26: Comparing the estimated density of T (red dashed) with the true density curve(black solid, normal distribution), where U is assumed as a Normal distribution and with a unknown standard deviation for U.

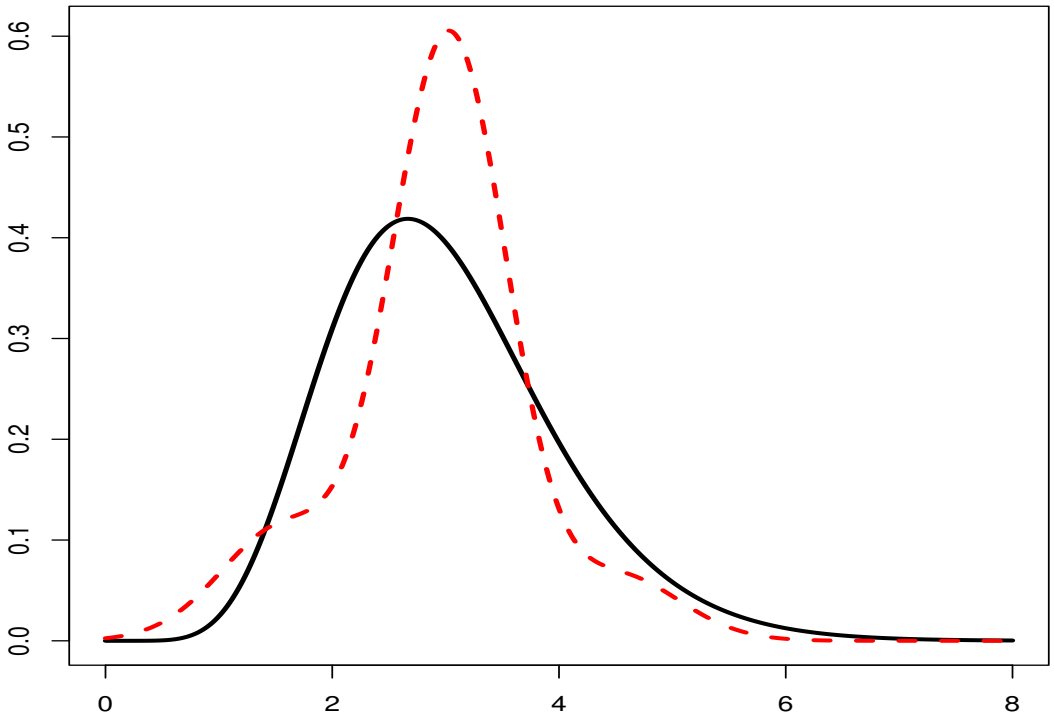


Figure 27: Comparing the estimated density of T (red dashed) with the true density curve(black solid, gamma distribution), where U is assumed as a Laplace distribution with an estimated standard deviation.

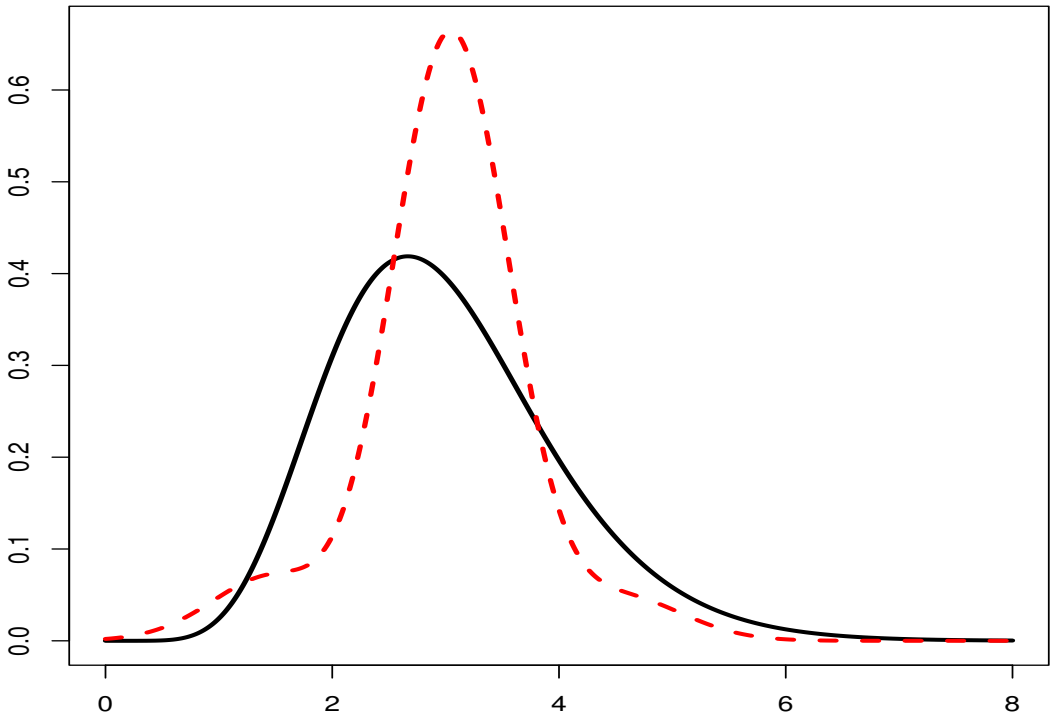


Figure 28: Comparing the estimated density of T (red dashed) with the true density curve(black solid, gamma distribution), where U is assumed as a Normal distribution with an estimated standard deviation.

3.3 Estimating the density of T with no information on the density of U

For this subsection, we will be relaxing the assumptions even more by not providing any assumption on both the distribution type and the standard deviation for U. In this case, U will also be represented as a sum of a series of basis functions where each basis function consists of one term of the scaled Hermite polynomial and a corresponding smoothing coefficient. Therefore in the likelihood function, we will be estimating both the smoothing coefficient for T and the smoothing coefficient for U. But similar to the previous subsection, we will only be presenting the estimated densities of T.

Once again, we start by looking at how well this method estimates the density of T when the true T is normally distributed. Figure 29 looks at the averaged density estimation of T when the true distribution for U has a Laplace distributed, and figure 30 looks at the cases where the true U has a Normal distribution. We see that with so little known information, the averaged estimated density of T has a larger variance than that of the true density curve.

We also looked at the cases where the true density of T is a gamma distribution, for both where the true U is normal (figure 32) and when U has a Laplace distribution (figure 31). For both cases, the average estimation for T has a smaller variance than the true density and is shifted a bit to the right.

4 Comparisons

In chapters IV and V, we have in total performed simulations for 5 scenarios using Hermite polynomials as a media to estimate any unknown densities:

1. No replications, known distribution for U, known σ_U,
2. No replications, known distribution for U, unknown σ_U,
3. Multiple replications, known distribution for U, known σ_U,
4. Multiple replications, known distribution for U, unknown σ_U,

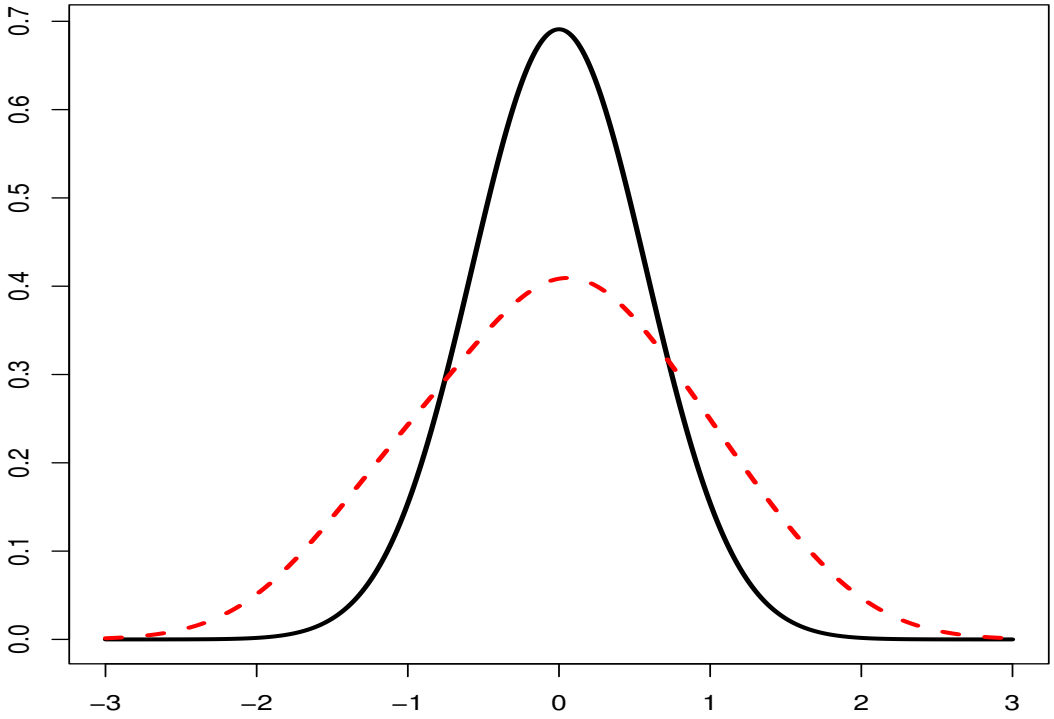


Figure 29: Comparing the average density estimate of T using Hermite polynomials(red line) of 350 simulations to the true density of T(black line), where T has a Normal distribution($T \sim$ Normal$(0, 1)$) and U has a Laplace distribution($U \sim$ Laplace$(0, 1/\sqrt{3})$).

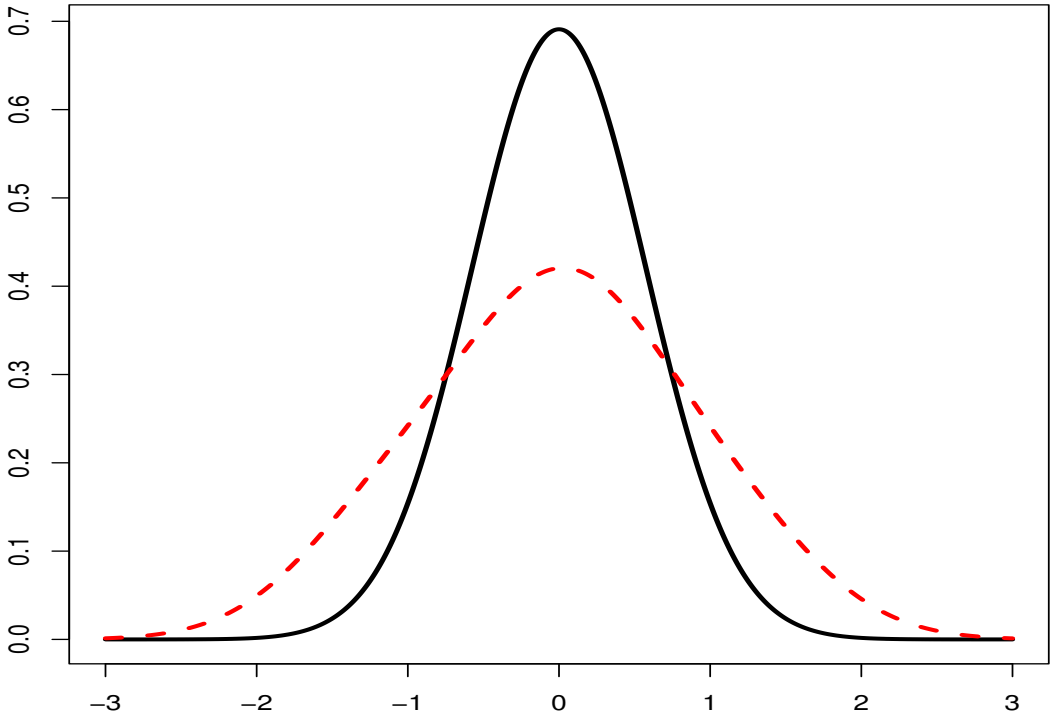


Figure 30: Comparing the average density estimate of T (red line) of 350 simulations to the true density of T(black line), where T has a Normal distribution($T \sim \text{Normal}(0, 1)$) and U also has a Normal distribution ($U \sim \text{Normal}(0, 1/\sqrt{3})$).

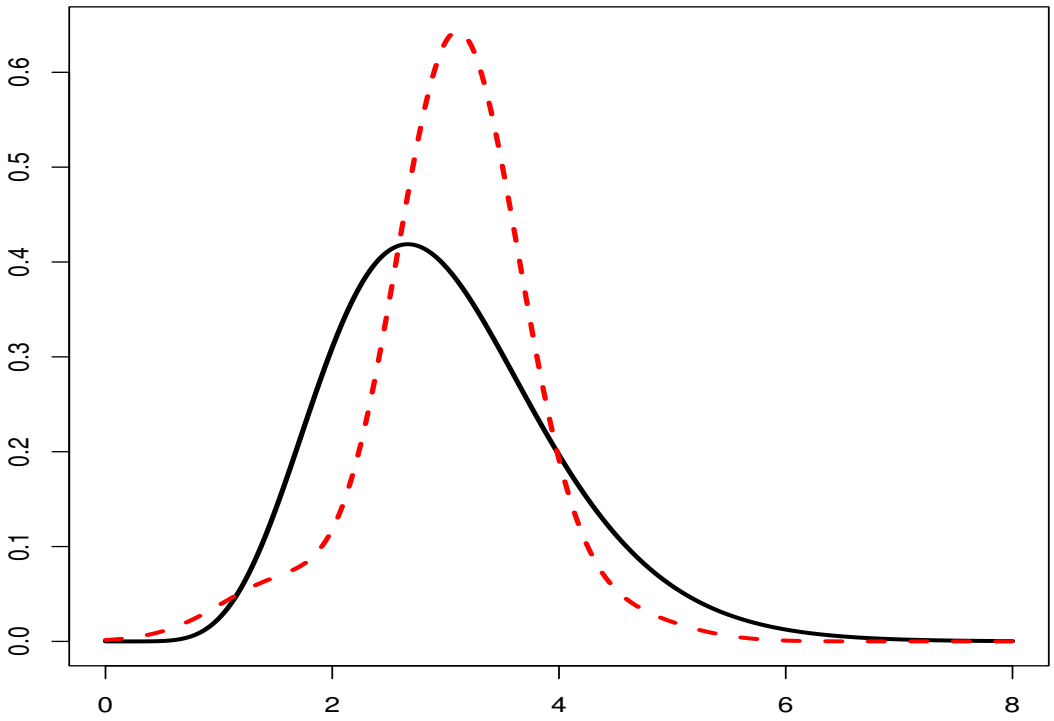


Figure 31: Comparing the average density estimate of T using Hermite polynomials(red line) of 350 simulations to the true density of T(black line), where T has a Gamma distribution($T \sim \Gamma(9, 1/3)$) and U has a Laplace distribution($U \sim \text{Laplace}(0, 1/\sqrt{3})$).

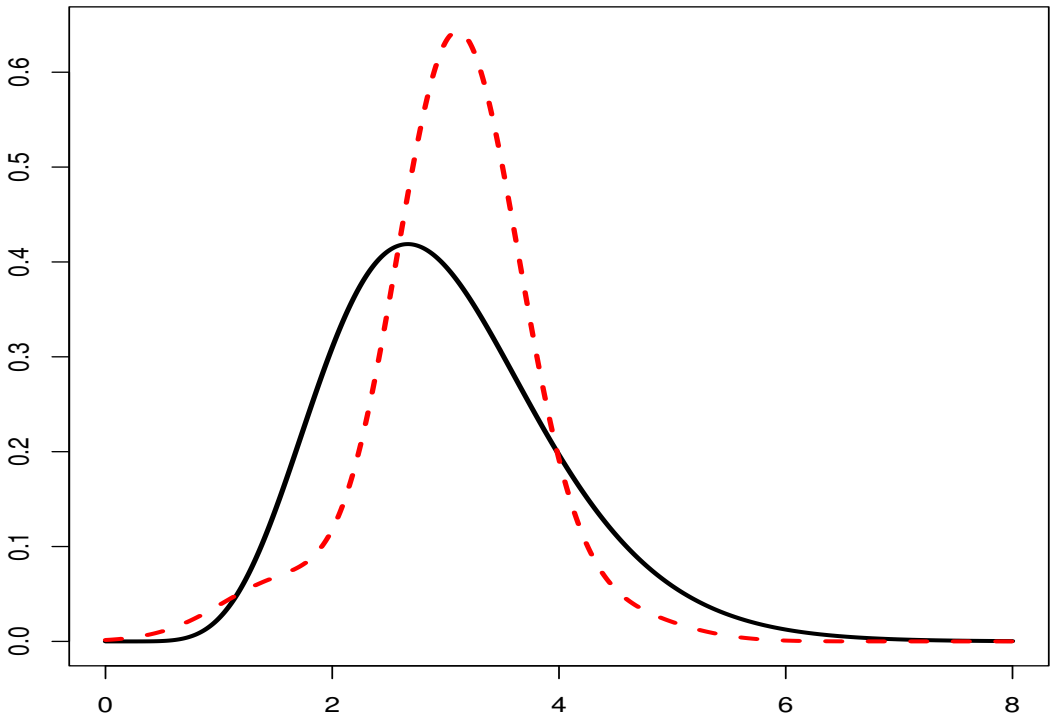


Figure 32: Comparing the average density estimate of T using Hermite polynomials(red line) of 350 simulations to the true density of T(black line), where T has a Gamma distribution($T \sim \Gamma(9, 1/3)$) and U has a Normal distribution($U \sim \text{Normal}(0, 1/\sqrt{3})$).

5. Multiple replications, unknown distribution for U, unknown σ_U.

For this section, I would like to perform comparisons between these scenarios to see how the information given to each scenario affects the result of the estimation of T. Although we have been looking at multiple examples in the previous simulations, in this section, we will only be using one example to perform these comparisons. We will be looking at the example where T is has a standard normal distribution, and U has a Laplace distribution with mean 0 and standard deviation of $1/\sqrt{3}$.

4.1 Knowing σ_U vs estimating σ_U

We will start with comparing the averaged estimated densities of T between scenarios where we know the standard deviation of U and the scenarios where we need to estimate the standard deviation of U. That is we will be comparing the results from scenario 2 with scenario 1 and the results from scenario 4 with scenario 3. For the following figures, the black solid curve will be representing the true density curve of T, the red dashed curve will be representing the averaged estimated density of T where the standard deviation of U is also estimated, and the magenta dot-dashed curve will be representing the averaged estimated density of T where we know the value of the standard deviation of U.

Figure 33 looks at the comparison where both scenarios contain no replicate for the observed data W. We see that with a correct assumption on the standard deviation of U, the averaged estimated density curve has a higher peak. Figure 34 looks at the comparison where both scenarios contain 4 replicates for W. We see that the red dashed curve and the magenta dot-dashed curve is almost overlapping each other, this brings us to the conclusion that when we have enough replicates in the observed data, knowing or estimating the standard deviation or variance of the error term does little to affect our estimation of the density of T, but with no replicates in W, having the correct information on σ_U will allow a more accurate estimation for the density of T.

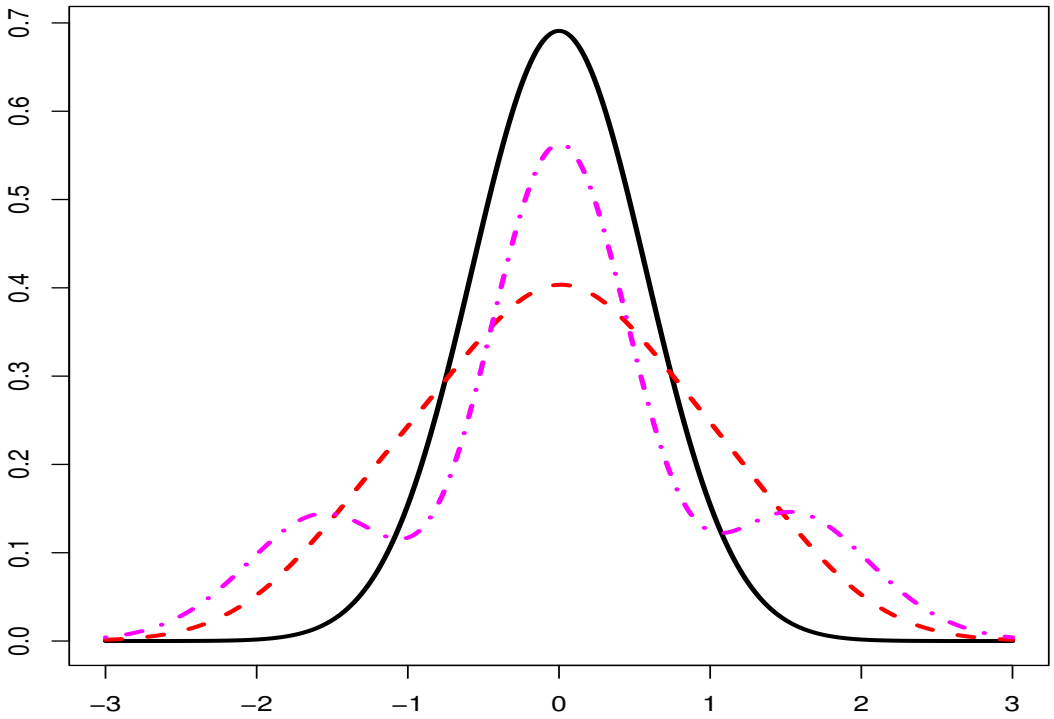


Figure 33: Given that both scenarios the data has no replicates and the distribution type for error term U is assumed, compare the averaged estimated density curve where the standard deviation of U is assumed (magenta dot-dashed) and where the standard deviation of U is estimated (red dashed). The black solid curve is the true density curve of T ($T \sim \text{Normal}(\mu_T = 0, \sigma_T = 1)$).

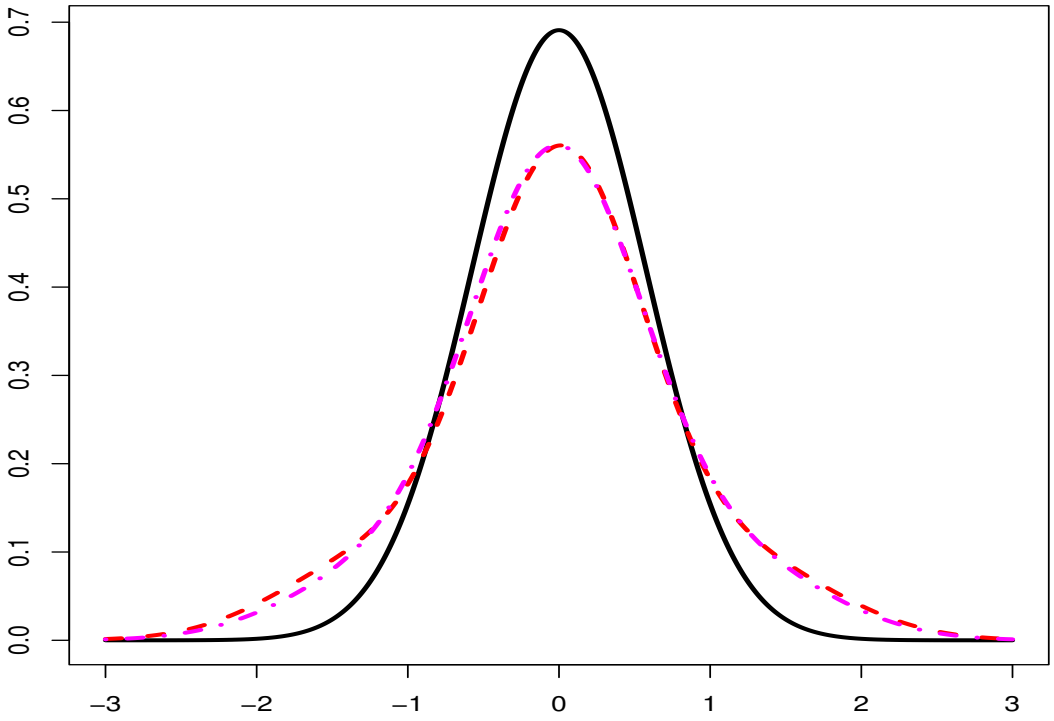


Figure 34: Given that both scenarios has 4 replicates and the distribution type for error term U is assumed, compare the averaged estimated density curve where the standard deviation of U is assumed (magenta dot-dashed) and where the standard deviation of U is estimated (red dashed). The black solid curve is the true density curve of T ($T \sim \text{Normal}(\mu_T = 0, \sigma_T = 1)$).

4.2 No replicates vs multiple replicates

We now look at how having multiple replicates may affect the estimation result for the density of T. We start with comparing scenario 1 with scenario 3 to see with a good understanding on the error term U, what type of effect replicates have on estimating density T, we then compare scenario 2 with scenario 4 to see whether the additional estimation of σ_U will alter or support our findings. For the following figures, the black solid curve will still be representing the true density of T, but now the red dashed curve will be representing the scenarios with 4 replicates of W and the magenta dot-dashed curve will be representing the scenarios where W have no replicates.

Figure 35 looks at no replicates for W versus 4 replicates for W where we assume the correct distribution type and standard deviation for error term U. From this figure, we can conclude that having multiple replicates for W aids in the smoothing of the estimated densities. Figure 36 works with the same comparison but on the scenarios where we also need to estimate the standard deviation of U, and this figure supports our previous findings.

4.3 MAE and MSE results

Similar to section 4.3 in the previous Chapter (Chapter 3), the MAE and MSE values are recorded for each scenario. In this section, we will be exploring how the MAE and MSE values change between each scenario.

Figure 37 compares the MAE values between each scenario and figure 38 compares the MSE values between each scenario. For both figures, there are in total five boxplots labeled 1 to 5, each one of these boxplots corresponds to the labels for each scenario. For each boxplot, we record the MAE and MSE values for all 350 simulations. In both figures, we see that the boxplots for scenarios 1 and 2 have a much smaller variation than scenarios 3, 4 and 5, this may cause some concern since it looks like results obtained from multiple replicates are more variable than results obtained from using only 1 replicate. This variation is caused due to the computation process. Through the whole simulation process for multiple replicates, a good amount of simulations produced inconclusive results, this also gave MAE and MSE values that are quite large, causing the illusion that multiple replicates are more variable,

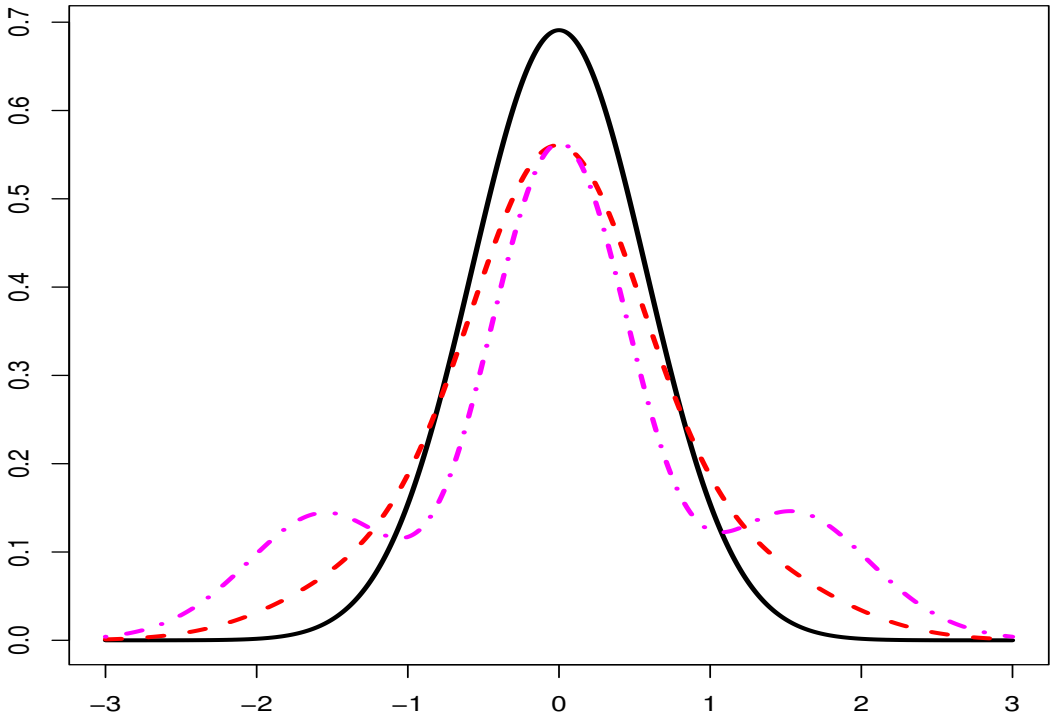


Figure 35: Given that both scenarios the distribution type and the standard deviation for error term U is assumed, compare the averaged estimated density curve where the data has no replicates (magenta dot-dashed) and where the data has 4 replicates (red dashed). The black solid curve is the true density curve of T ($T \sim \text{Normal}(\mu_T = 0, \sigma_T = 1)$).

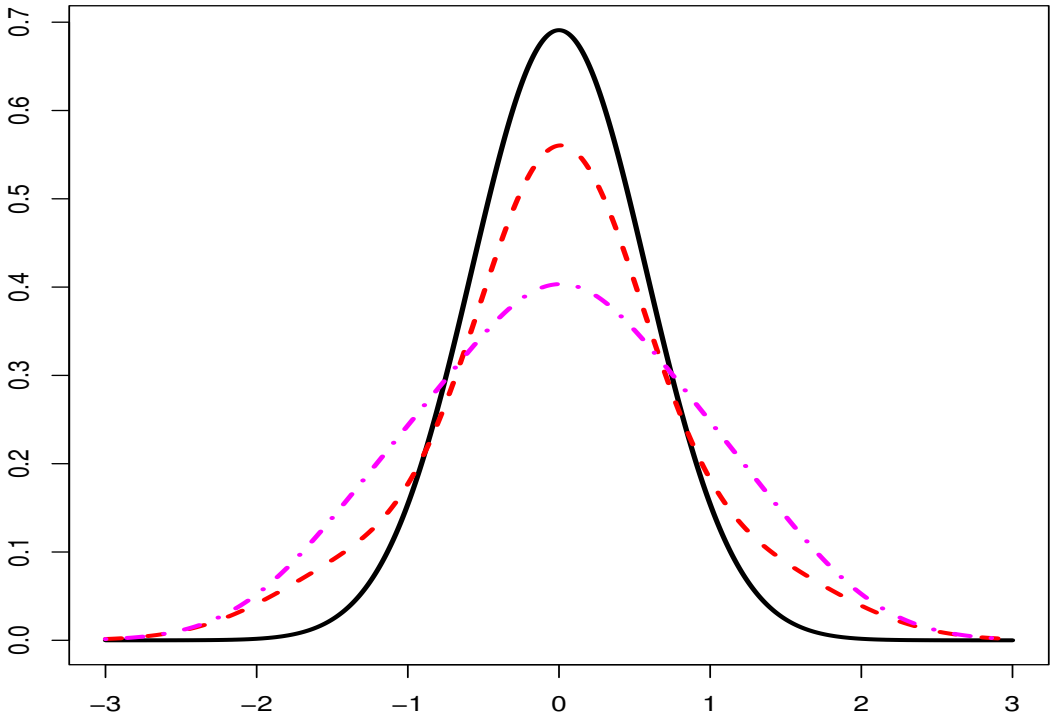


Figure 36: Given that both scenarios has assumed the distribution type for error term U and estimates the standard deviation for U, compare the averaged estimated density curve where the data has no replicates (magenta dot-dashed) and where the data has 4 replicates (red dashed). The black solid curve is the true density curve of T ($T \sim \text{Normal}(\mu_T = 0, \sigma_T = 1)$).

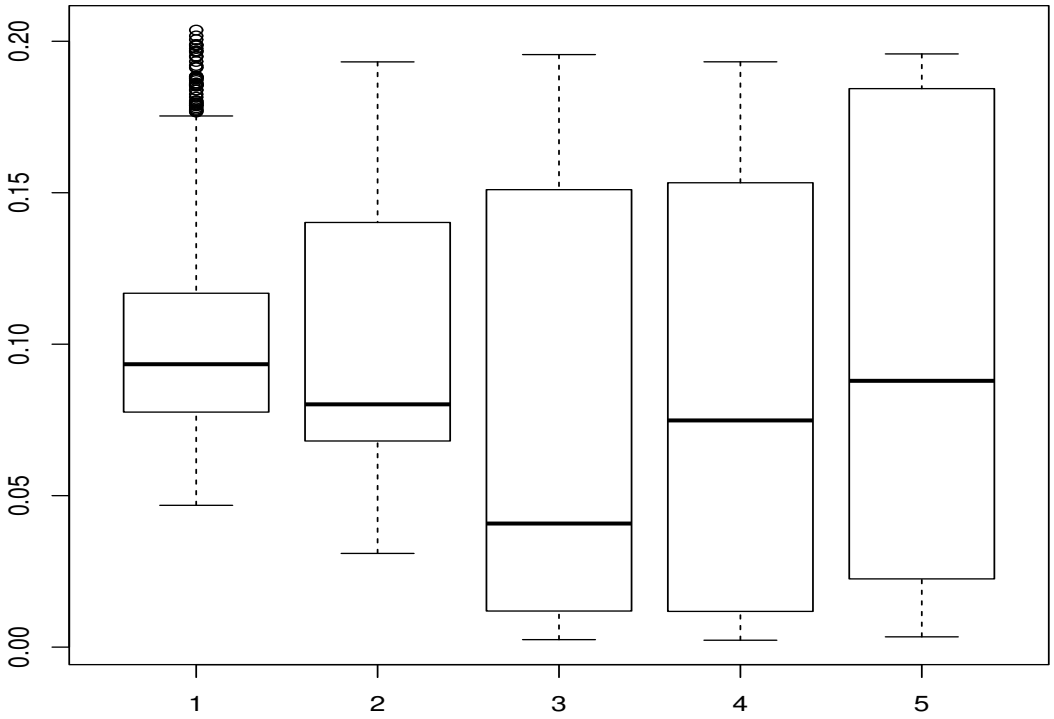


Figure 37: A series of boxplot that compares the mean absolute error values of each scenario, where the numbers labeled under each boxplot corresponds to the label for each scenario.

in fact if we look at the median of each box plot, we can see that scenarios where there are multiple replicates have smaller MAE and MSE values, indicating that an increase in complexity in computation will increase the amount of inconclusive results, but when the computation process give conclusive results, multiple replicates will provide a much more accurate result.

5 Real Data

5.1 Quick Introduction to EATS

Many nutritional surveys have used the 24-hour recall (24HR) to collect information on food intake (Dwyer et al. (2003)). 24HR recall collects a subject's food and nutritional intake for the past 24 hours. The main purpose of this section is to use these pieces of daily information to estimate how the population's long term daily intake of nutrients and foods is distributed. Eating at America's Table Study (EATS) is one of the studies that use 24HR to collect

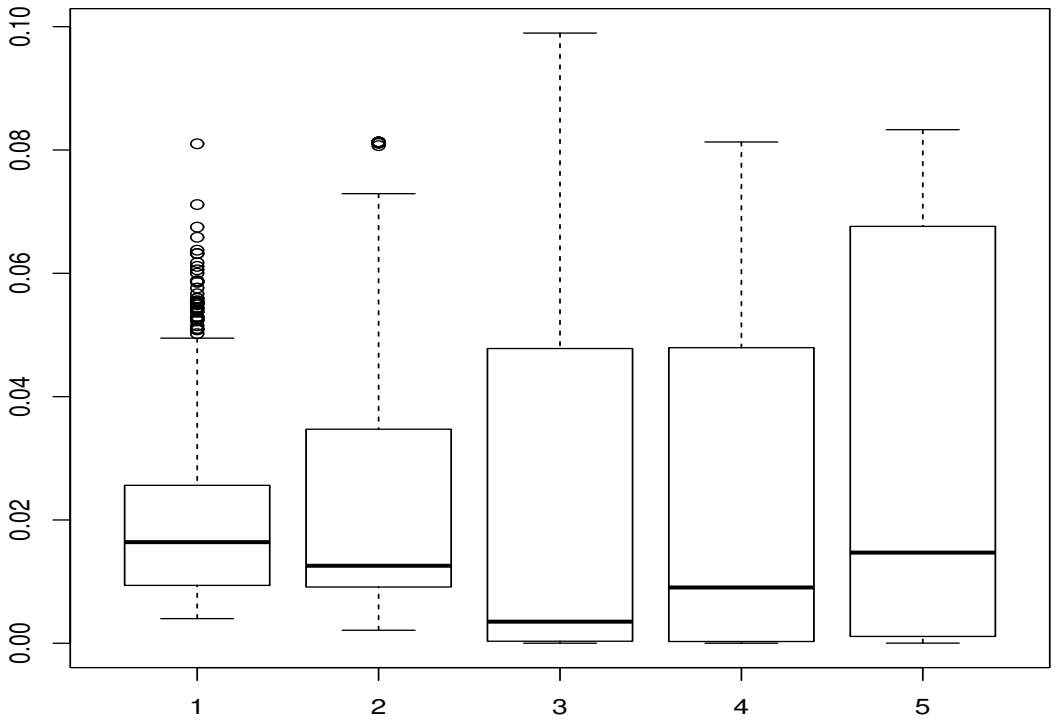


Figure 38: A series of boxplot that compares the mean squared error values of each scenario, where the numbers labeled under each boxplot corresponds to the label for each scenario.

data. Since 24HR records each subject intake for the past 24 hours, some data entries will regularly have zero input and some subjects may not have replicated, for this analysis we will need to choose either a food intake variable that we believe most people will eat almost daily (minimal zero input), or choose a nutritional intake such as calcium, protein or fat that we most likely will ingest daily through a variety of foods. Since EATS has 4 replicates, we will analyse the data by taking the average of all 4 replicates for each subject. We will use the data provided by Subar et al. (2001), and the variable that we are focusing on is the protein intake. For this example, I have separated the male and female participants, and provided two density plots, one for each gender. For both gender, as suggested by Subar et al. (2001), I have adjusted the data by dividing each input by their respective energy intake per thousand calories. That is, if a participant recorded to have consumed 12 grams of protein the past 24 hours and that their energy intake that particular 24 hours is 2500 calories, then we adjust the data by dividing 12 by 2.5, and therefore we will use the result 4.8 grams of protein per thousand calories as the adjusted data for analysis.

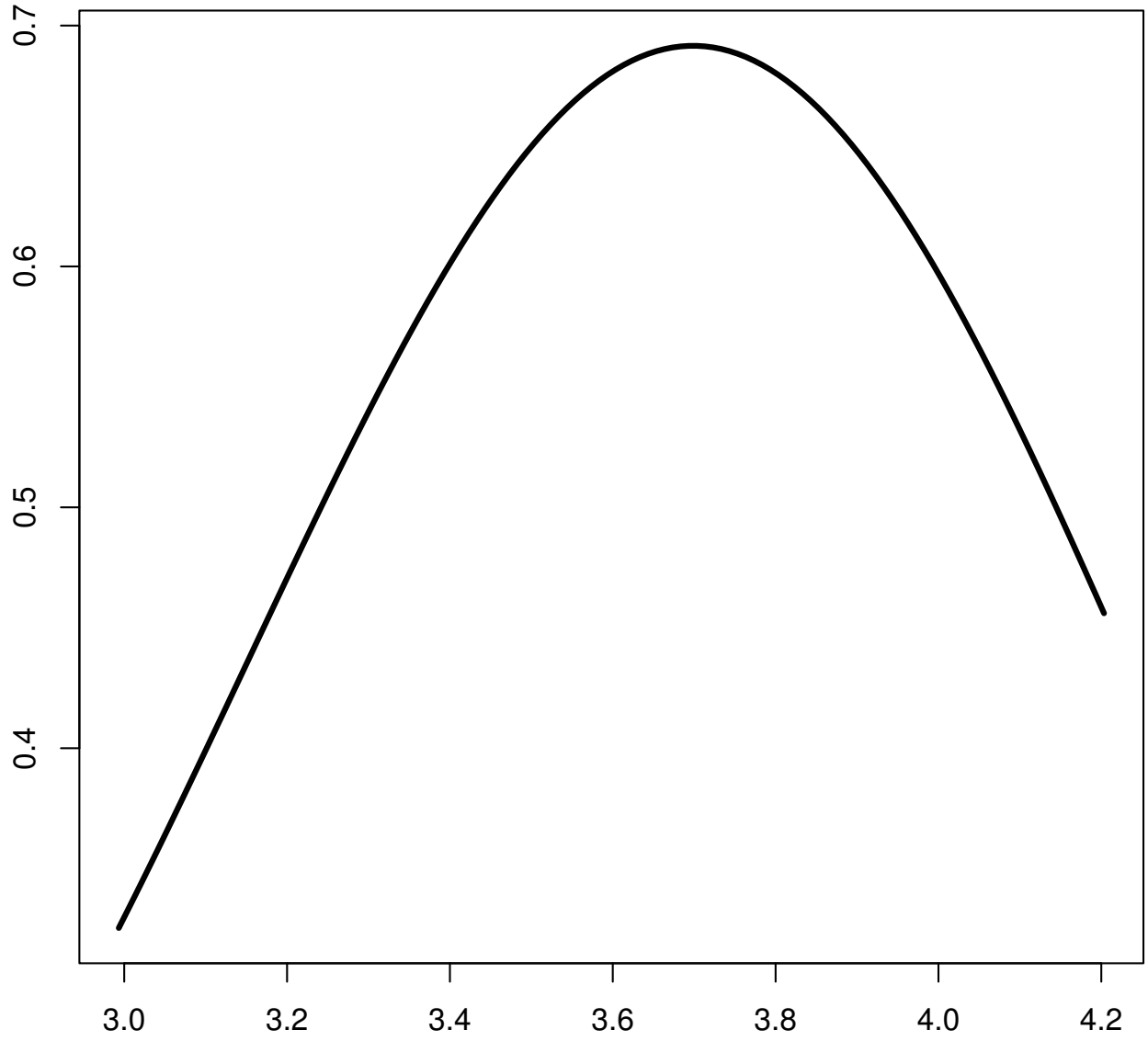


Figure 39: The estimated density curve of a populations long term usual average daily intake of proteins for females.

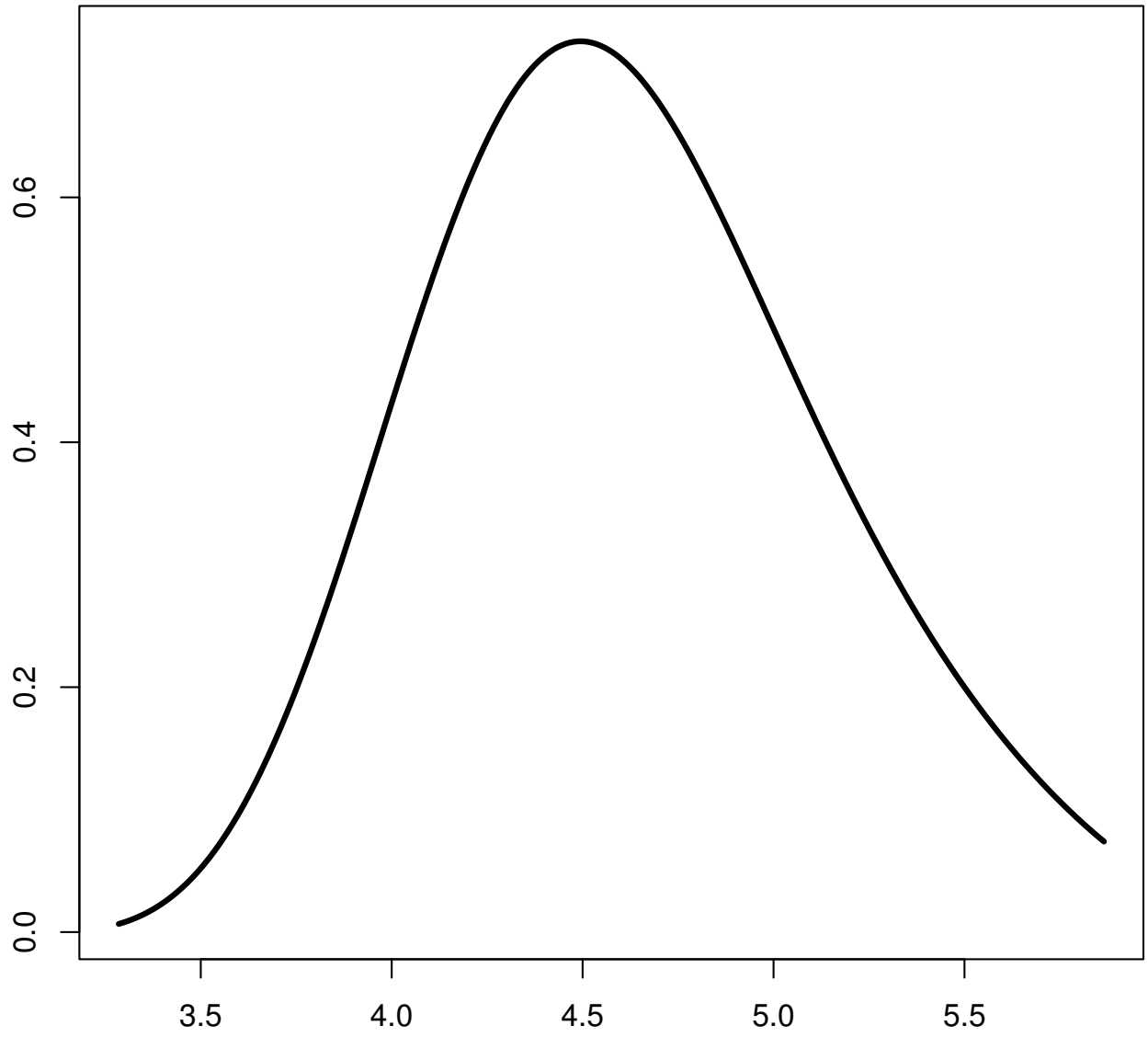


Figure 40: The estimated density curve of a populations long term usual average daily intake of proteins for males.

Figure 39 looks at the density curve for the populations usual daily intake of protein in females, and figure 40 looks at the density curve for the populations usual daily intake of protein in males. We see that in general, males take more protein than females, but not by too much. As this density is plotted after energy adjustment, we can say that the mode of males consume around 4.5 grams of protein per thousand calories, and the mode of females consume around 4 grams of protein per thousand calories. Now if we assume that a person's energy intake is 2000 calories per day, then we can then that the mode of males on average consumes 9 grams of protein per day, and the mode of females on average consume 8 grams of protein per day.

Part VI

Chapter 5: Semiparametric Density Deconvolution for Data with Excess Zeros

In the previous chapters, we explored the idea of estimating the density of the latent variable T first semi-parametrically, then non-parametrically, using a simple classical error model. The results show that this model gives accurate estimates when the observable variable W is continuous. Although continuous variables are quite common, through collecting data we see that the observed variables may take many forms, and not all forms perform well when analysed using only a simple classical error model. For this chapter, we will be exploring the density estimation for a specific type of data - data with excess zeros, specifically zero-inflated data for nutritional data. For this type of data, we will be using a more complicated model to describe the relationship between the observable variable W and the latent variable T, we will follow the lead of Tooze et al. (2006) and use the NCI (National Cancer Institute) model. The NCI model consists of two parts: part 1 accommodates the percentage of zeros that is in the data, part 2 models the non-zero part of the data, this part will be a linear model, but depending on the dataset, a transformation may be used on this linear model. And lastly, we combine these two parts as a way to estimate the density of the latent variable T, which, in the nutritional concept will often be a population's long term average of a certain intake.

The NCI model has been frequently used in health research, Tooze et al. (2006) looked at the density estimation of variable T using the NCI model assuming normality on all latent variables, they created separate latent variables for parts 1 and 2 of the model, allowing correlation to exist between the latent variables, therefore in order to estimate the density of variable T, they had to first estimate the latent variables and the correlation values. Kipnis et al. (2009) estimated the T value for each subject and then applied regression calibration to understand the relationship between T and various health outcomes. To decrease the

number of variables that required estimating, instead of allowing correlation they created a simple linear relationship between the latent variables in part 1 and part 2 of the model. For this chapter, we will be estimating the density of T semi-parametrically by using as little assumptions as possible. We will follow the lead of Kipnis et al. (2009) and simplify the NCI model by using a simple linear relationship on the latent variables between the two parts of the NCI model.

For this chapter, we will be using the simplified version of the NCI model which contains only two latent variables and one error variable. The idea is to obtain density estimations for all these unobservable variables and use these estimated densities to estimate the density of our final latent variable T. We will start with a detailed description of how to combine the Hermite density deconvolution with this simplified NCI model and how we estimate the final density using previously estimated densities of other variables, then we will present some simulation results.

1 Methodology

For this section, we will first give a brief description of the NCI model. We will then look at our simplified version of the NCI model, and discuss in detail the likelihood of this model and how densities represented by polynomial functions are incorporated into the likelihood. Lastly, we will be discussing how to estimate the density of T, once we obtain densities for all other latent variables.

1.1 Using Zero Inflated Data

The NCI model is a model which helps analyse a dataset with excess zero. The model is developed at the National Cancer Institute, the idea was to use two or more short term recalls of a subject's nutritional intake to estimate a population's long term consumption pattern. An example of a zero-inflated dataset is in surveys where you are given a yes or no question, where if you say no, the input will be zero, and if the answer is yes, an additional question of "how much?" will follow, and in this case, the amount will be recorded. For a

question such as "Have you ingested any alcohol in the past 24 hours?", it is most likely that a significant proportion of the subject may answer no, which will result in a dataset with a lot of zero inputs. Also when asked this same question to the same subject, the answer may change depending on the different days, since people are more likely to drink in the weekend than on weekdays, this suggests that having multiple recalls from the same subject will improve the accuracy when estimating a population's long term pattern.

For this paper we will assume that each subject will be answering the same question multiple times, where each subject will only be asked the question once in a 24 hour period, we call the results collected each time as a recall or a replicate. The NCI method performs on the bases that even though a subject may have recorded zero as their input in one or multiple of the recalls, this does not mean that this subject's record will stay as zero in the long term. That is when a subject has answered no to "Have you ingested any alcohol in the past 24 hours?", we do not assume that this subject is a non-alcoholic drinker, just that the subject happened to not drink on that particular day. By using a dataset with replicates, the NCI method takes into account how often a subject may have zero as an input, and when the input is non-zero, how large or small the amount may be. Using this information, we will determine what the density of the variable of interest would look like in the long term.

In this section, we will be looking at our simplified NCI method with the help of multiple replicates. For $i = 1, ..., n$ subjects, consider two latent variables, (X_i, U_i) and assume that they are independent to each other. For $j = 1, ..., J$ replicates, let W_{ij} be the observed data, we will be assuming the all subjects have the same amount of replicates. Then the observed data W will contain the input from all J replicates and all n subjects. Let ϵ_{ij} be the variability between each replicate for each subject on a transformed scale and C_{ij} be the binary indicator of the observed variable W_{ij}, where C_{ij} is 0 if W_{ij} has a zero input and C_{ij} is 1 if W_{ij} is a positive number. Our model based on Kipnis et al. (2009):

$$\begin{aligned} \text{pr}(C_{ij} = 1|X_i) &= \Phi(X_i) && (21) \\ \text{h}(W_{ij}|C_{ij} = 1, X_i, U_i) &= \alpha X_i + U_i + \epsilon_{ij} = W^*_{ij} && (22) \end{aligned}$$

where $\Phi(\cdot)$ is the probit function and $h(\cdot)$ is a transformation function. Kipnis et al. (2009) recommend the BoxCox transformation.

For this particular model, we can interpret X_i as the propensity of a subject consuming food or nutrition of interest, and U_i being the adjustment amount between subjects. Using alcohol consumption as an example, when looking at the mean amount of alcohol ingested using the recall values recorded, subject A and B may have similar values, but this does not mean that these two subjects will have the same consumption patterns. One subject may have a glass of wine every night after dinner, and the other may not drink any from Monday to Friday, but have long drinking sessions in the weekend. That is although these two subjects may have similar $\bar{W}_i$ values, the first subject will have a much larger X_i value and a lower U_i value, whereas the latter subject will have a lower X_i value and a larger U_i value.

Let $f_X(\cdot)$, $f_U(\cdot)$ and $f_\epsilon(\cdot)$ be the density functions corresponding to variables X, U and ϵ. Making the restriction that $\int u f_U(u)du = 0$, the likelihood for person i can be expressed as the probability of this person consuming $\{\Phi(X_i)\}^{C_{ij}} \times \{1-\Phi(X_i)\}^{(1-C_{ij})}$ times the likelihood of regression 22, the likelihood function that combines the information of 21 and 22 can be written as

$$\begin{aligned}
&\mathcal{L}_i(\alpha, f_x, f_U, f_\epsilon | C_{i1}, ..., C_{iJ}, W^*_{i1}, ..., W^*_{iJ}) \qquad (23)\\
&= \int\int \prod_{j=1}^{J} \Big[\{\Phi(X_i)\}^{C_{ij}} \times \{1-\Phi(X_i)\}^{(1-C_{ij})} \times \{f_\epsilon(W^*_{ij} - \alpha X_i - U_i)\}^{C_{ij}} \Big] \\
&\qquad \times f_X(X_i) \times f_U(U_i) dX_i dU_i.
\end{aligned}$$

Assuming the transformation results in approximate normal, we can then assume that $f_\epsilon(\cdot)$ has a normal distribution, but with no assumption on the mean or the variance. As for $f_X(\cdot)$ and $f_U(\cdot)$, except that the latter has mean zero, no additional assumptions are given and therefore the densities are considered as unknown. Following the previous chapter, we will estimate any unknown densities, in this case $f_X(\cdot)$ and $f_U(\cdot)$ by using Hermite polynomial basis functions

$$f_X(x) = \{\textstyle\sum_{k=0}^{K} \theta_{kX} p_k(x)\}^2; \quad f_U(u) = \{\textstyle\sum_{k=0}^{K} \theta_{kU} p_k(u)\}^2;$$

Here θ_{kX} and θ_{kU} are the smoothing parameters for the densities of variables X and U, and $p_k(\cdot)$ are the basis functions expressed in 2. In order for the densities functions to integrate

to one, we take advantage of the orthonormal properties of $p_k(\cdot)$ and we have the restrictions (Schennach and Hu, 2013) that

$$\sum_{k=0}^{K}\theta_{kX}^2 = \sum_{k=0}^{K}\theta_{kU}^2 = 1. \tag{24}$$

In order for $f_U(\cdot)$ to have mean zero, we have the restriction (Schennach and Hu, 2013) that

$$\int u f_U(u)du = 0 \quad \rightarrow \quad \sum_{k=0}^{K-1}\sqrt{2(k+1)}\theta_{kU}\theta_{(k+1)U} = 0.$$

1.1.1 Usual Intake

In the previous subsection, we looked at a simplified version of the first two parts of an NCI model. With this we are able to estimate the density of latent variables X and U by estimating the sets of smoothing parameters θ_{kX} and θ_{kU}, we are also able to estimate the mean and standard deviation of variable ϵ and the slope α. But this is not the final product we wish to estimate. For this subsection, we will be exploring how to use the values estimated in the previous subsection to estimate a population's long term pattern on the variable of interest. We will be calling a population's long term pattern as the usual intake.

The main assumption is $E(W_{ij}|\text{subject } i) = (\text{usual intake})_i = T_i$. From Kipnis et al. (2009), a person's usual intake is defined conditional on (X_i, U_i) as

$$T_i = \Phi(X_i)E\{h^{-1}(\alpha X_i + U_i + \epsilon_{ij})|X_i, U_i\}, \tag{25}$$

here $h^{-1}(\cdot)$ is the inverse transformation function. Here we can describe the usual intake function as the product of the probability that a subject has a non-zero input for any recall and the usual amount given that the subject has a non-zero input on that recall.

One of the biggest limitations for calculating (25) is that we do not have the values of X and U for each corresponding subject, and therefore can not obtain the usual intake value T for each individual subject. Our ultimate goal for this section is to use the information obtained on variables X and U from section 1.1, which are the empirical densities f_X and f_U and use them to obtain an estimate of the density of T (f_T). By obtaining f_T we will be able to understand what the distribution of the general populations long term average usual intake is like, and will be able to answer questions such as what percentage of our

population is under or over a certain threshold. Since this is a semi-parametric problem, we can calculate an estimated μ_X, σ_X and σ_U from the densities f_X and f_U, but we do not know the type of distribution for X and U, this means we can not obtain f_T analytically. We will, therefore, estimate f_T numerically. To do this, we create mock samples of variables X and U by sampling a large number of values from f_X and f_U (for example a sample size of 1000000) and inputting these newly sampled values into (25) in order to obtain a mock sample of variable T. We then fit a smooth density curve onto this large mock sample of T to get an empirical estimation for f_T.

For the next few paragraphs, I will explain in detail how to compute the expectation part of (25) and the inverse transformation $h^{-1}(\cdot)$ for the BoxCox transformation:

To compute the expectation conditional on variables X and U, we use Gauss-Hermite quadrature. Let Q be a standard normal distribution. So that $\epsilon = \mu_\epsilon + \sigma_\epsilon Q$. Then the expectation can be rewritten as:

$$\begin{aligned} E[h^{-1}(\alpha X_i + U_i + \epsilon_{ij})|X_i, U_i] &= \int h^{-1}(\alpha X_i + U_i + \mu_\epsilon + \sigma_\epsilon Q_i)\Phi(Q_i)dQ_i, \\ &= \pi^{-1/2} \int h^{-1}(\alpha X_i + U_i + \mu_\epsilon + \sqrt{2}\sigma_\epsilon V_i)\exp(-V_i^2)dV_i \end{aligned} \quad (26)$$

Here $\Phi(Q)$ is the distribution for variable Q where $Q = \sqrt{2}V$. By converting the expectation function into (26), we now have the function in the form $\int f(V)\exp(-V^2)dV$, in this case we can use Gauss-Hermite quadrature and calculate this integral by $\sum_{l=1}^{L} w_l f(V_l)$. Here $l = 1, ..., L$ is the number of nodes used, and w_l is the associated weights for the l^{th} node.

For computing the BoxCox inverse, let $h(x, \lambda) = (x^\lambda - 1)/\lambda$, if $\lambda \neq 0$ and $h(x, \lambda) = \log(x)$, if $\lambda = 0$. If $\lambda \neq 0, h^{-1}(x, \lambda) = (1 + \lambda x)^{1/\lambda}$ By simple algebra,

$$h^{-1}(W_{ij}, \lambda) = [1 + \lambda\{(\alpha X_i + U_i + \epsilon_{ij})\}]^{1/\lambda}. \quad (27)$$

Because of the possibility that the argument on the right hand side of (27) might contain negative values, and given that we are estimating food consumption, having consumed a negative amount of food is not feasible, we will let all negative values to be zero. Therefore we will be using the function

$$h^{-1}(W_{ij}, \lambda) = \max\left[0, 1 + \lambda\{(\alpha X_i + U_i + \epsilon_{ij})\}\right]^{1/\lambda}. \quad (28)$$

1.1.2 BoxCox transformation parameter selection

For this subsection, we will briefly discuss a method for selecting the best BoxCox transformation parameter λ by using a normal score. The idea is to choose the best λ value by comparing how well the BoxCox transformation performs with a normal score. In the next section, we will see how well this selection method works on simulations.

?? suggested that for designs where permutation testing is infeasible and maximum likelihood testing is used, inverse normal transformation can be used, and Blom score is one of the more popular transformation that is widely used in many statistical software. First we create a normal score that will be used for comparison. Let $B = (B_1, ..., B_{99})$ be a series of Blom score as $B_q = \Phi^{-1}((q - 3/8)/(Q + 1/4))$, here $Q = 99$. We then create a variable W_{non-0}, where we record all the non-zero inputs from subjects in W who have more than 2 replicates which contains non-zero inputs, for each subject in W_{non-0}, we find the mean intake for this subject ($\overline{W_{non-0,i}}$) and calculate $\epsilon_i^* = W_{non-0,i} - \overline{W_{non-0,i}}$. Now we compute the first through the 99^{th} percentiles of ϵ^*, let $P = (P_1, ..., P_{99})$ be the percentile values. Define $G(\lambda) = \{h(P_1, \lambda), ..., h(P_{99}, \lambda)\}$, here $h(\cdot)$ is the BoxCox transformation function.

For each λ value, we perform a linear regression of $G(\lambda)$ on B to see how the transformation fits the Blom score. To determine how well the transformation fits, we use the R^2 value from each regression. And we choose the most appropriate λ value based on the largest R^2 value.

Figure 41 looks at the comparison of estimating the λ parameter of the BoxCox transformation. We performed 50 simulations, where each simulation has a certain percentage of zero inputs in the data, the amount of zero inputs in each simulation is represented by the pink curve in percentage form. The gray line shows the true λ value, red line is the λ estimated by the use of Blom score mentioned above and the blue line is the λ estimated using an existing λ estimation function in the software 'R'. We see that existing packages do not work well with data that contains excess zero, therefore for the rest of this paper, we will be selecting λ using this Blom score.

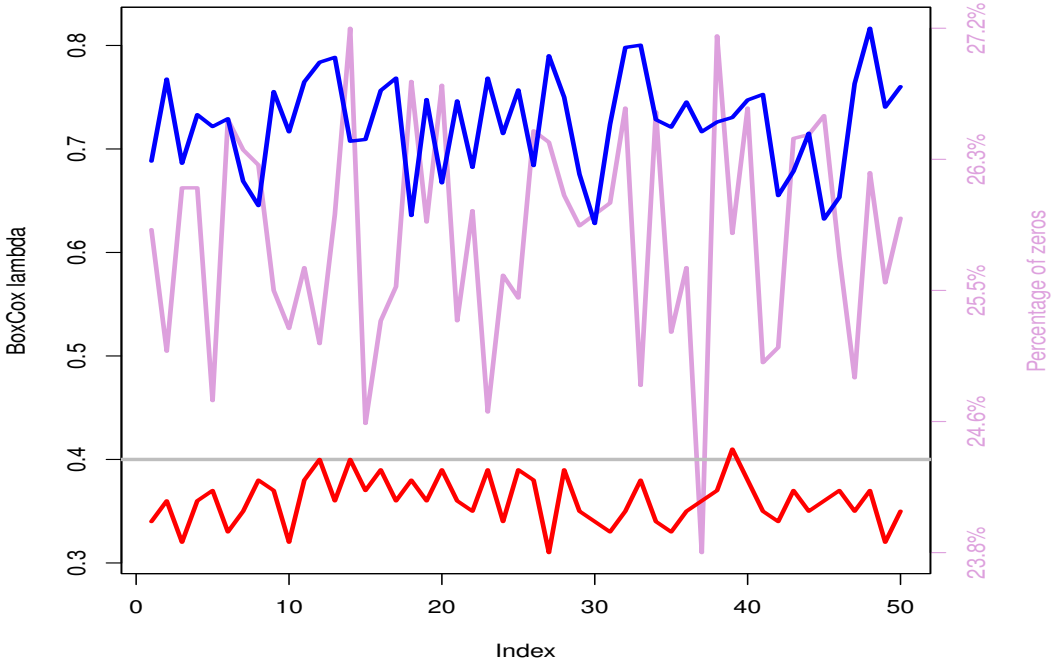


Figure 41: Comparing the estimation of BoxCox transformation parameter λ between an existing BoxCox package from the software 'R' (blue) with this new method containing Blom score (red). Where the grey line is the true transformation parameter. This figure also contains the percentage of zeros from each simulation (light purple) with its corresponding label on the right axis.

1.2 Bootstrapping

We also wish to perform blocked boostrap on the following simulations and on the real data applications. A **blocked bootstrap** was chosen since we wish to keep the replicates for each subject as a block, and bootstrap in respect to each subject instead of each input of the data. To do this we sample with replacement within the subjects to form a bootstrap sample of our simulated dataset. We run this bootstrap sample through the whole process of estimating f_X and f_U, then finally we create large mock samples and estimate f_T for each bootstrap. This step of block bootstrapping is performed for the purpose that we can understand how well this estimation process detects the true density of T.

2 Simulations

We first discuss how to generate our simulated datasets. We generate variables X, U and ϵ, and calculate variable W using these three generated variables. For this section, all simulations will have $j = 4$ replicates and $n = 1000$ subjects.

First we create random generations of 1000 for variables X and U, then we generate 4 sets of random normal values each with a sample size of 1000 for ϵ_j (ϵ_1, ϵ_2, ϵ_3, ϵ_4), this creates the 4 replicates that we need. For each replicate of ϵ we calculate a corresponding set of values for W^*, as $W_j^* = \alpha X + U + \epsilon_j$ for $j = 1, ..., 4$. To determine which W^* value will become a zero input in the observed variable W, we first calculate $\Phi(X)$ and 4 sets of random values using a uniform distribution, each set will have a sample size of 1000 (R_{ij}, $i = 1, ..., 1000$, $j = 1, ..., 4$), one set of random values for each set of replicates. We then compare the values in R_{ij} to $\Phi(X_i)$, if $R_{ij} < \Phi(X_i)$ then $W_{ij} = 0$, otherwise, W_{ij}^* will be transformed into W_{ij} by using the formula of a reverse BoxCox transformation $(\lambda W_{ij}^* + 1)^{1/\lambda}$.

2.1 Simulation 1 results

For the first simulation, let $\alpha = 1$, $X \sim \text{Normal}(\mu_x = 0.7, \sigma_x^2 = 0.4^2)$, $U \sim \text{Normal}(\mu_u = 0, \sigma_u^2 = 0.18^2)$, $\epsilon \sim \text{Normal}(\mu_\epsilon = 0, \sigma_\epsilon^2 = 0.28^2)$. For the BoxCox transformation, let $\lambda = 0.4$. For this simulation the generated variable W will have around 26% of the data set as zero

inputs. We created 321 simulated data sets. As we see later in the figures, we will be plotting the 5th and 95th percentile for the estimated densities, having 321 simulations will allow the 5th and 95th percentile to land of the 16th and 305th simulation respectively, when we organise the simulation results in ascending order. This removes any additional calculations.

We first look at how well we can estimate the BoxCox transformation parameter with about a quarter of the data being zero inputs. For the Blom score λ selection method, we give λ a grid from 0 to 1 where each point is 0.01 width apart. Figure 42 shows a bar chart of the λ values for all 321 simulations, we see that the majority of the simulations considers that 0.37 is the most appropriate λ value, indicating that with a significant amount of zero inputs in the data, the method slightly underestimates the λ value, the estimations also have a 0.02 standard deviation, which tells us that the estimations are fairly consistent, with the largest estimation at 0.43 and the smallest at 0.32. Now due to the long computational time it takes for all the following simulations, we will not be re-estimating λ for each simulation, we will only be using the λ of the first simulation.

The final results that we will be presenting are the density curves of T for each simulation. From (25), we see that it is quite challenging to obtain the distribution of variable T, even if we have known the distributions for the latent variables X and U. This means that in order to get the density of true T ($f_{t,TRUE}$), we will need to create large mock samples from the true densities of X and U, and estimate $f_{t,TRUE}$ using the method mentioned in section 1.1.1, and we do this once for each simulation. Therefore, even though we are calling this as the true density of T, it is actually an average estimation of density T using true densities of X and U. Therefore the results that we will present later on will be the mean density of T using all 321 simulations. And similarly, we obtain the density of the estimated T ($f_{t,Est}$) using the same method as that we did to estimate the true density of T, but this time we will be generating large mock samples from the estimated densities of X and U for each simulation.

For both the true and estimated T, once we obtain the large mock samples, we fit a smooth density curve by using the "bkde" function from the "KernSmooth" package in "R" specifying that the bandwidth is calculated through the "dpik" function also from the same package.

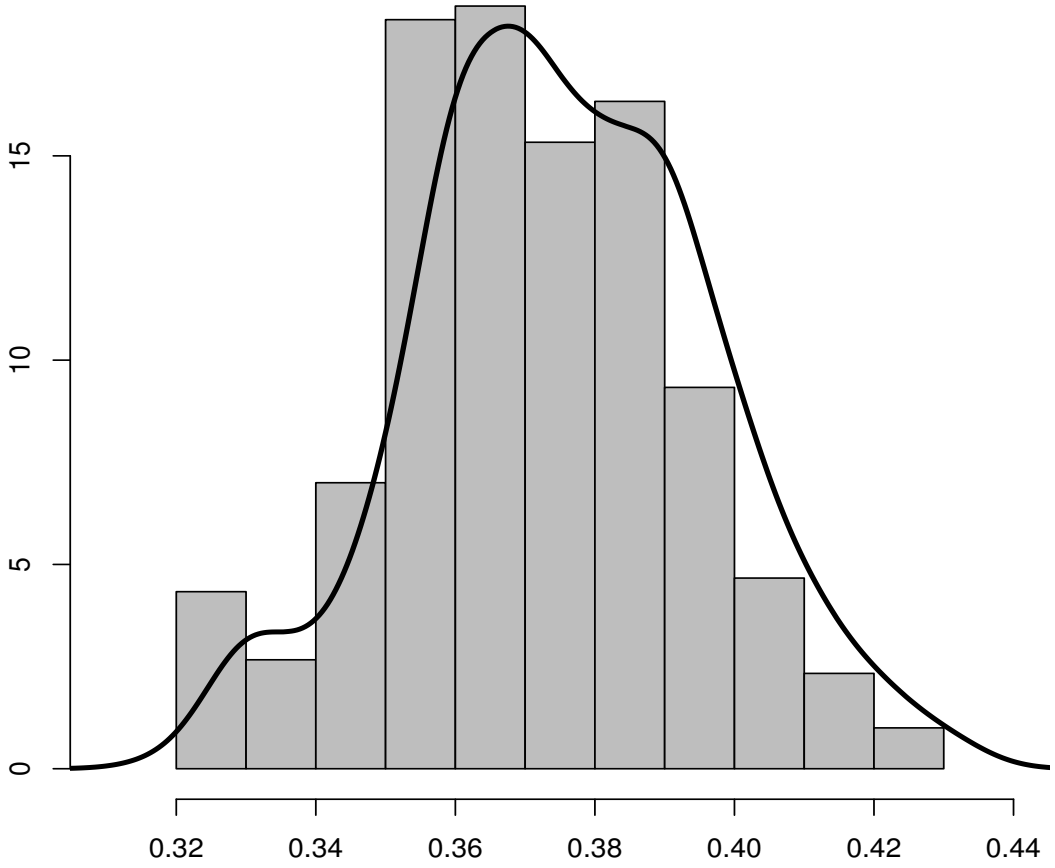


Figure 42: A density plot on the estimation of the BoxCox transformation parameter value λ (solid), where the true $\lambda = 0.4$ (dashed), using a dataset where X is normally distributed and there is on average 26% zero inputs.

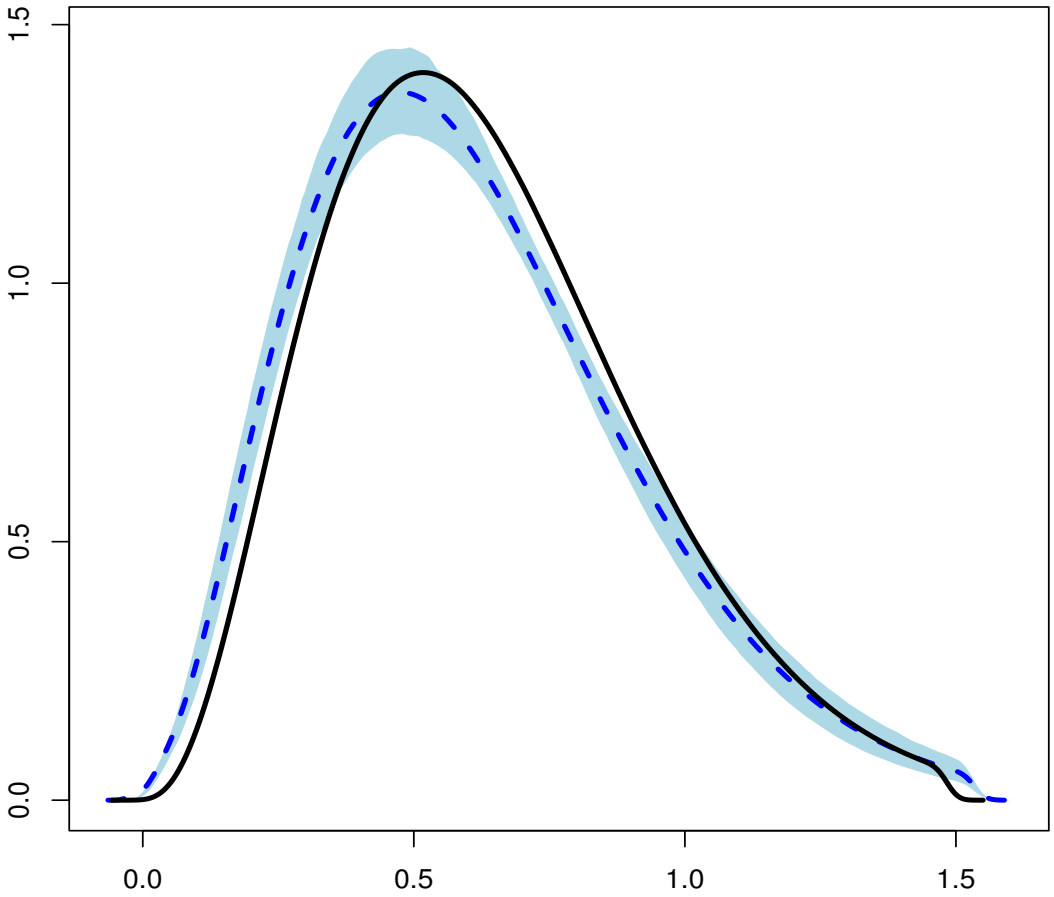


Figure 43: Comparing the mean of the true density curves (solid) to the mean of the estimated density of T (dashed). Here the light blue shade is the 5th to 95th percentile of the estimated density. All density curves are calculated through using the "bkde" function in "KernSmooth" package specifying that the bandwidth is calculated using the "phik" function. For this plot, we have forced $\lambda = 0.4$.

We start by determining how well the density of T can be estimated when the transformation parameter is constant for all simulations, that is forcing the λ value to be 0.4 for all. We do this to see with a consistent transformation, how well does the likelihood work in estimating the density of T. Figure 43 compares the average of the estimated densities of T (dashed) to the average of the estimated true densities of T (solid). We can see that both curves have a similar shape and similar variance, though the estimated average density (f_{Est}) is slightly shifted to the left compared to the true average density (f_{TRUE}). The figure also shows the empirical 5th to 95th percentile of $f_{t,Est}$ (light blue) for all simulations. We see that the shade has more variation around the peak of the curve and much less variation

around the two tails.

Table 1 shows the mean cumulative T values for every 10 percentile point of the true and estimated density of T, also the standard deviation for T at each decile point. For each simulation, 100 blocked bootstraps were performed and the standard error for the decile values was calculated using these bootstraps. We can see that the standard error for the estimated density increases as the decile points go up, we believe that this is because the density forces the density curve to stop at zero on the left tail, whereas there is much more freedom around the other tail, allowing more fluctuation as the percentiles increase. The bootstrapping confirms this idea, as the standard deviation also increases in a similar manner as the estimated standard deviations. Although the standard error result from the bootstrap more conservative than that from estimated standard error. We believe that this will decrease with more samples or an increase in replicates.

The following tables will also contain information on the coverage for both the estimated densities from the simulations, and the estimated densities from the blocked bootstraps. The coverage for the simulated data is calculated by determining whether the average of the estimated density $f_{T,Est}$ is within 95% of each simulated result, and the result is reported as a percentage.

We then look at the same example, but this time we will be using the Blom score as a selection method for the BoxCox transformation parameter for each simulation. Figure 44 shows a comparison between the true density curve (solid) with the average of the estimated density curves (dashed), where the red dashed curve is the estimated density curve where we estimated a transformation parameter for each simulation, and the blue dashed curve is the estimated density curve where we forced a particular transformation value ($\lambda = 0.4$) to all simulations, with the 5th to 95th percentile of the estimated densities as a shade of a lighter color. Looking at the red dashed curve, we see that in general the shape and variation of the density curve are still quite similar to the true density, but is skewed more to the left. This suggests that the estimation of the λ value has a large impact toward the skewness of the curve. Also with a much more variable λ estimation, we have a much wider shade of the 5th to 95th percentile.

Table 2 shows the mean and standard deviation of each decile value for all simulations

	X ∼ Normal(mean=0.7, st. dev=0.4)					
	True	Estimated			Bootstrap	
	mean	mean	st.error	coverage	st.error	coverage
10%	0.289	0.251	1.59×10^{-2}	10.0%	3.62×10^{-2}	48.6%
20%	0.381	0.344	1.80×10^{-2}	22.1%	3.76×10^{-2}	57.1%
30%	0.457	0.422	1.91×10^{-2}	37.1%	4.03×10^{-2}	62.9%
40%	0.530	0.495	2.00×10^{-2}	47.7%	4.24×10^{-2}	62.9%
50%	0.602	0.571	2.10×10^{-2}	57.9%	4.40×10^{-2}	65.7%
60%	0.680	0.652	2.22×10^{-2}	66.4%	4.54×10^{-2}	80.0%
70%	0.769	0.745	2.35×10^{-2}	74.8%	4.71×10^{-2}	82.9%
80%	0.880	0.862	2.62×10^{-2}	81.9%	5.07×10^{-2}	82.9%
90%	1.044	1.038	3.57×10^{-2}	93.4%	6.44×10^{-2}	91.4%

Table 1: The decile values of the true and estimated density of a populations usual intake where the true X and U both have normal distributions. This table also contains the mean and standard deviation from bootstrapping and the coverage percentage when compared to the average true density and the average estimated density. All estimated and bootstrapped densities are analysed with fixed $\lambda = 0.4$

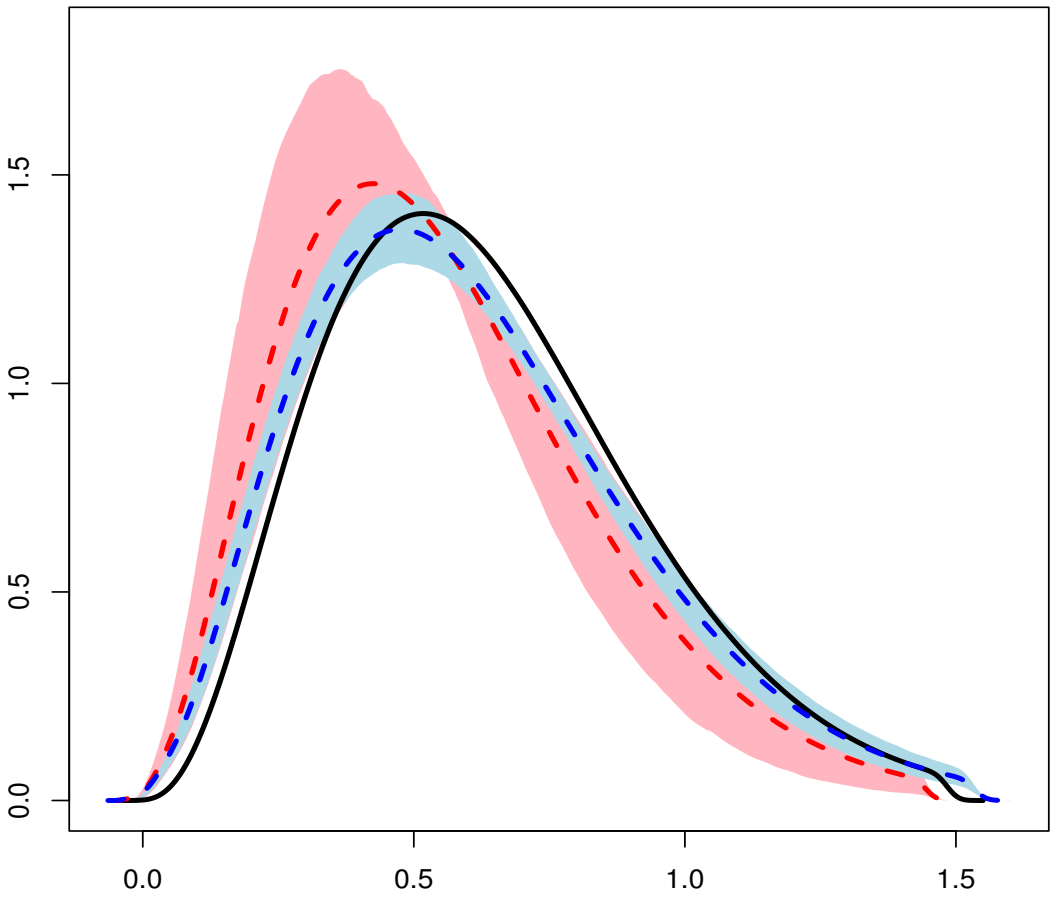


Figure 44: Comparing the true density curve (solid) to the estimated density of T (dashed) where X has a Normal distribution, where the red curve is estimated using an estimated λ, and the blue curve is estimated when $\lambda = 0.4$. Here the light pink shade is the 5th to 95th percentile of the estimated density with estimated λ values and the light blue shade is the 5th to 95th percentile of the estimated density with $\lambda = 0.4$. All density curves are calculated through using the "bkde" function in "KernSmooth" package specifying that the bandwidth is calculated using the "phik" function.

	X ∼ Normal(mean=0.7, st. dev=0.4)					
	True	Estimated			Bootstrap	
	mean	mean	st.error	coverage	st.error	coverage
10%	0.289	0.229	2.49×10^{-2}	20.9%	3.83×10^{-2}	37.1%
20%	0.381	0.314	3.20×10^{-2}	32.7%	4.40×10^{-2}	51.4%
30%	0.457	0.385	3.78×10^{-2}	39.6%	5.00×10^{-2}	57.1%
40%	0.530	0.453	4.32×10^{-2}	45.2%	5.54×10^{-2}	65.7%
50%	0.602	0.522	4.87×10^{-2}	53.3%	6.05×10^{-2}	65.7%
60%	0.680	0.597	5.45×10^{-2}	57.3%	6.58×10^{-2}	71.4%
70%	0.778	0.683	6.09×10^{-2}	62.9%	7.18×10^{-2}	74.3%
80%	0.880	0.792	6.89×10^{-2}	66.0%	7.99×10^{-2}	88.6%
90%	1.044	0.956	8.17×10^{-2}	72.6%	9.58×10^{-2}	88.6%

Table 2: The decile values of the true and estimated density of a populations usual intake where the true X and U both have normal distributions. This table also contains the mean and standard deviation from bootstrapping and the coverage percentage when compared to the average true density and the average estimated density.

where we estimate a λ value. A blocked bootstrap is performed on each simulation and the mean and standard deviation for each decile value is recorded for each bootstrap. Through the bootstrapping, we see that there is a much larger variation on the density estimation, and the bootstrapped density curves are shifted more to the left. A coverage percentage is calculated to see if the bootstrapping can detect the true and estimated densities, it seems that the coverage increases as the decile value increase and the coverage is quite poor on the left tails with more than half of the bootstrap missing the true value at the 10% mark but much better on the right tail. Similar to the previous table, the bootstrap results are more conservative in comparison to the estimated.

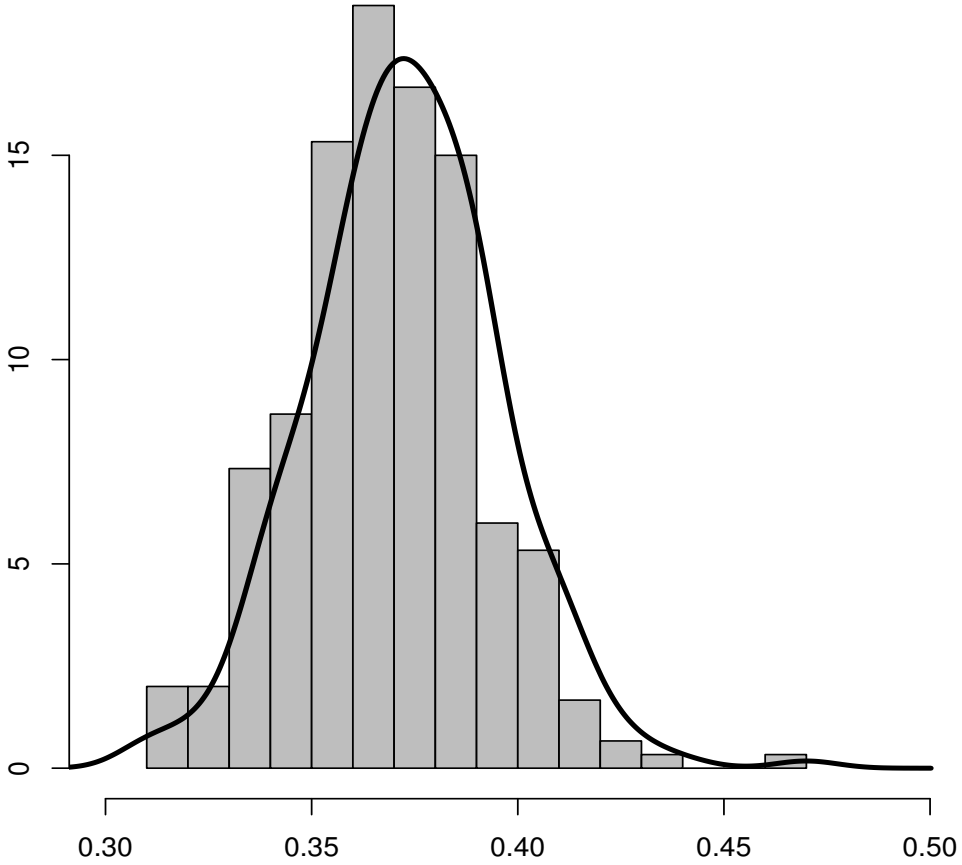


Figure 45: A density plot on the estimation of the BoxCox transformation parameter value λ (solid), where the true $\lambda = 0.4$ (dashed) for example where Y is a Gamma distribution.

2.2 Simulation 2 results

Another simulation was performed, where $\alpha = 1$, $Y \sim \text{Gamma}(\text{shape} = 8, \text{scale} = 1)$, $U \sim \text{Normal}(\mu_u = 0, \sigma_u^2 = 0.18^2)$, $\epsilon \sim \text{Normal}(\mu_\epsilon = 0, \sigma_\epsilon^2 = 0.28^2)$ and $\lambda = 0.4$. For a better comparison between this simulation and the previous one, also to allow better control on the number of zero inputs in the data, we let $X = 0.7 + 0.4 * (Y - \mu_Y)/\sigma_Y$, where the 0.7 and 0.4 are the mean and standard deviation of X from the previous simulation. Once again, 321 simulations are performed where each simulation has around 25% of the data input as zero.

Once again we start by look at how well the λ value is estimated. Figure 45 shows that the λ has a mean around 0.37. Like the previous example, it seems that this BoxCox transformation parameter selection method has underestimated the λ value. But also like

	X ∼ Gamma(shape=8, scale=1)					
	True	Estimated			Bootstrap	
	mean	mean	st.error	coverage	st.error	coverage
10%	0.306	0.244	2.97×10^{-2}	14.6%	6.20×10^{-2}	20.7%
20%	0.378	0.329	3.36×10^{-2}	88.2%	4.56×10^{-2}	90.2%
30%	0.442	0.399	3.46×10^{-2}	94.1%	4.99×10^{-2}	95.4%
40%	0.504	0.465	3.44×10^{-2}	94.1%	5.45×10^{-2}	95.7%
50%	0.572	0.531	3.30×10^{-2}	88.2%	5.88×10^{-2}	87.1%
60%	0.648	0.602	3.02×10^{-2}	63.2%	6.28×10^{-2}	66.6%
70%	0.740	0.683	2.55×10^{-2}	23.4%	6.60×10^{-2}	20.7%
80%	0.862	0.787	2.24×10^{-2}	3.1%	6.77×10^{-2}	5.4%
90%	1.061	0.943	4.74×10^{-2}	5.9%	6.56×10^{-2}	2.1%

Table 3: The decile values of the true and estimated density of a populations usual intake where the true Y has a Gamma distribution and U has a normal distribution. Also the standard deviation using bootstraps and the coverage percentage when compared to the true and estimated densities.

the previous example, the λ values estimated from each simulation has a small variation with a standard deviation of 0.02 and a range between 0.31 to 0.47 with no extreme values.

Figure 46 shows the density estimations for the mean true and estimated density of T. Once again the dashed curve represents the average estimated density of T where we fix $\lambda = 0.4$, and the solid curve represents the average estimated true density of T. We can see that the estimated density of T is also a little skewed to the left compared to its true density. The light blue shade is the 5^{th} to 95^{th} percentile of all the estimated intake densities. We can see that 90% of the data is behaving well. Table 3 shows the deciles of the true and estimated density of T for the case where Y has a Gamma distribution. The tables also contain the standard deviation for each decile value of the estimated densities. We see that the estimated density is mostly biased on the left tail, indicating that this model seems to overestimate the number of individuals who have very small intake values in the long term.

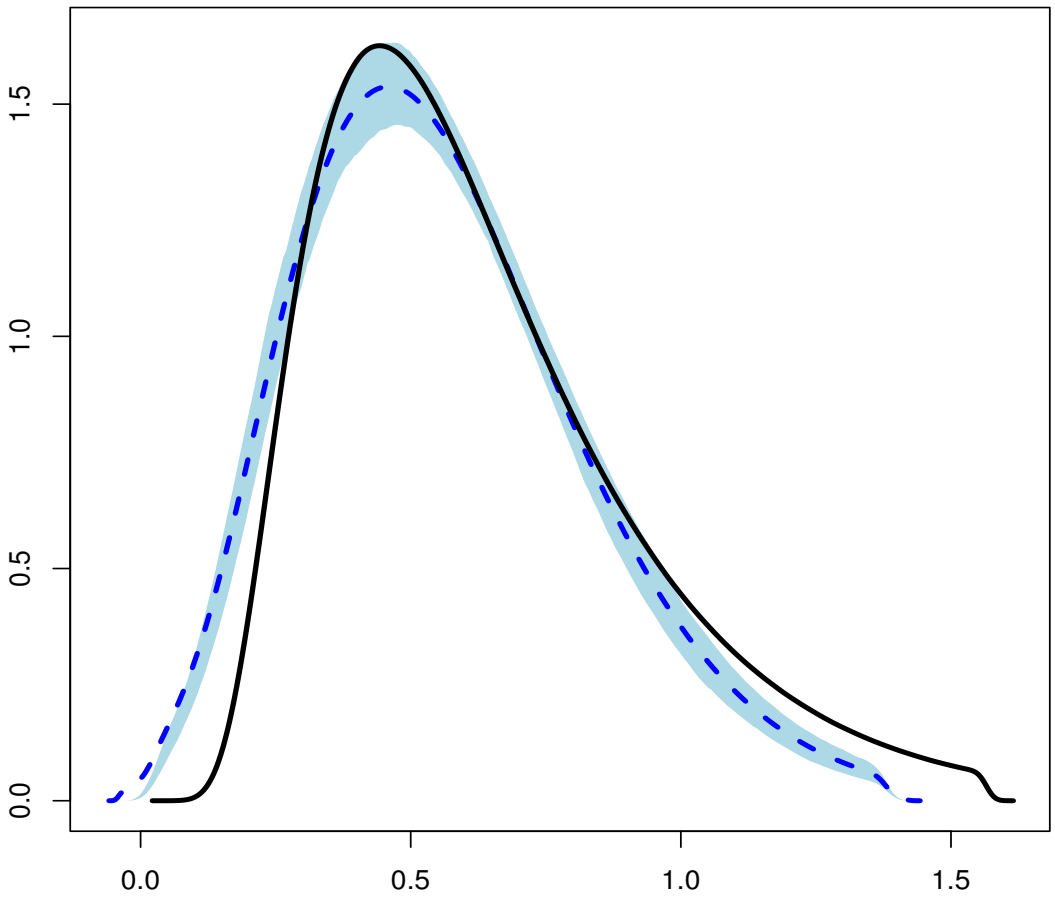


Figure 46: Comparing the true density curve (solid) to the estimated density of T (dashed) for the case where Y has a Gamma distribution. This figure also shows the 5th to 95th percentile of the estimated density (light blue). All density curves are calculated through using the "bkde" function in "KernSmooth" package specifying that the bandwidth is calculated using the "phik" function.

Figure 47 now shows the average true density of T (solid) and the average estimated densities of T (dashed) where the blue line is estimated with a fixed λ term and the red curve requires an estimation on λ for each simulation. For both types of estimation, the 90% range is plotted by shading each dashed curve with the same color but of a lighter shade. We see that for both types of estimations, we seem to have difficulties obtaining an unbiased density estimation on the left tail. Also by looking at the red shade, we see that since the λ values are being underestimated, this seems to contribute greatly to the skewness of the density curve. But in general, the average estimated densities have similar shape and variation as the true density.

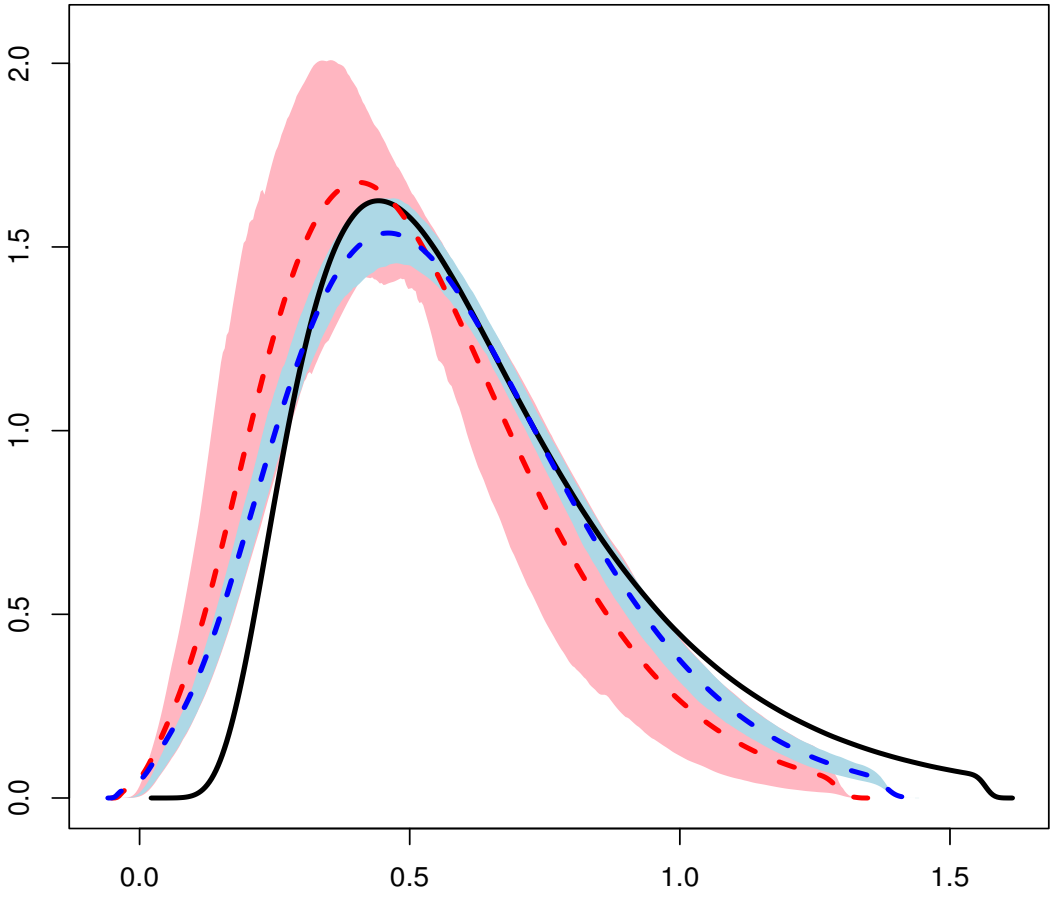


Figure 47: Comparing the true density curve (solid) to the estimated density of T (dashed) where Y has a Gamma distribution. Here the light pink shade is the 5th to 95th percentile of the estimated density with estimated λ values and the light blue shade is the 5th to 95th percentile of the estimated density with $\lambda = 0.4$. All density curves are calculated through using the "bkde" function in "KernSmooth" package specifying that the bandwidth is calculated using the "phik" function.

Part VII

Chapter 6: EATS Data application

1 Introduction

One data that contains zero-inflated data is the Eating at Americans Table Study (EATS), Subar et al. (2001). One of the method which EATS collects data is called 24HR recall, where during the course of a year, each subject was asked to recall their past 24 hour food intake on four different days, each day three months apart. Each food group was recorded using grams, cups or ounces depending on the food that is recorded. If a food group was not consumed in that particular 24 hours, the value for that food group is zero.

Such a data collection method has it's pros and cons. One of the most obvious con is that since it only takes input from such a small time frame, many of the data will contain zero inputs. For example, a subject may have had some beef steak for dinner, then their seafood, pork and lamb input will most likely be zero for that particular recall. With only four recalls per subject, and with a good portion of the input being zeros, the 24HR recall in EATS is not a good representation of each subjects long term consumption pattern. But 24HR recall also have a obvious pro, which is recalling memory from such a small time frame allows us to obtain much more accurate data. Ofter all, when asked to recall precisely how many cups of water you ingested or what exactly is in the salad you just ate, it is much easier to get an accurate response if the recall time frame is only 24 hours. Using the density estimation method from the previous chapter, we wish to reduce the cons that occurs from the 24HR recalls, and make use of the accuracy in EATS data to obtain an accurate estimate of the populations long-term average daily usual intake pattern. In the EATS data that is available to me, 965 subjects participated in this questionnaire, and all subjects were able to complete all four recalls. It should also be mentioned that all subjects are between the age of 20 and 70.

In this EATS data, the recalls are organised in categories, such as alcohol, dark green vegetables, deep yellow vegetables, total vegetables, total grains, organs and so on. All

categories will have a certain amount of zeros. Although most categories will have roughly 20% to 40% of zero inputs, some categories will have very small or extremely large amounts of zeros. For example, almost all subjects will consume some sort of grains daily, which means that the total grains category will have a very small percentage of zero inputs (0.67%), and most participants will not choose to have organs as part of their everyday meal, therefore in the organ category, nearly all input are zeros (99.1%). This density estimation method for excess zeros performs with the understanding that: a) this is a zero-inflated data, and b) that all participants are consuming the food of interest regularly in the long run. Therefore data such as total grains will not be ideal for using this estimation method, since although a small amount of inputs are zero, total grains can be considered more as a continuous data than a zero-inflated data. Also we will not use this density estimation method on data for food groups such as organs, with the large percentage of zeros, we can assume that a significant proportion of the subjects are non-consumers on that particular food. For this chapter, we will be choosing to analyse food groups which are consumed regularly but not necessarily consumed daily.

For this data analysis, we looked at three food groups with three drastic different percentage of zero inputs: alcohol, total fruits, and total vegetables. In order to get more stable results, we standardized the data from all three food groups before doing any analysis. EATS not only records short term food intakes, but also records other relevant variables such as each subject's energy expense for each recall (in the units kilocalories). Therefore we standardized the data by dividing each subject's food intake by their energy expense per thousand kilocalories. For each food group, we will be looking at the density curve of the populations long term average daily usual intake per thousand kilocalories spent.

Since all of these food groups have a number of zero inputs, we will be using the modified NCI method (equation 21 and 22) in hopes that we can obtain an accurate pattern of how much or how little the population consumes the food of interest on average each day.

	number of replicates with zero inputs				
	0 rep	1 rep	2 reps	3 reps	4 reps
Alcohol	80 (8.29%)	69 (7.15%)	100 (10.36%)	153 (15.85%)	563 (58.34%)
Total Fruit	519 (53.78%)	228 (23.63%)	144 (14.92%)	56 (5.80%)	18 (1.87%)
Total Vegetable	834 (86.42%)	119 (12.33%)	11 (1.14%)	1 (0.10%)	0 (0.00%)

Table 4: Separating the subject into how many recalls containing zeros, where the percentage is presented in brackets. This is done for all data which we will be using for analysis.

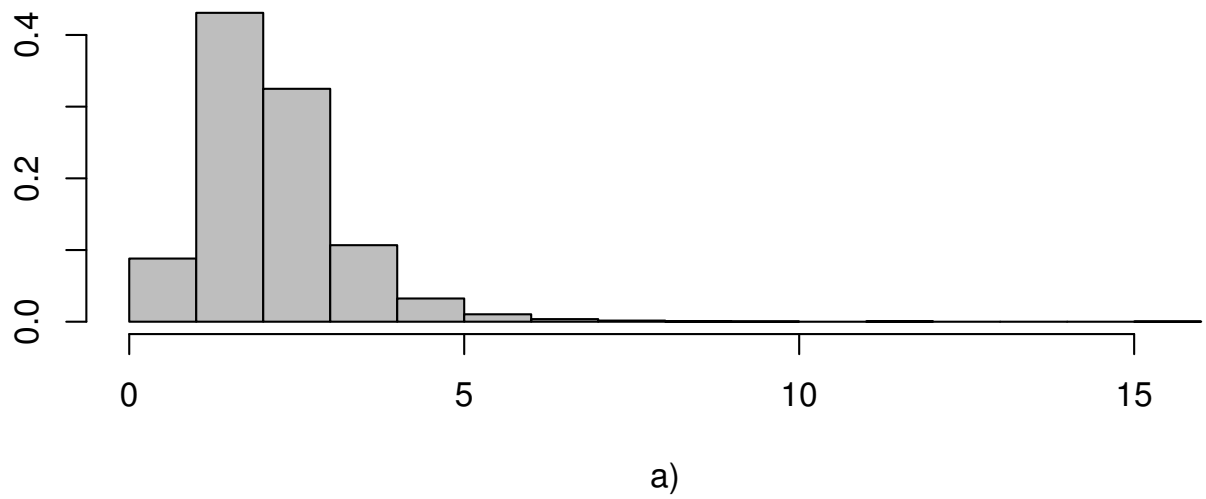
1.1 Energy

We first start with a rough summary on the energy recorded in EATS. For each participant, their energy spent daily is recorded along with their detailed food intake for each recall.

Figure 48 shows the histograms of each subjects energy spent, the unit is in 1000 kilocalories. Figure a) is a histogram of all the energy recorded from all recalls and all subjects. We see that the majority of energy spent is under 5000 kilocalories per day with a mean around 2100 kilocalories. The least energy spent within all replicates is only around 10 kilocalories and the most energy spent is around 15400 kilocalories per day. As we can see that the energy spent per day varies quite a bit, so we might get a better understanding towards how active the participants are by looking at the average energy spent between the recalls for each subject. Figure b) shows the histogram of the average energy spent per day for all subjects. We can see that this figure has less variation with no extremely large or small values. On average, the mean energy spent between subjects is still around 2100 kilocalories. Now the subject with the least active life style spends on average 310 kilocalories per day, and the most active subject spends on average 8700 kilocalories per day.

1.2 Alcohol

For alcohol intake, the unit used to measure alcohol is in grams, this data records the amount of pure alcohol each subject has ingested for the past 24 hours from only alcoholic beverages. For alcohol, the initial thought is that most people will under estimate how much they drink,



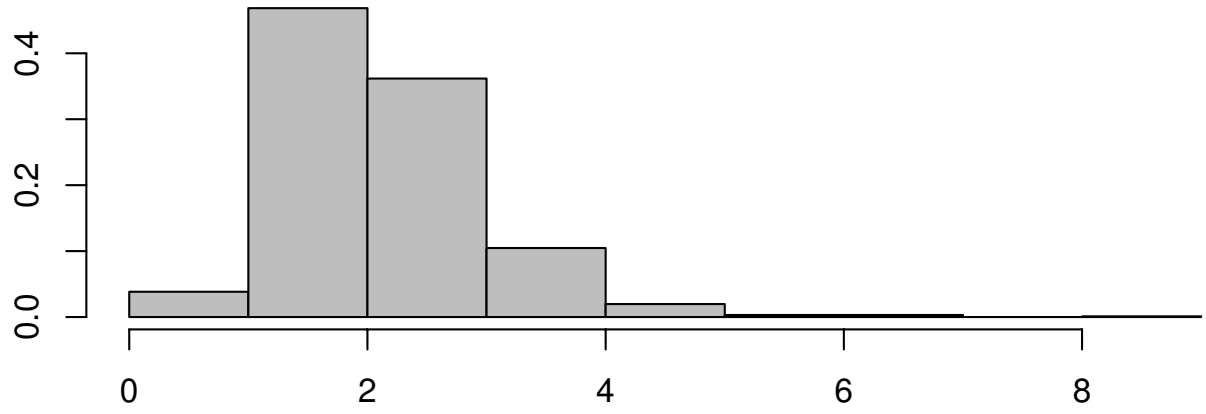


Figure 48: a) a histogram on all the energy expense recorded with in unit of 1000 kilocalories, b) a histogram on the mean energy expense for each participant.

especially when they binge drink, and this will create a "human" error problem.

For this data, we made modifications before standardising the data. The modification is to let any input that is less than 1.42 become 0. We do this step because 1.42 grams of pure alcohol is a very small amount, it is equivalent to one eighth of a standard beer in the US. Therefore for any recall that reports a consumption of pure alcohol less than 1.42, we believe that those recalls are not recalls on how much the participants drank for the past 24 hours, but may be consumed in some other form such as a trifle or a liqueur filled chocolate, and that results from these recalls will not be beneficial in obtaining an unbiased estimate of the populations long-term alcoholic drinking pattern.

After converting all the alcohol recall values that are less than 1.42 grams into zeros, around 77% of the data entries are zeros. From table 4 we see that although there is over half of subjects have stated they did not consume any alcoholic beverage for the past 24 hours in all of their recalls, we will still assume that they are alcoholic drinkers in the long run.

Looking at only the non-zero inputs from the standardised data, we are able to obtain some basic statistics. Where the average alcohol consumption is 15 grams per thousand kilocalories spent with the smallest recall as 0.5 grams per thousand kilocalories and the largest recall as 80 grams per thousand kilocalories.

Figure 49 shows the density curve of the populations long term average daily usual average daily intake per thousand kilocalories spent of pure alcohol from alcoholic beverages without converting any overly small values to zero. We see that the density curve is unimodal where the mode of the density is around 5 grams of pure alcohol per thousand kilocalories spent. From the previous subsection, we see that a very rough average of the daily energy spend is around 2000 kilocalories. Assuming that the average person will be spending 2000 kilocalories per day, this means that the mode of the general populations long term usual daily average intake of pure alcoholic from alcoholic beverages is around 10 grams, this indicates that most of the population between age 20 - 70 tend to have a moderate amount of alcoholic beverages and only a small number of people drink to excess. But we can also see that the curve can go all the way up to more than 25 grams per thousand kilocalories, indicating that there is a small proportion of the population that has a habit of drinking to

excess.

Figure 50 shows the long term usual daily average intake of pure alcohol after converting any amount less than 1.42 grams into 0. We see that there is now a large peak around the 0 value. Indicating that although the measurement error model used does not account for non-drinkers, there is still a large proportion of the population that does not regularly drink alcohol. Compared to figure 49, it is less skewed. And also now the curve only extends to around 12 grams of alcohol per thousand kilocalories, showing favorable results that the population doesn't actually have subjects that regularly drink so much.

Alcohol has a larger percentage of zero inputs in comparison to other foods, therefore we expect the mode to be close to 0, but we can also see that using the modified NCI method for excess zeros, it has shrunk the max alcohol consumption down quite a bit, giving us a better understanding of how much our society drinks on average.

1.3 Total Fruits

For total fruits, the data records the total amount of daily consumption from every type of fruit, the results is in the unit cups. This data contains 20% zero inputs. The initial thought is that fruits are often considered as nutritious and is recommended to be consumed daily, but not everyone reaches for fruit on a daily basis. But as many people like to show that they are more healthy than they really are, we believe that they may over estimate their consumption of fruits.

According to the Healthy Eating Index (HEI) (Britten et al., 2006), for every thousand kilocalories spent each day the recommended amount of fruits consumed is 0.8 cups. So for total fruits, 0.8 will be our cut-off point in determining the proportion of population who do not consume enough fruits daily. This value will also assist us in standardising the data. To standardise the data, we divide the results from each recall by 0.8 times the corresponding energy expense. From all the non-zero inputs of the standardised data, the average daily consumption is around 1.3 cups of fruit per thousand kilocalories with 1.8 cups per thousand kilocalories as the most consumption and 4.7×10^{-6} cups per thousand kilocalories as the least non-zero consumption. As the least non-zero consumption is so close to zero, we can

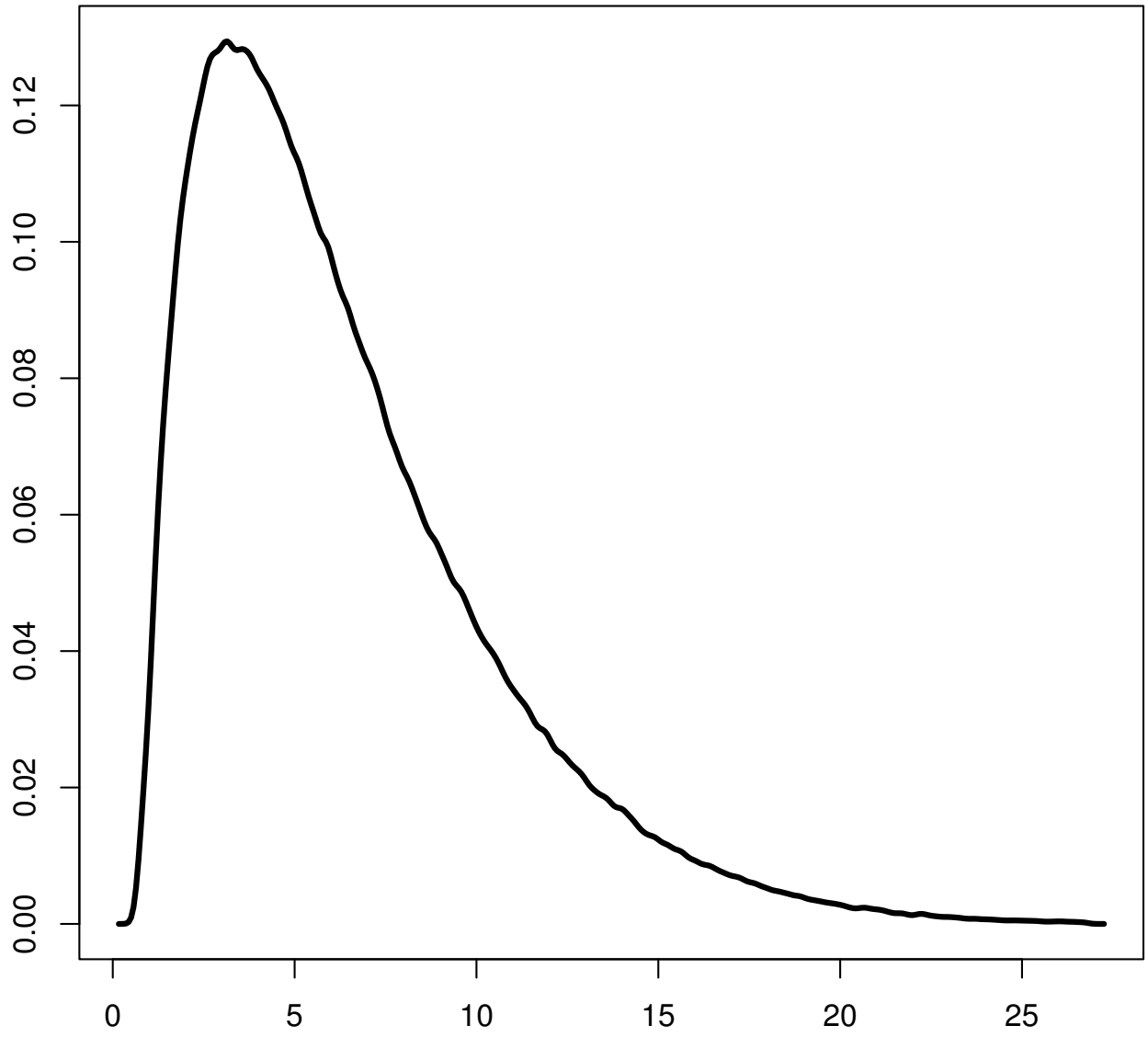


Figure 49: The estimated density curve of a populations long term usual average intake of pure alcohol from alcoholic beverages (in grams) with only energy adjustment.

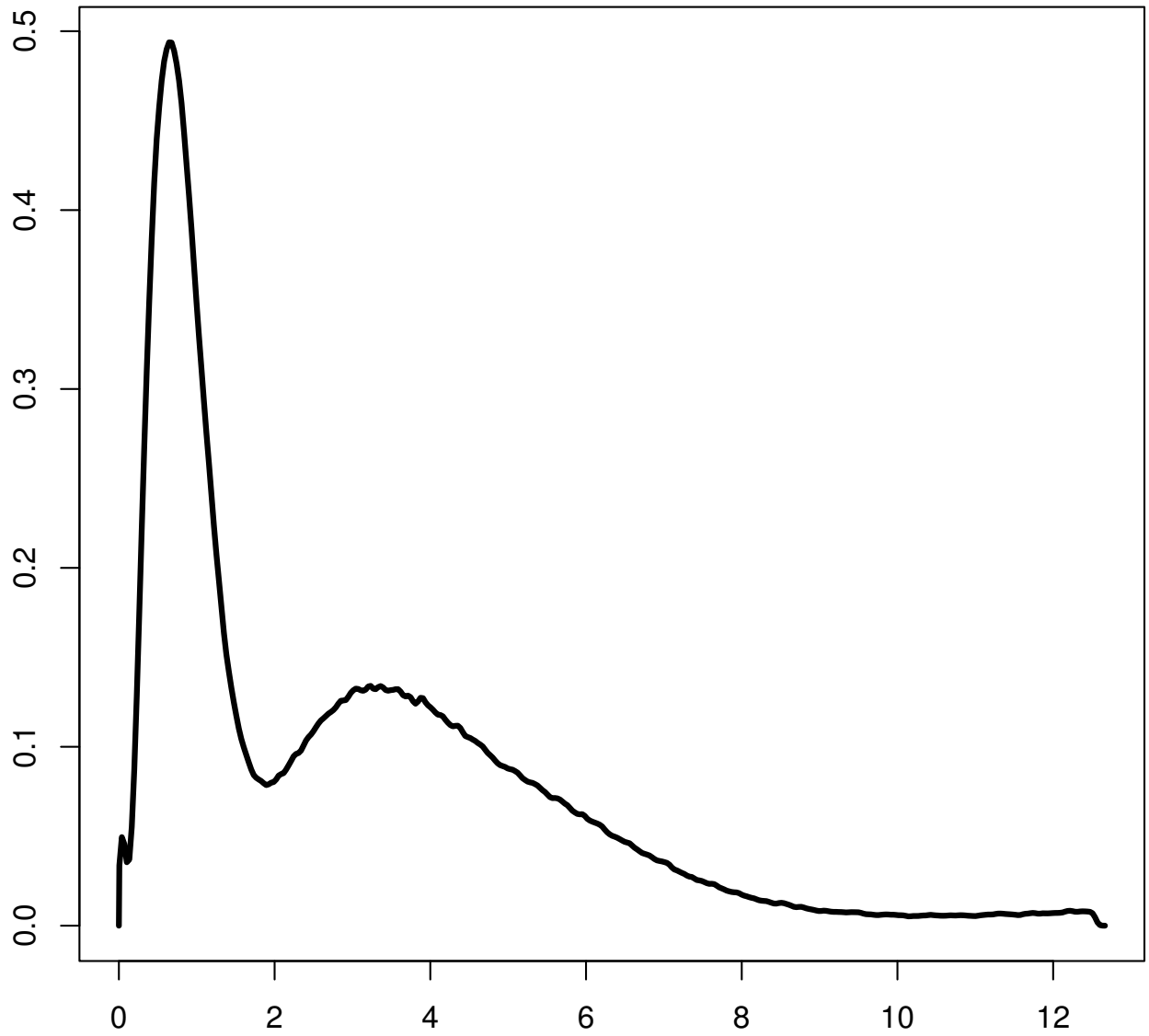


Figure 50: The estimated density curve of a populations long term usual average intake of pure alcohol from alcoholic beverages (in grams), where all amount less than 1.42 grams in converted to zero.

speculate that this may have been obtained though other foods such as blueberries from a blueberry muffin or dehydrated berries from a granola bar.

Figure 51 shows the density curve of the populations long term average daily usual intake of total fruits. Similar to the alcohol density plots, the estimated density curve for total fruits is also a unimodal curve. This curve indicates that the mode of the population has about half a cup of fruit per thousand kilocalories spent, and nearly no one consumes on average more than 1.5 cups per thousand kilocalories spent per day. Using the Healthy Eating Index (HEI) which suggested that for every thousand calories spent each day the recommended fruit consumption is 0.8 cups. We can see that around 74% of the population do not consume the recommended amount fruit each day. As we suspected, many people over estimated how much fruit they consumed.

1.4 Total Vegetables

Total vegetables records the daily consumption amount of all types of vegetables, this should include leaf vegetables, root vegetables and legume. Similar to the fruits, total vegetables are measured in cups. For this data, there is only 3.7% zeros. The initial thoughts on vegetables is that most people would have some variety of vegetable as part of their meals, therefore the percentage of zero inputs for total vegetables should be low, but similar to fruits, many people may over estimate how much they consume daily due to vegetables being full of essential nutrition.

Once again we chose the cut-off point using HEI. According to HEI the recommended consumption amount for vegetables is 1.1 cups daily per thousand kilocalories. Following the example of the fruits, we standardise total vegetables by dividing each recall by 1.1 times the corresponding energy expense. For the non-zero values of the standardised data, the mean is 1.8 cups per thousand kilocalories with the largest value as 12.5 cups per thousand kilocalories and the smallest non-zero value as 5.5×10^{-4} cups per thousand kilocalories. Once again since the smallest non-zero value is so small, we believe that this particular recall is recording the consumption of another food which contains a small amount of vegetables.

Figure 52 shows the density curve of a populations long term usual daily average intake

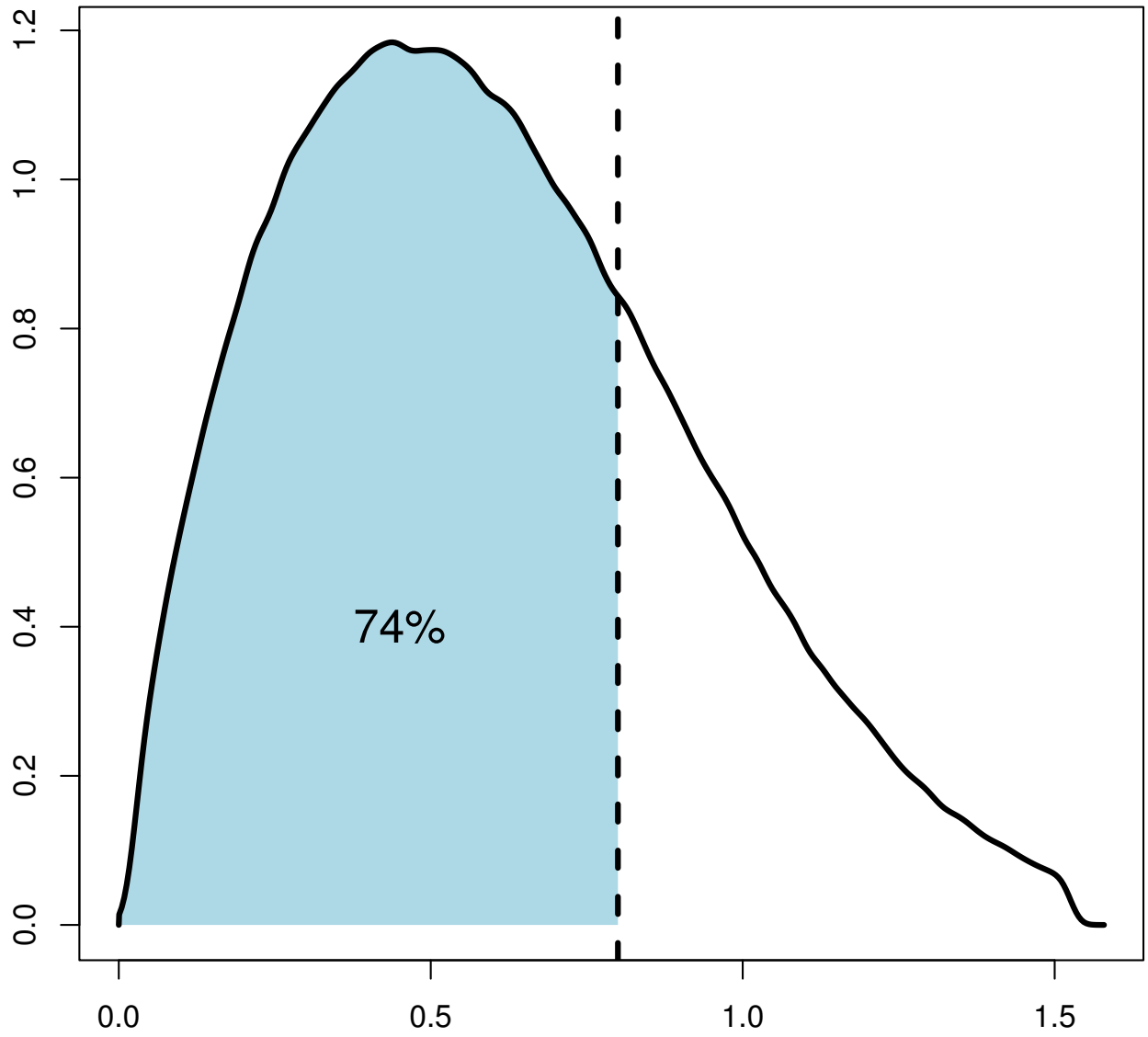


Figure 51: The density curve of the populations long term usual average intake of total fruits. Where the light blue shaded area indicates the percentage of population that does not reach the daily recommended amount.

of total vegetables. The estimated density curve for the total vegetables indicates that the mode of the population has about 1 cup of vegetable per thousand kilocalories spent, most people consume at least half a cup per thousand kilocalories and the population consumes on average no more than 1.5 cups per day per thousand kilocalories. Using 1.1 cups of vegetable for every thousand kilocalories spent as the cut-off point, this suggests that 65% of the population do not consume up to the recommended amount of vegetables each day. Once again our initial thought has been confirmed. This food is a very good example on the impact of ignoring the excess zero problem, as we only have around 4 percent of zero inputs, if we ignore the excess zero problem, the mean would be closer to 1.8 than 1, indicating that we will be still be over estimating the problem.

2 Bootstrapping

To understand how well the density estimation has worked, we created a series of blocked bootstrap for each food variable. We choose to perform a block bootstrap since each participant has 4 replicates and these replicates should not be separated in a bootstrap simulation. If we simulate without binding the 4 replicates together, we may create participants that have very different consumption patterns, and that would not be an accurate representation of the current data. To perform this block bootstrap, we sampled with replacement within the subjects in the data keeping each subjects 4 recalls together as a block. We do this to each standardised data. Then for each newly sampled bootstrap data, we perform the whole density estimation, by first estimating the parameters α, μ_ϵ, σ_ϵ, θ_{kX} and θ_{kU}. Using these parameters, we obtain f_X and f_U for each bootstrapped data and ultimately we estimate f_T for each bootstrap. We perform this bootstrap 100 times and we present the values for the cumulative grid value at three percentage points (10%, 50% and 90%), also the 90% confidence interval range are calculated using the mean and standard deviation of these three percentages.

We can see in table 5 that there is much more fluctuation at the two ends of the curve, than the middle for all three datasets. This corresponds to the simulations, indicating that it is a lot harder to obtain a good estimate of the two extremes.

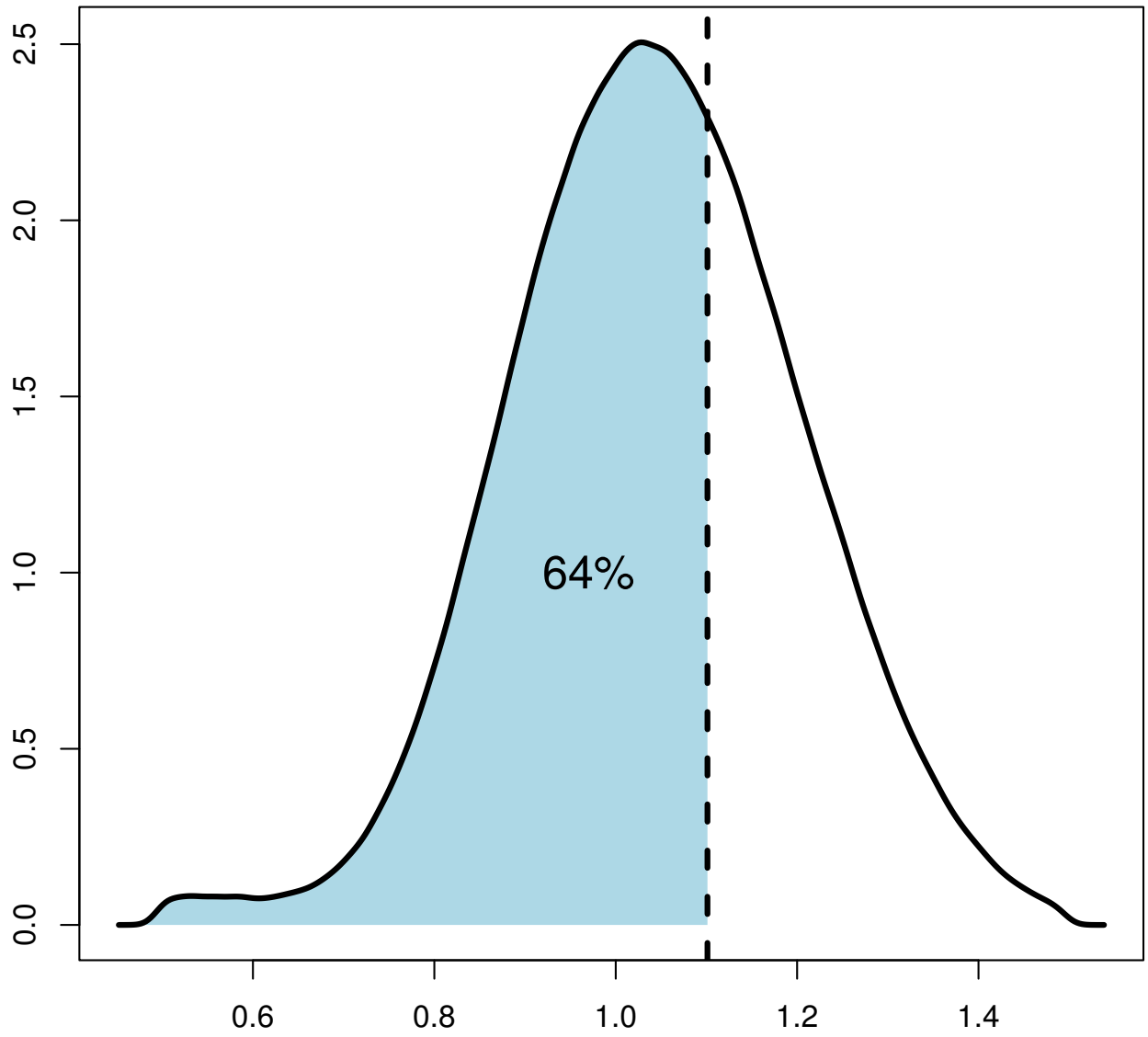


Figure 52: The density curve of the populations long term usual average intake of total vegetables, where the shaded area indicates the percentage of population that does not meet the daily recommended amount.

		10%	50%	90%
Alcohol	Estimated density	2.13	5.43	11.900
(grams)	95% C.I	(1.73, 2.54)	(5.00, 5.86)	(10.489, 13.31)
Total	Estimated density	0.20	0.56	1.048
Fruits	95% C.I	(0.03,0.36)	(0.45,0.67)	(0.573,1.523)
Total	Estimated density	0.840	1.04	1.249
Vegetables	95% C.I	(0.70,0.98)	(0.70, 1.39)	(0.78, 1.71)

Table 5: The 10%, median and 90% values for each food density and the 90% confidence interval for these three percentage points.

Part VIII

Future plans

For our thesis we have managed to develop a method which can obtain a distribution of a populations nutritional consumption pattern. This method utilised existing measurement error models and existing techniques such as log likelihood, Chebyshev quadrature, Hermite polynomials and so on.

We started with a simple classical error model which works on continuous data. This model would be a good candidate for nutritional data such as protein intake, fibre intake and iron intake. These are nutrition that we are sure to consume everyday and therefore will work well with a classical error model. We compared our method with an existing method (KDD) and concluded that depending on the type of data, one may work better than the other, we have also pointed out the flexibility of our method and we can customise it to our liking such as allowing this method to be non-parametric or to use different measurement error models. Although it is not demonstrated in our thesis, we believe that we can easily convert a simple classical error model into a Berkson error model in our methodology depending on which model is needed to analyse the data. Though we did increase the complexity of the classical error model into a two part error model for excess zero data.

Although in our introduction, we mentioned two types of collection method for nutritional data, we worked only with 24 hour recall for our thesis, this is because we were interested in how we well we can analyse zero inflated data, and see how we can get around so many zero input and still produce an unbiased estimate of the distribution of a populations nutritional intake pattern. For our method using a two part measurement error model, we can use it to analyse 24 hour food intake data such as alcohol intake, meat intake and dairy intake, since although we regularly consume such food, they may not be consumed daily. Through our simulation works, we saw that in general the estimation for the distribution of a populations daily average food intake is under estimated. This is largely due to the fact that the two part model contains a BoxCox transformation, and given the large amount of zero inputs, the transformation parameter is often under estimated.

For future extensions, we would like to develop a method to solve this under estimation problem. Another problem we would like to solve in the future is the stability of the optimisation process, as we have seen in the classical error model case, a few simulations have unstable results, and the amount of unstable simulations have increased as the measurement error model becomes more complicated.

As we have mentioned through out the thesis, we are working with data with equal replicates, that is all the subjects have the same amount of replicates, but as data collection goes, this is an unrealistic expectation, since a lot of the times a subject will stop communicating and therefore will have less replicates that the rest of the subjects, the information that this subject provides will still be useful and will help increase the accuracy of the estimations, therefore as another future extension, we wish to develop our method more to allow unequal number of replicates for each subject.

When working with zero inflated data, one assumption we had was that all subject are consumers of the food of interest, that is if we analyse alcohol intake, we are assuming all subject at some point drinks alcohol periodically, and if we analyse meat intake, we are assuming none of the subjects are vegetarian or vegan. With no direct information of each subjects dietary choices, we can not differentiate the non-consumers from the "small amount" consumers. Therefore as a future extension, we wish to work with an even more complicated measurement error model that will be able to differentiate the non-consumers and obtain a more accurate distribution for the population.

For future extensions, we would like to develop a method to solve this under estimation problem. Another problem we would like to solve in the future is the stability of the optimisation process, as we have seen in the classical error model case, a few simulations have unstable results, and the amount of unstable simulations have increased as the measurement error model becomes more complicated.

As we have mentioned through out the thesis, we are working with data with equal replicates, that is all the subjects have the same amount of replicates, but as data collection goes, this is an unrealistic expectation, since a lot of the times a subject will stop communicating and therefore will have less replicates that the rest of the subjects, the information that this subject provides will still be useful and will help increase the accuracy of the estimations,

therefore as another future extension, we wish to develop our method more to allow unequal number of replicates for each subject.

When working with zero inflated data, one assumption we had was that all subject are consumers of the food of interest, that is if we analyse alcohol intake, we are assuming all subject at some point drinks alcohol periodically, and if we analyse meat intake, we are assuming none of the subjects are vegetarian or vegan. With no direct information of each subjects dietary choices, we can not differentiate the non-consumers from the "small amount" consumers. Therefore as a future extension, we wish to work with an even more complicated measurement error model that will be able to differentiate the non-consumers and obtain a more accurate distribution for the population.

References

Achilleos, A. and Delaigle, A. (2012). Local Bandwidth Selectors for Deconvolution Kernel Density Estimation. *Statistics and Computing*, 22, 563–577.

Armstrong, B. (1985). Measurement Error in the Generalised Linear Model. *Communications in Statistics-Simulation and Computation*, 14, 529–544.

Berkson, J. (1950). Are There Two Regressions? *Journal of the American Statistical Association*, 45, 164–180.

Berry, S. M., Carroll, R. J., and Ruppert, D. (2002). Bayesian Smoothing and Regression Splines for Measurement Error Problems. *Journal of the American Statistical Association*, 97, 160–169.

Block, G., Wakimoto, P., Jensen, C., Mandel, S., and Green, R. R. (2006). Peer reviewed: Validation of a food frequency questionnaire for hispanics. *Preventing chronic disease*, 3.

Britten, P., Marcoe, K., Yamini, S., and Davis, C. (2006). Development of Food Intake Patterns for the MyPyramid Food Guidance System. *Journal of nutrition education and behavior*, 38, S78–S92.

Carroll, R. J., Chen, X., and Hu, Y. (2010). Identification and Estimation of Nonlinear Models Using Two Samples with Nonclassical Measurement Errors. *Journal of Nonparametric Statistics*, 22, 379–399.

Carroll, R. J. and Hall, P. (1988). Optimal Rates of Convergence for Deconvolving a Density. *Journal of the American Statistical Association*, 83, 1184–1186.

Carroll, R. J., Küchenhoff, H., Lombard, F., and Stefanski, L. A. (1996). Asymptotics for the SIMEX Estimator in Nonlinear Measurement Error Models. *Journal of the American Statistical Association*, 91, 242–250.

Carroll, R. J., Ruppert, D., Crainiceanu, C. M., Tosteson, T. D., and Karagas, M. R. (2004). Nonlinear and Nonparametric Regression and Instrumental Variables. *Journal of the American Statistical Association*, 99, 736–750.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC press.

Carroll, R. J. and Stefanski, L. A. (1990). Approximate Quasi-likelihood Estimation in Models with Surrogate Predictors. *Journal of the American Statistical Association*, 85, 652–663.

Castro, M., Bolfarine, H., and Galea, M. (2013). Bayesian Inference in Measurement Error Models for Replicated Data. *Environmetrics*, 24, 22–30.

Chan, H. T. Welcome to the Harvard T.H. Chan School of Public Health Nutrition Department's File Download Site. `https://regepi.bwh.harvard.edu/health/nutrition.html`.

Chen, X. and Pouzo, D. (2009). Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals. *Journal of Econometrics*, 152, 46–60.

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical association*, 89, 1314–1328.

Delaigle, A. and Gijbels, I. (2004a). Bootstrap Bandwidth Selection in Kernel Density Estimation From A Contaminated Sample. *Annals of the Institute of Statistical Mathematics*, 56, 19–47.

Delaigle, A. and Gijbels, I. (2004b). Practical Bandwidth Selection in Deconvolution Kernel Density Estimation. *Computational Statistics & Data Analysis*, 45, 249–267.

Delaigle, A. and Gijbels, I. (2007). Frequent Problems in Calculating Integrals and Optimizing Objective Functions: A Case Study in Density Deconvolution. *Statistics and Computing*, 17, 349–355.

Delaigle, A. and Hall, P. (2016). Methodology for Non-parametric Deconvolution When the Error Distribution Is Unknown. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 231–252.

Devanarayan, V. and Stefanski, L. A. (2002). Empirical Simulation Extrapolation for Measurement Error Models with Replicate Measurements. *Statistics & Probability Letters*, 59, 219–225.

Dodd, K. W., Guenther, P. M., Freedman, L. S., Subar, A. F., Kipnis, V., Midthune, D., Tooze, J. A., and Krebs-Smith, S. M. (2006). Statistical Methods for Estimating Usual Intake of Nutrients and Foods: A Review of the Theory. *Journal of the American Dietetic Association*, 106, 1640–1650.

Dwyer, J., Picciano, M. F., Raiten, D. J., Committee, S., et al. (2003). Collection of Food and Dietary Supplement Intake Data: What We Eat in America–NHANES. *The Journal of nutrition*, 133, 590S–600S.

Eckert, R. S., Carroll, R. J., and Wang, N. (1997). Transformations to Additivity in Measurement Error Models. *Biometrics*, 52, 262–272.

Fan, J. (1992). Deconvolution with Supersmooth Distributions. *Canadian Journal of Statistics*, 20, 155–169.

Fuller, W. A. (2009). *Measurement Error Models*. John Wiley & Sons.

Gauss, C. F. (1815). *Methodus Nova Integralium Valores per Approximationem Inveniendi*. apvd Henricvm Dieterich.

Geman, S. and Hwang, C. (1982). Nonparametric Maximum Likelihood Estimation by the Method of Sieves. *The Annals of Statistics*, 10, 401–414.

Grenander, U. (1981). *Abstract Inference*. Wiley New York.

Hall, P. and Ma, Y. (2007). Semiparametric Estimators of Functional Measurement Error Models with Unknown Error. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 429–446.

Hermite, C. (1864). *Sur un Nouveau Développement en Série des Fonctions*. Mallet-Bachelier.

Hernández-Avila, M., Romieu, I., Parra, S., Hernández-Avila, J., Madrigal, H., and Willett, W. (1998). Validity and reproducibility of a food frequency questionnaire to assess dietary intake of women living in mexico city. *Salud publica de Mexico*, 40, 133–140.

Holmes, C. C. and Mallick, B. K. (2003). Generalized Nonlinear Modeling with Multivariate Free-knot Regression Splines. *Journal of the American Statistical Association*, 98, 352–368.

Hu, Y. and Schennach, S. M. (2008). Instrumental Variable Treatment of Nonclassical Measurement Error Models. *Econometrica*, 76, 195–216.

Institute, N. C. Diet History Questionnaire: Web-based DHQ. `http://riskfactor.cancer.gov/DHQ/webquest/index.html`.

Institute, N. C. Dietary Assessment Primer. `https://dietassessmentprimer.cancer.gov/profiles/recall/`.

Institute, N. C. NutritionQuest. `http://www.nutritionquest.com/`.

Kipnis, V., Midthune, D., Buckman, D. W., Dodd, K. W., Guenther, P. M., Krebs-Smith, S. M., Subar, A. F., Tooze, J. A., Carroll, R. J., and Freedman, Laurence, L. S. (2009). Modeling Data with Excess Zeros and Measurement Error: Application to Evaluating Relationships Between Episodically Consumed Foods and Health Outcomes. *Biometrics*, 65, 1003–1010.

Küchenhoff, H. and Carroll, R. J. (1997). Segmented Regression with Errors in Predictors: Semi-parametric and Parametric Methods. *Statistics in Medicine*, 16, 169–188.

Laplace, P. S. (1820). *Théorie Analytique des Probabilités*. Courcier.

Mallick, B., Hoffman, F. O., and Carroll, R. J. (2002). Semiparametric Regression Modeling with Mixtures of Berkson and Classical Error, with Application to Fallout from the Nevada Test Site. *Biometrics*, 58, 13–20.

Mallick, B. K. and Gelfand, A. E. (1996). Semiparametric Errors-in-variables Models A Bayesian Approach. *Journal of Statistical Planning and Inference*, 52, 307–321.

Mocanu, O. D. and Oliver, J. (1999). Fault-Tolerant Memory Architecture Against Radiation-Dependent Errors: A Mixed Error Control Approach. *Journal of Electronic Testing*, 14, 169–180.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996). A Semiparametric Transformation Approach to Estimating Usual Daily Intake Distributions. *Journal of the American Statistical Association*, 91, 1440–1449.

Nusser, S. M., Fuller, W. A., Guenther, P. M., et al. (1995). Estimating Usual Dietary Intake Distributions: Adjusting for Measurement Error and Nonnormality in 24-hour Food Intake Data. Technical report, Center for Agricultural and Rural Development (CARD) at Iowa State University.

Reeves, G. K., Cox, D. R., Darby, S. C., and Whitley, E. (1998). Some Aspects of Measurement Error in Explanatory Variables for Continuous and Binary Regression Models. *Statistics in medicine*, 17, 2157–2177.

Rosner, B., Willett, W. C., and Spiegelman, D. (1989). Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Systematic Within-person Measurement Error. *Statistics in Medicine*, 8, 1051–1069.

Salvini, S., Hunter, D. J., Sampson, L., Stampfer, M. J., Colditz, G. A., Rosner, B., and Willett, W. C. (1989). Food-based validation of a dietary questionnaire: the effects of week-to-week variation in food consumption. *International journal of epidemiology*, 18, 858–867.

Sarkar, A., Mallick, B. K., Staudenmayer, J., Pati, D., and Carroll, R. J. (2014). Bayesian Semiparametric Density Deconvolution in the Presence of Conditionally Heteroscedastic Measurement Errors. *Journal of Computational and Graphical Statistics*, 23, 1101–1125.

Schennach, S. M. and Hu, Y. (2013). Nonparametric Identification and Semiparametric Estimation of Classical Measurement Error Models without Side Information. *Journal of the American Statistical Association*, 108, 177–186.

Shen, X. (1997). On Methods of Sieves and Penalization. *The Annals of Statistics*, 25, 2555–2591.

Sinha, S., Mallick, B. K., Kipnis, V., and Carroll, R. J. (2010). Semiparametric Bayesian Analysis of Nutritional Epidemiology Data in the Presence of Measurement Error. *Biometrics*, 66, 444–454.

Stefanski, L. A. and Carroll, R. J. (1990). Deconvolving Kernel Density Estimators. *Statistics*, 21, 169–184.

Stefanski, L. A. and Cook, J. R. (1995). Simulation-extrapolation: The Measurement Error Jackknife. *Journal of the American Statistical Association*, 90, 1247–1256.

Subar, A. F., Dodd, K. W., Guenther, P. M., Kipnis, V., Midthune, D., McDowell, M., Tooze, J. A., Freedman, L. S., and Krebs-Smith, S. M. (2006). The Food Propensity Questionnaire: Concept, Development, and Validation for Use as A Covariate in A Model to Estimate Usual Food Intake. *Journal of the American Dietetic Association*, 106, 1556–1563.

Subar, A. F., Thompson, F. E., Kipnis, V., Midthune, D., Hurwitz, P., McNutt, S., McIntosh, A., and Rosenfeld, S. (2001). Comparative Validation of the Block, Willett, and National Cancer Institute Food Frequency Questionnaires: the Eating at America's Table Study. *American journal of epidemiology*, 154, 1089–1099.

Subar, A. F., Thompson, F. E., Kipnis, V., Mithune, D., Hurwitz, P., McNutt, S., McIntosh, A., and Rosenfeld, S. (2001). Comparative Validation of the Block, Willett, and National Cancer Institute Food Frequency Questionnaires: The Eating at America's Table Study. *American Journal of Epidemiology*, 154, 1089–1099.

Tooze, J. A., Grunwald, G. K., and Jones, R. H. (2002). Analysis of Repeated Measures Data with Clumping at Zero. *Statistical methods in medical research*, 11, 341–355.

Tooze, J. A., Midthune, D., Dodd, K. W., Freedman, L. S., Krebs-Smith, S. M., Subar, A. F., Guenther, P. M., Carroll, R. J., and Kipnis, V. (2006). A New Statistical Method for Estimating the Usual Intake of Episodically Consumed Foods with Application To Their Distribution. *Journal of the American Dietetic Association*, 106, 1575–1587.

Wand, M. P. (1998). Finite Sample Performance of Deconvolving Density Estimators. *Statistics & Probability Letters*, 37, 131–139.

Wang, N., Lin, X., Gutierrez, R. G., and Carroll, R. J. (1998). Bias Analysis and SIMEX Approach in Generalized Linear Mixed Measurement Error Models. *Journal of the American Statistical Association*, 93, 249–261.

Zhang, S., Midthune, D., Guenther, P. M., Krebs-Smith, S. M., Kipnis, V., Dodd, K. W., Buckman, D. W., Tooze, J. A., Freedman, L., and Carroll, R. J. (2011). A New Multivariate Measurement Error Model with Zero-Inflated Dietary Data, and Its Application to Dietary Assessment. *Annals of Applied Statistics*, 5, 1456–1478.