Deconvolution Estimation of a Mixture Distribution with Boundary Effects Motivated by Mutation Effect Distribution

by
Mihee Lee

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2009

Approved by:

J. S. Marron, Advisor

Haipeng Shen, Advisor

Edward Carlstein, Committee Member

Christina L. Burch, Committee Member

Jon W. Tolle, Committee Member

Young K. Truong, Committee Member
ABSTRACT

MIHEE LEE: Deconvolution Estimation of a Mixture Distribution with Boundary Effects Motivated by Mutation Effect Distribution.
(Under the direction of J. S. Marron and Haipeng Shen.)

Density estimation in measurement error models has been widely studied. However, most existing methods consider only continuous target variables, hence they cannot be applied directly to many real problems. Motivated by an evolutionary biology study, we consider more general cases: the target distribution is a mixture of a continuous component and finite numbers of pointmasses, which can cover most of practical problems. In this dissertation, we approach the estimation of the distribution in three different ways under the framework of measurement error models.

Our first proposal is of the Fourier type, which is obtained by generalizing Liu and Taylor (1989). The proposed estimator has a closed form, and gives continuous and smooth density estimators for the continuous mixture component. In addition, its convergence rate is comparably fast. However, when the target distribution has non-smooth boundaries, it suffers from a strong boundary effect. This motivates us to propose two other methods of the sieve type; one is based on maximum likelihood (ML), and the other uses least squares (LS). By easily reflecting the known boundary information, they remarkably reduce the boundary problems, which is another major contribution of this dissertation. Moreover, the use of penalization improves the smoothness of the resulting estimator, especially the ML based estimator, and reduces the estimation variance.

For each estimator, some asymptotic properties are explored by mathematical computation, and finite sample performances are illustrated via simulation studies. In addition, the proposed estimators are applied to the virus lineage data in Burch et al. (2007), which originally motivates this study. In this application, we not only estimate the mutation effect distribution, but also visually validate the classical exponential assumption on the mutation
effect distribution, using density envelope plots.

**Keywords**: Boundary effect, Deconvolution, Fourier transformation, Mixture distribution, Measurement error, Penalization.
ACKNOWLEDGEMENTS

This is a great opportunity to express my respect and gratitude to my advisors, Professor J. S. Marron and Professor Haipeng Shen. I have been exceedingly fortunate to have advisors who gave me the freedom to explore my own ideas, and offer invaluable guidance when my steps faltered. They introduced me to new ways to express and cultivate my ideas, and to deeply examine and question thoughts. Their patience and support helped me overcome many trials leading to the completion of this dissertation. I hope one day to become as good of an advisor to my students as they have been to me.

I would like to thank my other committee members, Christina Burch, Edward Carlstein, Jon Tolle and Young Truong for their valuable suggestions on this dissertation. Especially, I thank Christina for kindly providing the data set that led to an interesting application, and Jon for helping me with the implementation of the proposed method. Furthermore, I would like to express my gratitude to Peter Hall, who has been involved from the very beginning of this work.

I also wish to acknowledge all the friends I met while in Chapel Hill. These friendships helped me to weather the long hours spent in my office. Also, I want to extend appreciation to my friends in Korea, who make me feel that they are always close even though we were physically far apart. Finally, the acknowledgements would not be complete without a heartfelt thanks to my family, especially my parents, for their love and belief in me. I cannot thank them enough.
## Contents

ACKNOWLEDGEMENTS v

List of Figures ix

List of Tables x

1 Introduction 1

2 Direct Deconvolution Estimation 5

2.1 Introduction ........................................................................ 5

2.2 The Proposed Estimators ..................................................... 6

2.2.1 Estimation of the Pointmass $p$ ......................................... 7

2.2.2 Density Estimation of the Continuous Component $f_c$ ........... 8

2.3 Asymptotic properties of the Proposed Estimators .................. 9

2.4 Simulation Studies .............................................................. 14

2.4.1 Case 1: Mixture of $N(3, 1)$ and the Pointmass ................. 15

2.4.2 Case 2: Mixture of $N(0, 1)$ and the Pointmass .................. 18

2.4.3 Case 3: Mixture of Exp(1) and the Pointmass .................... 21

2.5 Application to the Virus-lineage Data ................................. 23

2.5.1 Description on the data .................................................. 23

2.5.2 Analysis Results .......................................................... 24

2.6 Theoretical Proofs ............................................................. 26

vi
3 Sieve Type Deconvolution Based on Maximum Likelihood

3.1 Introduction .................................................. 36
3.2 Model and Methodology ....................................... 37
  3.2.1 Model ....................................................... 38
  3.2.2 Basic Description of the Estimation Method .......... 39
  3.2.3 A Standard Sieve Estimator .......................... 40
  3.2.4 A Penalized Sieve Estimator ......................... 42
3.3 Consistency of the Proposed Estimators ..................... 43
3.4 A Simulation Study .......................................... 46
  3.4.1 Simulation Description .................................. 46
  3.4.2 Pointmass Estimation .................................. 47
  3.4.3 Continuous Distribution Estimation .................. 48
3.5 Application to the Virus-lineage Data ....................... 50
  3.5.1 Estimation Results ..................................... 50
  3.5.2 Validation of the Exponential Assumption .......... 51
3.6 Proofs and Technical Details ................................ 53
  3.6.1 Proof of the Theorems ................................ 54
  3.6.2 Optimization details in the Sieve ML Estimation .. 60

4 Sieve Type Deconvolution Based on Least Squares

4.1 Introduction .................................................. 63
4.2 The Proposed Estimators ..................................... 64
  4.2.1 LS on Cumulative Distribution Functions ............ 66
  4.2.2 LS on Characteristic Functions ....................... 67
4.3 Theoretical Properties ..................................... 68
4.4 Simulation Studies .......................................... 69
  4.4.1 Case 1: Correctly Specified Error Distribution ..... 70
  4.4.2 Case 2: Misspecified Error Distribution ............. 73
4.5 Application to Virus-Lineage Data .......................... 75
List of Figures

2.1 Simulation 1: Direct deconvolution pointmass estimation .................. 15
2.2 Simulation 1: Direct deconvolution density estimation with the true $p$ . . . . . 17
2.3 Simulation 1: Direct deconvolution density estimation with $\hat{p}$ ................ 18
2.4 Simulation 2: Direct deconvolution pointmass estimation ...................... 19
2.5 Simulation 2: Direct deconvolution density estimation with the true $p$ ....... 20
2.6 Simulation 2: Direct deconvolution IMSE plot ....................................... 21
2.7 Simulation 3: Direct deconvolution pointmass estimation ...................... 22
2.8 Simulation 3: Direct deconvolution density estimation with the true $p$ ....... 22
2.9 Direct deconvolution pointmass estimation in the virus data analysis .......... 25
2.10 Direct deconvolution density estimation in the virus data analysis ............ 26
3.1 Simulation: Sieve ML pointmass estimation ......................................... 48
3.2 Simulation: Sieve ML density estimation .............................................. 49
3.3 Virus lineage data - Comparison of three estimation methods ................. 51
3.4 Density-envelope plots - validation of the exponential assumption ............ 52
3.5 Bias and variance tradeoff for the PS-ML estimator ............................. 54
4.1 Simulation 1: Performance of the LS-cdf with uniform weight function .... 70
4.2 Simulation 1: LS-cdf with non-uniform weight function .......................... 72
4.3 Simulation 1: Performances of the LS-chf method ................................ 72
4.4 Simulation 1: Effects of the parameter $M$ in LS-chf ............................ 73
4.5 Simulation 2: Performances of LS-chf under misspecified error assumption .. 74
4.6 Virus-lineage data: Performances of the proposed sieve LS methods ......... 76
List of Tables

4.1 Numerical comparison: LS-cdf, LS-chf and ML ......................... 75
Chapter 1

Introduction

Estimating the probability distribution of a random variable is one of the classical problems in statistics. When some realizations of the random variable are given, many methods to estimate its distribution have been developed from the classical statistical viewpoint. First, if the distribution of the random variable comes from a known parametric family, or there are good reasons to assume some specific distribution family as the truth, an assumed probability distribution can be fit to the data by estimating only the unknown parameters. This process is called parametric estimation. See, for example, Casella and Berger (2001) and Lehmann and Casella (1998). When a parametric model is inappropriate, several methods have been developed to estimate the distribution (Hastie et al., 2001). There is a large body of work in this direction called nonparametric density estimation. Common methods in that area include histogram, kernel smoothing, splines, etc.

When the information about the underlying distribution is correct, parametric estimation is better than the nonparametric methods in two aspects. First of all, it is easier and needs less amount of computation than the latter. Moreover, the convergence rate of the estimator is faster in the case of parametric estimation. The drawback of parametric estimation is it highly depends on the true distribution. That is, when the assumed true distribution is not correct, its performance can be poor. On the other hand, nonparametric density estimation does not need any assumption for the true distribution of the data except continuity or smoothness. So, when there is little information about the underlying distribution, the nonparametric approach is recommended.
The estimation of the distribution is particularly challenging when the true realizations of the target variable are not observed. In this case, the classical distribution estimations described above cannot be directly applied. Let $X$ be the random variable whose distribution we want to estimate. Because of some reasons, we can observe only the error contaminated variable $Y$ where

$$ Y = X + Z, \quad (1.1) $$

where $Z$ is a measurement error with known probability density function $f_Z$, and is independent of the target variable $X$. To fit the underlying distribution based on the error contaminated data, one naive method is to ignore the existence of the measurement error. When the measurement error is much smaller than the target variable, the performance of this approach is not too bad. However, when the measurement error is comparatively large, so not negligible, this method can give an arbitrarily poor conclusion.

A better way is to take into account the existence of the measurement error and reduce its effect from the observed data by using the known distribution information of $Z$, which is called deconvolution method. Such a problem has been widely studied when $X$ is a continuous random variable having a continuous probability density function. There are large volume of studies for the deconvolution problem.


The second group is for non-Fourier based methods. Many of such methods use basis functions such as B-splines or wavelets to expand the target density (or distribution function), before estimating the coefficients of the basis using various approaches. For example, see Mendelsohn and Rice (1982), Koo and Park (1996), Cordy and Thomas (1997), Pensky and Vidakovic (1999), Carrasco and Fleorens (2002), Johnstone et al. (2004), Hall and Qiu (2005)
and Staudenmayer et al. (2008). In addition, other alternatives have been proposed such as NPMLE (Groeneboom and Wellner, 1992; van Es and van Zuijlen, 1996; Jongbloed, 1998), SIMEX (Stefanski and Bay, 1996; Wagner and Stadtmüller, 2008), and TAYLEX (Carroll and Hall, 2004; Wagner and Stadtmüller, 2008).

In this dissertation, motivated by an evolutionary biology study, we are particularly interested in $X$ which has a mixture distribution with both discrete and continuous components. That is, $X$ can be represented as

$$X = \begin{cases} a_j, & \text{with probability } p_j, \text{ for } 1 \leq j \leq \nu, \\ X_c, & \text{with probability } p_{\nu+1}, \end{cases}$$

(1.2)

where the values of $a_1, \ldots, a_\nu$ are known constants, and $X_c$ is a continuous random variable with the density function $f_c$. Then, estimating the distribution of $X$ is equivalent to estimating both $p = (p_1, \ldots, p_{\nu+1})^T$ and the density $f_c$. Here, each $p_j$ is the unknown mixing probability, hence the estimation of $p$ can be understood as the estimation of the mixture proportion. However, our problem is quite different from the challenging classical mixture distribution estimation problems (McLachlan and Basford, 1988; McLachlan and Peel, 2000) because $\nu$-mixture components are exactly known. Since most of the above literatures, especially Fourier-type deconvolution methods, only consider the case where the target variable $X$ has a continuous distribution, they can not appropriately address the mixture structure.

Another important contribution of this dissertation is efficient handling of boundary effects. Suppose that $f_c$ is supported in a finite interval, and has a jump discontinuity at boundaries. One of the most common assumption of Fourier-type methods is that the target density is continuous (over the whole real line), they usually show serious boundary problems even when $X$ is purely continuous, like an exponential variable.

In the following three chapters, we approach the estimation of $(p, f_c)$ in three different ways: the direct deconvolution estimation in Chapter 2, the sieve type estimation based on maximum likelihood in Chapter 3, and another sieve estimation using least squares in Chapter 4. The direct deconvolution estimator is of the Fourier-type, which is obtained by generalizing Liu and Taylor (1989). It enjoys nice properties of other Fourier deconvolution
type estimators, for example, closed forms of the estimators and comparably fast convergence rates. However, it still suffers from the boundary problem. On the other hand, the estimators proposed in Chapters 3 and 4 are based on the sieve methods (Grenander, 1981) which can easily reflect the known boundary information of the target distributions. As a result, they remarkably reduce the boundary problems. For the sieve methods, we focused on the case where $f_c$ is supported on a finite interval. However they can be easily extended to general $f_c$ with little loss of estimation precision, from the tightness property of any single probability distribution.

In each chapter, we investigate the asymptotic properties of each estimator, and its finite sample properties are studied via simulation studies. In addition, we apply the proposed estimators to virus lineage data Burch et al. (2007), and compare the results.
Chapter 2

Direct Deconvolution Estimation

2.1 Introduction

Mutations provide the raw material for evolution, so it is of fundamental importance to study the distribution of the mutation effects in order to understand evolutionary dynamics Elena et al. (1998). However, there is a limited literature on the estimation of the distribution so far. In cases where measurements of individual mutation effects have been obtained, the most common method is to fit exponential (or gamma) distributions to the difference of fitness between unmutated and mutated individuals (Elena et al., 1998; Burch et al., 2007; Sanjuan et al., 2004). This parametric approach is simple and easy, but it ignores the existence of measurement errors that are not usually negligible. As a result, it fails to detect small effects (Burch et al., 2007). Moreover, no serious work has been done to validate the parametric fit.

Instead of this parametric method, we approach the same problem via a nonparametric deconvolution idea, especially based on the Fourier type method. Fourier deconvolution methods have been widely studied, but most of the existing methods consider only the case where the target variable has a continuous density function. In our motivating evolutionary study in Section 2.5, two types of mutations exist: silent mutations that have no effect on the fitness, and deleterious mutations that reduce the fitness. Both the frequency of deleterious mutations and the size of the deleterious mutation effect are of biological interest. Hence, we propose to model the underlying mutation effect distribution as a mixture of a pointmass at 0, which corresponds to the silent mutation effect or no mutation, and a continuous distribution
for the deleterious mutation effect that is supported only on the positive real line. In this case, existing methods from the deconvolution literature cannot be directly applied.

In this chapter, we focus on the case that the distribution of the target variable $X$ is a mixture of a pointmass and a continuous distribution, i.e. $\nu = 1$ in (1.2). For notational convenience, we will use the symbols $p$ and $a$, instead of $p_1$ and $a_1$, from now on. Then, the generalized density ( Cuevas and Walter, 1992) of $X$, say $f_X$, can be expressed as

$$f_X(x) = p\delta_a(x) + (1 - p)f_c(x),$$

(2.1)

where $\delta_a$ denotes the Dirac delta at $a$, and $f_c$ is the density of $X_c$.

In Section 2.2, we propose the estimators for both the pointmass $p$ and the continuous density $f_c$ on top of measurement error models, by extending the idea of the classical Fourier deconvolution estimation. Their asymptotic properties are also provided in Section 2.3, with the technical proofs given in Section 2.6. Section 2.4 presents several simulation results to illustrate the performance of the proposed estimators. In Section 2.5, the estimators are applied to the virus-lineage data of Burch et al. (2007). In both simulations and real data analysis, we only consider the case $a = 0$.

### 2.2 The Proposed Estimators

In this section, we propose the direct deconvolution estimators of $p$ and $f_c$ of (2.1). The estimators are derived below in Sections 2.2.1 and 2.2.2 respectively, along with theorems about their asymptotic properties. Detailed proofs are provided in Section 2.6.

Deconvolution estimation of mixture densities is a natural approach, and our proposal directly extends the method of Liu and Taylor (1989) to cases of mixtures of discrete and continuous components. Let $X$ be the variable with the mixture structure in (1.2) with $\nu = 1$, and let $Y$ denote the corresponding variable contaminated by the measurement error $Z$, defined as (1.1). Our procedure starts with estimating the density of $Y$, say $f_Y$, based on the observations $\{Y_i : i = 1, \ldots, n\}$. Afterwards, the generalized density of $X$ can be obtained by directly deconvoluting $f_Z$ from $f_Y$, due to the independence assumption of $X$.
and $Z$.

The proposed estimators are attractive in the sense that they take into account the measurement errors, and have closed form expressions that are easy to implement. Our experience suggests that the estimator for $f_c$ performs well except near non-smooth boundaries. This is a common problem that is shared by the existing deconvolution estimators. For example, in our motivating application, the support is known to be positive. In this case, our density estimator has some problem near the origin, but works well in the rest of the support. The use of boundary information in deconvolution problem has been studied by Pensky (2002), Hall and Qiu (2005) and Meister (2007). However they only consider the case where $X$ has a continuous distribution, and it is not clear how to extend their methods to general models.

### 2.2.1 Estimation of the Pointmass $p$

We consider the pointmass estimation first. The basic idea comes from the Inverse-Fourier transformation (Billingsley, 1995). Since $p$ is the probability that $X$ takes the value $a$, it can be obtained as

$$p = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} \exp(-ita) \varphi_X(t) \, dt,$$

where $\varphi_X$ is the characteristic function of $X$.

From (2.2), the pointmass $p$ can be estimated by replacing $\varphi_X$ with its estimator $\hat{\varphi}_X$. Hence we need to estimate the characteristic function of $X$. For that, we make use of the relation $Y = X + Z$, and the independence between $X$ and $Z$. It follows that $\varphi_X = \varphi_Y / \varphi_Z$, where $\varphi_Z$ is the known characteristic function of $Z$, and $\varphi_Y$ is the characteristic function of $Y$ that can be estimated by the empirical characteristic function of $Y$ based on the observations, i.e.

$$\hat{\varphi}_Y(t) = \frac{1}{n} \sum_{j=1}^{n} \exp(itY_j).$$
As a result, a naive estimator of $p$ is proposed as

\[
\tilde{p} = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} \exp(-ita) \hat{\varphi}_X(t) \, dt,
\]

\[
= \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} \exp(-ita) \cdot \frac{\hat{\varphi}_Y(t)}{\varphi_Z(t)} \, dt,
\]

\[
= \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} \frac{1}{n} \sum_{j=1}^{n} \frac{\exp(it(Y_j - a))}{\varphi_Z(t)} \, dt.
\]  

(2.3)

One thing to be noted is that $p$ is a probability, and hence a real number. However the integrand of (2.3) contains a complex term, so it is not guaranteed that $\tilde{p}$ is a real number. Therefore we take only the real part of $\tilde{p}$ as the estimator. Another problem is the computational challenge caused by the limiting operation. To ease the difficulty, we replace $T$ by $T_n$, a sequence of positive real numbers which goes to infinity as $n$ goes to infinity. Hence we can get the final estimator $\hat{p}$ of the pointmass as

\[
\hat{p} = \frac{1}{2nT_n} \sum_{j=1}^{n} \text{Re} \int_{-T_n}^{T_n} \frac{\exp(it(Y_j - a))}{\varphi_Z(t)} \, dt,
\]  

(2.4)

where $\text{Re}$ denotes the real part of the complex integral.

2.2.2 \textbf{Density Estimation of the Continuous Component} \(f_c\)

To estimate $f_c$, we also use the Inverse-Fourier transformation. In particular, when $\varphi_X$ is an integrable function, it is known (Billingsley, 1995) that the random variable $X$ has a density function $f_X$ of the form

\[
f_X(x) = \lim_{M \to \infty} \frac{1}{2\pi} \int_{-M}^{M} \exp(-itx)\varphi_X(t) \, dt.
\]

In our problem, $X_c$ is assumed to have a continuous density $f_c$, so its characteristic function $\varphi_c$ is integrable. In addition, the mixture structure of $X$ suggests that $\varphi_c(t)$ can be expressed as

\[
\varphi_c(t) = \frac{\varphi_X(t) - p \cdot \exp(ita)}{1 - p},
\]  

(2.5)
where $\varphi_X(t)$ can be estimated in the same manner as discussed above in Section 2.2.1. Then, $f_c$ can be estimated as

$$\hat{f}_c(x) = \lim_{M \to \infty} \frac{1}{2\pi} \int_{-M}^{M} \hat{\varphi}_c(t) \exp(-itx) dt$$

$$= \lim_{M \to \infty} \frac{1}{2\pi} \int_{-M}^{M} \left[ \hat{\varphi}(t) - p \exp(it\alpha) \right] \exp(-itx) dt$$

$$= \lim_{M \to \infty} \frac{1}{2\pi(1-p)} \int_{-M}^{M} \left[ \frac{1}{n} \sum_{j=1}^{n} \exp(it(Y_j-x)) \right] dt.$$

As in the pointmass estimation, $\hat{f}_c$ is not guaranteed to be a real-valued function. Moreover, the computation of $\hat{f}_c$ also involves the limit operation. Therefore, we take only the real part of the above integration, and replace $M$ by $M_n$, a sequence of positive numbers converging to infinity. In addition, since $p$ is usually unknown, we plug in $\hat{p}$ to replace $p$. Hence the final form of the estimator $\hat{f}_c$ is given as

$$\hat{f}_c(x) = \frac{1}{2\pi n(1-\hat{p})} \sum_{j=1}^{n} \text{Re} \int_{-M_n}^{M_n} \left[ \frac{\exp(it(Y_j-x))}{\varphi_Z(t)} - \hat{p} \exp(it(a-x)) \right] dt. \quad (2.6)$$

If the true probability $p$ is known, then $\hat{f}_c$ can be obtained using that value, which improves the estimation performance.

### 2.3 Asymptotic properties of the Proposed Estimators

The estimator $\hat{p}$ can be shown to be consistent as stated in Theorem 2.3.3. Below we first derive the mean and the variance of the estimator in Lemmas 2.3.1 and 2.3.2. All the proofs are given in Section 2.6.

**Lemma 2.3.1.** Let $\hat{p}$ be the estimator of $p$ as defined in (2.4), and assume that $\varphi_Z(t)$ does not equal to 0 for any $t \in [-T_n, T_n]$. Then the expectation of the estimator is given by

$$E(\hat{p}) = p + \frac{1-p}{2T_n} \text{Re} \int_{-T_n}^{T_n} \varphi_c(t) \exp(-ita) dt,$$
where \( \phi_c \) is the characteristic function of \( X_c \), the continuous component of \( X \).

Remark 1. Note that \( T_n \) goes to infinity as \( n \to \infty \), and \( X_c \) is a continuous random variable with \( P(X_c = a) = 0 \). Hence the expectation of \( \hat{p} \) converges to \( p \) as \( n \) goes to infinity, which suggests that \( \hat{p} \) is asymptotically unbiased.

The following Lemma 2.3.1 derives the variance of \( \hat{p} \). We assume that the distribution of the measurement error \( Z \) is symmetric about 0, which is a common assumption in measurement error models.

**Lemma 2.3.2.** Suppose that the distribution of \( Z \) is symmetric about 0. Then the variance of \( \hat{p} \) is given by

\[
\text{Var}(\hat{p}) = \frac{1}{2nT_n^2} \int_0^{T_n} \int_0^{T_n} \left[ \frac{\text{Re}\{\varphi_V(s + t) + \varphi_V(s - t)\} - 2\text{Re}\{\varphi_V(s)\} \text{Re}\{\varphi_V(t)\}}{\varphi_Z(s)\varphi_Z(t)} \right] \text{d}s\text{d}t,
\]

where \( V = Y - a \), and \( \varphi_V(\cdot) \) is the characteristic function of \( V \).

Remark 2. Note that the variance of the density estimator in Liu and Taylor (1989) is

\[
\text{Var}(\hat{f}_n(x)) = \frac{1}{n\pi^2} \int_0^{T_n} \int_0^{T_n} \left[ \frac{1}{2} \text{Re}\{\varphi_V(s + t) + \varphi_V(s - t)\} - \text{Re}\{\varphi_V(s)\} \text{Re}\{\varphi_V(t)\} \right] \times \frac{\varphi_K(sh_n)\varphi_K(th_n)}{\varphi_Z(s)\varphi_Z(t)} \text{d}s\text{d}t.
\]

The variance of the pointmass estimator \( \hat{p} \) has a very similar structure as that of \( \hat{f}_n(x) \) when \( h_n = 0 \). However \( \text{Var}(\hat{p}) \) converges to 0 much faster than \( \text{Var}(\hat{f}_n(x)) \). In fact,

\[
\frac{\text{Var}(\hat{p})}{\text{Var}(\hat{f}_n(x))} = O(T_n^{-2}), \quad \text{as } n \to \infty.
\]
Based on the above two lemmas, we conclude that \( \hat{p} \) is consistent under some suitable conditions in Theorem 2.3.3.

**Theorem 2.3.3.** Suppose that \( \varphi_Z(t) \) is not equal to 0 for any \( t \), \( f_c(a) \) has a finite value, and the distribution of \( Z \) is symmetric about 0. In addition, suppose that there is a sequence \( T_n \) satisfying

\[
T_n \to \infty, \quad \frac{1}{n^{1/2}T_n} \int_0^{T_n} \frac{1}{\varphi_Z(t)} dt \to 0
\]

as \( n \) goes to infinity. Then \( \hat{p} \) converges to \( p \) in probability as \( n \to \infty \), i.e. \( \hat{p} \) is a consistent estimator of \( p \).

**Remark 3.** Theorem 2.3.3 suggests that the distribution of the measurement error \( Z \) highly affects the choice of \( T_n \), hence the convergence rate of the estimator. For example, when \( Z \) has the standard normal distribution, \( T_n = \log^{1/2} n \) satisfies (2.7). In this case, the variance of the estimator is of the order \( \log^{-3/2} n \), i.e. \( \text{Var}(\hat{p}) = O(\log^{-3/2} n) \) as \( n \to \infty \).

Theorems 2.3.4 and 2.3.5 below provide some asymptotic properties of \( \hat{f}_c(x) \). For any \( x \neq a \), we show in Theorem 2.3.4 that the proposed density estimator is a consistent estimator of \( f_c(x) \) under some suitable conditions. In addition, under stronger conditions, Theorem 2.3.5 establishes the consistency of \( \hat{f}_c(x) \) at \( x = a \). The proofs of the theorem are provided in Section 2.6.

**Theorem 2.3.4.** Suppose that the conditions in Theorem 2.3.3 hold. In addition, suppose that

\[
M_n \to \infty, \quad n^{-1/2} \int_0^{M_n} \frac{1}{\varphi_Z(t)} dt \to 0
\]

as \( n \) goes to infinity. Then \( \hat{f}_c(x) \) converges to \( f_c(x) \) in probability for any \( x \neq a \).
Theorem 2.3.5. Suppose that $\varphi_Z(t)$ is not equal to zero at any $t$, $f_c(a)$ is finite, and the distribution of $Z$ is symmetric about 0. In addition to (2.8), suppose that
\[ M_n = o(T_n), \quad n^{-1/2} \int_0^{T_n} \frac{1}{\varphi_Z(t)} dt = O(1), \tag{2.9} \]
as $n \to \infty$. Then $\hat{f}_c(x)$ is a consistent estimator of $f_c(x)$ at $x = a$.

Remark 4. When comparing Theorem 2.3.5 with Theorem 2.3.4, the consistency of $\hat{f}_c(x)$ at $x = a$ requires stronger conditions, which guarantee $M_n(\hat{p} - p) \to 0$ in probability. This is stronger than $\hat{p} - p$ converges to 0, which is required in Theorem 2.3.4.

After obtaining $\hat{p}$ and $\hat{f}_c(x)$, the generalized density estimator of (2.1) is easily obtained as
\[ \hat{f}_X(x) = \hat{p}\delta_a(x) + (1 - \hat{p}) \hat{f}_c(x). \tag{2.10} \]
Under the conditions in Theorem 2.3.4, the consistency of $\hat{f}_X(x)$ is easily shown by the consistency of $\hat{p}$ and $\hat{f}_c(x)$, and Lévy’s continuity theorem.

Corollary 2.3.6. Suppose that the conditions in Theorem 2.3.4 hold. Then, for any $x \neq a$, $\hat{f}_X(x)$ in (2.10) is a consistent estimator of $f_X(x)$.

In addition to the consistency, we obtain the actual convergence rate of $\hat{f}_X$ in terms of the mean squared error (MSE). There are two factors which affect the convergence rate: the smoothness of the error distribution, and the smoothness of $f_c$. We use the order of the characteristic functions $\varphi_Z(t)$ and $\varphi_c(t)$ as $t \to \infty$ in order to describe the smoothness of the corresponding distributions.

In Lemma 2.3.7, we obtain the order of $\text{Bias}(\hat{f}_X(x))$, which is determined by the tail property of $\varphi_c$. Here, we consider two types of $\varphi_c$:

\begin{itemize}
    \item[(B1)] $|\varphi_c(t)||t|^{\beta_1} \leq d_1$, \quad for some $\beta_1 > 1$ and $d_1 > 0$ as $t \to \infty$;
\end{itemize}
(B2) $|\varphi(t)| \exp(|t|^{\beta_1}/\gamma_1) \leq d_1$, for some $\beta_1 \geq 1$ and $d_1 > 0$ as $t \to \infty$;

**Lemma 2.3.7.** Suppose that $\varphi_Z(t)$ is not equal to zero for any $t$. Then, for any $x \neq a$,

$$\text{Bias}(\hat{f}_X(x)) = \begin{cases} O(T_n^{-1} + M_n^{-\beta_1+1}), & \text{under } (B1); \\ O\left(T_n^{-1} + M_n^{-\beta_1+1} \exp\left(\frac{-M_n^{\beta_1}}{\gamma_1}\right)\right), & \text{under } (B2). \end{cases}$$

The following Lemma 2.3.8 shows the order of the variance of $\hat{f}_X(x)$, which depends on the tail property of $\varphi_Z(t)$. Again, we consider three types of error distributions:

(V1) $|\varphi_Z(t)| |t|^\beta_2 \geq d_2$, $t \to \infty$, for some $\beta_2 > 1$ and $d_2 > 0$;

(V2) $|\varphi_Z(t)| \exp(|t|^\beta_2/\gamma_2) \geq d_2$, $t \to \infty$, for some $\beta_2 > 0$, $\gamma_2 > 0$ and $d_2 > 0$;

In (V1), the constraint $\beta_2 > 1$ comes from the fact that $f_Z$ is a continuous density, so that its characteristic function $|\varphi_Z|$ is integrable.

**Lemma 2.3.8.** Suppose that $\varphi_Z(t)$ is symmetric about 0. Then,

$$\text{Var}(\hat{f}_X(x)) = \begin{cases} O\left(\frac{T_n^{2\beta_2-1}}{n} + \frac{M_n^{2\beta_2+1}}{n}\right), & \text{under } (V1); \\ O\left(\frac{1}{nT_n^{\beta_2}} \exp\left(\frac{2T_n^{\beta_2}}{\gamma_2}\right) + \frac{1}{nM_n^{\beta_2-2}} \exp\left(\frac{2M_n^{\beta_2}}{\gamma_2}\right)\right), & \text{under } (V2), \end{cases}$$

where $\beta^* = 1$ if $\beta_2 < 1$, and $\beta^* = 2\beta_2$ for $\beta_2 \geq 1$.

From the above Lemmas 2.3.7 and 2.3.8, we can get the convergence rate of $\hat{f}_X(x)$. Theorem 2.3.9 below is for the case where (B1) and (V2) are satisfied. Note that the normal distribution, which is the most common model for measurement errors, satisfies the condition (V2) with $\beta_2 = 2$. Exponential or gamma distributions can be examples of (B1). The results for other combinations can be obtained by the proof procedures similar to Theorem 2.3.9.
Theorem 2.3.9. Suppose that $\varphi_c$ and $\varphi_Z$ satisfy (B1) and (V2), respectively. And assume that $\varphi_Z(t) = \varphi_Z(-t) \neq 0$ for any $t$. Then, for any $x \neq a$, by choosing

$$M_n = (\gamma/4)^{1/\beta_2}(\log n)^{1/\beta_2} \quad \text{and} \quad T_n = (\gamma/4)^{1/\beta_2}(\log n)^{\alpha/\beta_2},$$

we can get the follows: when $\alpha = \min(\beta_1 - 1, 1)$,

$$E\left(\hat{f}_X(x) - f_X(x)\right)^2 = O\left((\log n)^{-2\alpha/\beta_2}\right), \quad \text{as} \quad n \to \infty.$$

Note that Fan (1991b) shows that when $X$ is a continuous variable with the density function $f_X$, and $Z$ is a super smooth error corresponding to (V2), the convergence optimal convergence rate of $\hat{f}_X$ has an order $O((\log n)^{-2\alpha^*/\beta_2})$, where $0 \leq \alpha^* < 1$. Our result established in Theorem 2.3.9 is very similar to this optimal convergence rate, even the assumptions on the target distribution are different.

2.4 Simulation Studies

In this section, we perform three simulation studies to investigate the performance and properties of the estimators proposed in Section 2.2. All subsections have similar simulation schemes: the pointmass $p = 0.5$ at 0, the sample size $n = 300$, the distribution of the measurement error, etc. The only change is the distribution of the continuous component, which is $N(3, 1)$, $N(0, 1)$ and Exp(1), respectively. These simulation setups cover a wide range of scenarios, including overlapping mixture components and non-smooth boundaries. Details are explained in each subsection.

An important issue in the deconvolution estimation is the choice of the integration range parameters, $T_n$ for estimating $p$, and $M_n$ for estimating $f_c$. Instead of selecting one pair of such parameters, we adopt the scale space approach suggested by Chaudhuri and Marron (2000). The idea is that we will try a range of parameters, and see the change of the estimators as the parameters change.
2.4.1 Case 1: Mixture of $N(3, 1)$ and the Pointmass

We start with a variable $X$ whose distribution is the mixture of a normal distribution with mean 3 and standard deviation 1, and the pointmass at 0, with the mixing probability being 0.5, i.e.

$$X \sim \begin{cases} N(3, 1), & \text{with probability 0.5,} \\ 0, & \text{with probability 0.5.} \end{cases}$$

In this case, the two components are not strongly overlapping. Moreover, the continuous part is supported on the whole real line, so there is no boundary problem.

We assume the independent measurement error variable $Z$ has a normal distribution with mean 0 and standard deviation $\sigma = 0.1$. We simulate $L = 100$ random samples with size $n = 300$ from the distribution of $Y = X + Z$, which is the convolution of the target distribution and the distribution of $Z$.

![Figure 2.1](image.png)

Figure 2.1: (Case 1) The left panel shows 100 simulated pointmass estimators (the gray curves), and their average (the black solid curve), as functions of $T_n$ which is the integration range parameter on the horizontal axis. In the right panel, each point is an individual pointmass estimator, and the solid curve is a kernel density estimate of these 100 estimators. The dotted and dashed vertical lines show the true value of the pointmass and the average estimator, respectively.

Figure 2.1 summarizes the performance of the pointmass estimator for 100 simulated data sets. Figure 2.1(a) shows the change in the pointmass estimator as a function of the integration parameter $T_n$ of (2.4), where the vertical axis shows the value of $\hat{p}$. The gray curves show the pointmass estimators from the 100 samples and the black solid curve is the...
average of the 100 estimators. According to Figure 2.1(a), as $T_n$ increases, the estimator $\hat{p}$ first decreases from 1, and increases slightly before stabilizing around the true pointmass 0.5 for $T_n$ larger than 3. Once it stabilizes, the average estimator $\hat{p}$ lies within the interval $[0.4983, 0.5010]$, which suggests a small bias when $T_n$ is large enough. On the other hand, the variance of the estimator increases as $T_n$ increases.

For Figure 2.1(b), we choose a specific value of $\hat{p}$, from each gray curve Figure 2.1(a). Since we found that the pointmass is usually overestimated in several simulation studies, and too large $T_n$ results in instability of the estimation, we choose the first local minimum of each $\hat{p}$ as our estimate, if it lies between 0 and 1. Otherwise, e.g. there is no local minimum, we choose $T_n$ which gives the smallest difference of $\hat{p}$, and the corresponding $\hat{p}$ is used as our estimate. Figure 2.1(b) shows the scatter plot of these 100 estimators. To show the distribution of these estimators, their kernel density estimator (the solid curve) is plotted together. In addition, the dotted and dashed vertical curves show the true value 0.5 and the average of the 100 estimators, respectively. For this selection method, the pointmass estimator tends to have a slightly smaller value than the true value (the average of the 100 estimators is about 0.48). Note that we use the same range for the horizontal axis in the corresponding panels of Figures 2.1, 2.4 and 2.7 to make the comparison clear.

Figure 2.2 plots the density estimator in (2.6) for various values of $M_n$. Here, the true value of $p = 0.5$ is used in estimating the density $f_c$, in order to study the performance of density estimation with no influence from the pointmass estimation. In each panel, the black solid curve is the average of the estimators from the 100 samples, while the gray dash-dot curve is the true density $f_c$. In addition, the gray solid curves are the average estimator $\pm 2$ standard error, which play a role as a confidence band based on the 100 samples. Note that in some panels the curves are completely overlapping.

Similar to the pointmass estimation, a large value of the integration range parameter $M_n$ corresponds to a small estimation bias. However, when $M_n$ is too big, the estimator is very wiggly, and some periodic component dominates the entire structure of the target function. On the other hand, a small $M_n$ gives a small estimation variance, but a large bias due to over-smoothing. When $M_n = 1.5^2$, the estimator is almost the same as the true value of $f_c$. 

Figure 2.2: (Case 1) Estimation of the continuous density (with known $p$). This plot shows the proposed estimator of $f_c$. Each panel corresponds to the estimator based on $M_n = 1.5^0, 1.5^2, 1.5^6$ and $1.5^8$. In each plot, the dash-dot curve is the true density, the black solid curve is the average estimator, and the gray solid curves show the average estimator ± 2 standard error, based on the 100 random samples.

Interestingly, the standard error of $\hat{f}_c(x)$, reflected by the width of the confidence band, near $x = 0$ is much larger than near $x = 6$. Since the normal density curve is symmetric about its mean (3 in this case), one might expect the variations of the estimators $\hat{f}_c(0)$ and $\hat{f}_c(6)$ would be similar, but this is not the case. The pointmass at 0 adds additional noise to the estimation of the density function at 0. This is consistent with Remark 4, which states that the consistency of $\hat{f}_c(x)$ at $x = a$ requires more assumptions than $x \neq a$.

Figure 2.3 shows the density estimators which are computed using the pointmass estimator $\hat{p}$, plotted in Figure 2.1(b). Compared to the density estimators obtained from the true $p$, the curves in Figure 2.3 show larger values near $x = 0$, which is the location of the pointmass. This result can be explained by the underestimation of the pointmass (0.478 on average). Except for the neighborhood of $x = 0$, the performance of the density estimator based on $\hat{p}$
Figure 2.3: (Case 1) Estimation of the continuous density (with estimated $\hat{p}$). This plot shows the proposed estimator of $f_c$, based on the pointmass estimator $\hat{p}$. Each panel corresponds to the estimator based on $M_n = 1.5^0, 1.5^2, 1.5^6$ and $1.5^8$. In each plot, the dash-dot curve is the true density, the black solid curve is the average estimator, and the gray solid curves on the 100 random samples.

is similar to that based on $p$.

2.4.2 Case 2: Mixture of $N(0, 1)$ and the Pointmass

The second simulation considers the mixture of the standard normal distribution and the pointmass at 0 with a mixing probability of 0.5. Different from the first simulation, the location of the pointmass 0 is now the same as the mode of the standard normal distribution, so the two components are highly overlapping. We expect the pointmass $p$ strongly affects the estimation of $f_c$, and $\hat{p}$ is also affected by $f_c(x)$ near $x = 0$, which are confirmed below. We make the same assumption about the measurement error variable $Z$. The sample size and the number of iterations are also the same as the previous simulation, i.e., $n = 300$ and $L = 100$. 
Figure 2.4: (Case 2) This plot shows the bias and variance in the estimation of the pointmass \( p \). In the panel (a), the black solid curve shows the average estimator, and the gray curves are individual estimators. The horizontal axis \( T_n \) is the integration range parameter. The panel (b) shows a kernel density estimator of 100 pointmass estimators. The dotted vertical line shows the true value, and the dashed line shows the average estimator.

As in the previous case, in Figure 2.4(a), each gray curve shows an individual estimator and the black solid curve is the average estimator. The overall trend of \( \hat{p} \) is similar to the previous simulation, but the performance is worse as we expected: a slightly larger bias and a much larger variance. Especially, the estimation variation is much larger when a large \( T_n \) is used. This can be explained by the overlapping of the two mixture components. To estimate the pointmass, we use the same criterion in choosing \( T_n \) as discussed in Case 1. As shown in Figure 2.4(b), we can see the pointmass is overestimated (the average of the 100 estimators is around 0.52).

Similar to Figure 2.2, Figure 2.5 shows the result of the density estimation, which is based on the true \( p = 0.5 \). One big difference from the previous simulation is the trend of the standard error. In the current simulation, both the estimator \( \hat{f}_c \) and the standard error are almost symmetric about 0. In addition, the standard error has the biggest value at 0. This is because the location of the pointmass is the center point of the continuous component. So its effect on the estimation of \( f_c(x) \) is the largest when \( x = 0 \), and decreases as \( x \) departs from 0. Like the first case, \( M_n = 1.5^2 \) gives the best fit, almost overlapping the target.

We also estimate \( f_c \) based on the pointmass estimator \( \hat{p} \), instead of \( p \). As we discussed in the previous simulation, overall performance of the density estimator \( \hat{f}_c(x) \) is similar to that
Figure 2.5: (Case 2) This plot shows the direct deconvolution estimator of $f_c$. Each panel corresponds to the estimator based on $M = 1.5^0, 1.5^2, 1.5^6$ and $1.5^8$. In each plot, the gray dash-dot curve is the true density, the black solid curve is the average estimator, and the gray solid curves show the average estimator ± 2 standard error.

In addition, we investigate the performance of the estimation of $f_X(x)$, which is our essential estimand, in terms of the integrated squared bias, variance and mean squared error. Figure 2.6 shows the above three numerical measures in log scale for various values of $M_n$. Each panel of Figure 2.6 contains three curves. The gray dashed curve is for the case where $f_X(x)$ is assumed to be a continuous, i.e. the pointmass component is ignored. The black dashed curve displays the case where the true $p = 0.5$ is known, and hence used in estimating $f_c(x)$. The black solid curve is used for the case where $p$ is estimated, which fully reflects our mixture assumption on $f_X(x)$ with unknown $p$. 

Based on the true $p$, except near $x = 0$. In opposition to Case 1, $\hat{f}_c(x)$ is underestimated on the neighbor of $x = 0$. It might be because that $\hat{p}$ is overestimated (0.523 on average) in this case.
Figure 2.6: (Case 2) In Panels (a)-(c), the integrated squared bias, variance and mean squared error of \( \hat{f}_X(x) \) are plotted versus \( M_n \) in log-scale. In each panel, the gray dashed curve is for the case where \( p = 0 \) is assumed when estimating \( f_c(x) \), the black dash-dot curve corresponds to the case \( p = 0.5 \), and the black solid curve shows the case where \( \hat{p} \) is used in estimating \( f_c(x) \).

As one can expect, the performance (in terms of the mean squared error) is the best when the true pointmass \( p = 0.5 \) is used, and the worst when the pointmass component is ignored. When the estimated \( p \) is used, the estimation result is worse than the case \( p = 0.5 \), but is much better than the case \( p = 0 \). In addition, when \( M_n \) is too large, all three cases perform poorly.

2.4.3 Case 3: Mixture of Exp(1) and the Pointmass

The last simulation considers the mixture of the standard exponential distribution and the pointmass at 0. The rest of the simulation setup is the same as the previous two cases, in terms of the mixing probability, the measurement error distribution, the sample size, and the number of iterations.

The difficulty in estimating the exponential density is that it has a non-smooth left boundary, so the estimation would not be accurate near the left boundary (at 0). Moreover, the location of the pointmass is near the peak of the exponential component. Like the second case, the estimation of both \( p \) and \( f_c \) is highly related, which makes the task harder.

As shown in Figure 2.7, the pointmass is a little overestimated. The same criterion as in Case 1 is used to select \( T_n \). The estimation variance and bias are similar with those of the
Figure 2.7: (Case 3) This plot shows the bias and variance in the estimation of the pointmass $p$. In the left panel, each gray curve is an individual estimator, and the black solid curve shows the average estimator. The panel (b) shows the 100 estimators with its density. Here, the dotted/dashed lines show the true/average estimator, respectively.

Figure 2.8: (Case 3) This plot shows the direct deconvolution estimator of $f_c$. Each panel corresponds to the estimator based on $M = 1.5^0, 1.5^2, 1.5^6$ and $1.5^8$. In each plot, the gray dash-dot curve is the true density, the black solid curve is the average of the estimators, and the gray solid curves show the average estimator ±2 standard error.
second case, but slightly larger. The estimation of \( f_c \) also has similar trend with the other cases, in terms of a bias or a variance. In addition, it gives us very important information on the boundary effect; Since the exponential density is supported only on the positive real line, it is desirable that the estimator has only positive support. However, the support of \( \hat{f}_c \) includes the negative real line, and \( \hat{f}_c \) is underestimated near 0, especially when \( M_n \) is small. For large \( M_n \), as shown in Figures 2.8(c) and (d), the estimator changes sharply near 0, and oscillates on the negative real line. The variation on the negative real line can be considered as noise in these cases, so the boundary problem is weaken. Hence a larger \( M_n \) is preferred if the target density has any bounded support. In this simulation, the density estimator performs best when \( M_n = 1.5^6 \), which is much bigger than the previous two cases.

2.5 Application to the Virus-lineage Data

In this section, we illustrate the performance of the proposed estimators via an application to the virus-lineage data in Burch et al. (2007). In this analysis, our goal is to estimate the distribution of mutation effects on virus fitness.

2.5.1 Description on the data

For this data set, 10 virus lineages were grown in the lab for 40 days, in a manner that promoted the accumulation of mutations in discrete random events. Plaque size was used as a measure of viral fitness and measured everyday for each lineage.

Let \( Y \) be the reduction between two consecutive plaque size measurements. This \( Y \) is the sum of two facts; one is the real mutation effect on plaque sizes, say \( X \), and another is a measurement error \( Z \) which comes from technical difficulties in measuring plaque sizes. When the distribution of \( Z \) is known, the distribution of \( X \) can be estimated by the proposed method. The relationship between two variables \( X \) and \( S \) is measured by Burch and Chao (2004) as

\[
X = 22.73 \log(1 + S). \tag{2.11}
\]

Note that our final goal is to estimate the distribution of mutation effects on fitness, say \( S \),
not of $X$. It can be obtained from the relationship (2.11) and a simple change of variable operation on the estimation results for $X$. Due to the scientific interest in the mutation effects on fitness $S$, all the results are reported in terms of $S$.

In addition, the lineages were founded with a high fitness virus to ensure that during any given time interval, there are only two possibilities in terms of mutations:

(i) No mutation occurs, or only silent mutations occur.

(ii) A deleterious mutation occurs.

The silent mutation is defined as a mutation that has no effect on virus fitness, hence the theoretical mutation effect $X$ of the case (i) is 0. On the other hand, deleterious mutations reduce the plaque sizes, so the deleterious mutation effect on plaque sizes takes only positive values. The probability distribution of the deleterious mutation effects is usually considered as continuous. Hence the distribution of mutation effects can be expressed as the mixture of a point mass at 0, corresponding to case (i), and a continuous distribution (for the deleterious mutations) which is supported only on the positive real line. Unfortunately, we cannot observe the mutation effects without measurement errors, hence it is necessary to consider the measurement-error model on top of the mixture structure.

2.5.2 Analysis Results

We consider the pointmass estimation result first. Figure 2.9 plots the $\hat{p}$ versus $T_n$. In Figure 2.9(a), the pointmass $p$ is estimated for $T_n$ in the range $[0.1, 10]$; since we assume the normality for the measurement error $Z$, the integrand of (2.4) may have very large value near tails. So a large value of $T_n$ results in instability of the estimation and too long computation time, which is the reason we restrict the upper bound of $T_n$ by 10. The estimator $\hat{p}$ changes sharply when $T_n$ is large, which makes it difficult to see the precise change of $\hat{p}$ for small values of $T_n$. So in Figure 2.9(b), the picture is zoomed into the region to the left of the vertical bar, i.e. for $T_n$ between 0.1 and 4. From the simulation studies in Section 2.4, we have observed that $\hat{p}$ is usually overestimated, that is, it tends to be larger than the true parameter $p$. In
addition, when $T_n$ is large, variation of the estimation is very large. Hence we select the first local minimum as the estimator for $p$, which is 0.9363 at $T_n = 2.4$.

![Graph](image_url)

Figure 2.9: The panel (a) plots the estimator of the pointmass $\hat{p}$ versus the range parameter $T_n = (0.1, 0.2, \ldots, 10)$. The panel (b) shows $\hat{p}$ only for $T_n = (0.1, 0.2, \ldots, 4)$ to get a more precise view in the region of interest. The black dotted lines highlight the suggested $T_n$ and $\hat{p}$.

For the estimation of $f_c$, we again use the scale space approach suggested by Chaudhuri and Marron (2000). Figure 2.10 shows the density estimator $\hat{f}_c$ for different values of $M_n$, the integration range parameter. Each panel shows two density curves: for the black dash-dot curve, $\hat{p} = 0.9363$ is used, and for the gray solid curve, we use $\hat{p} = 0.9027$ which is the pointmass estimator given by Burch et al. (2007).

When the smaller pointmass is used, the peak location of the density curve estimator is closer to 0. It is due to the difficulty of separating small deleterious mutation effects from silent mutation effects. If we underestimate $p$, the proportion of silent mutations, some silent mutations are considered as deleterious mutations that have small effects. Except this, the effect of pointmass estimation is small on estimating the density curve. As shown in Figure 2.10, the two curves in each panel look very similar, and the curves change in the same way as $M_n$ changes.

We now discuss the effect of the integration range parameter. When $M_n$ is small (for example, $M_n = 0.5$), the estimator shows the overall trend of the density curve well. According to Figure 2.10(a), the deleterious mutation effects are mostly distributed near 0. In this case,
the density estimator has positive values even on the negative real line, which contradicts the fact that the deleterious mutation effects are always nonnegative. This boundary effect shows up better when $M_n$ is larger. In Figure 2.10(d), the density estimator changes very sharply near 0, and oscillates on the negative real line. The variation of the density curves on the negative part can be considered as noise, and the true underlying density curve is supported only on the positive real line.

### 2.6 Theoretical Proofs

In this section, we provide technical proofs for the lemmas and theorems in Section 2.3. Note that the estimators $\hat{p}$ and $\hat{f}_c$ have similar structures with the density estimator of Liu and Taylor (1989). Hence the proofs of the theorems are similar to their proof.
Proof of Lemma 2.3.1  It can be seen by the simple change of variable technique. According to the expression of $\hat{\beta}$,

$$E(\hat{\beta}) = E \left[ \frac{1}{2T_n} \text{Re} \left\{ \int_{-T_n}^{T_n} \frac{\exp(ita)}{\varphi_Z(t)} dt \right\} \right]$$

$$= \int_{-\infty}^{\infty} \int_{-T_n}^{T_n} \frac{1}{2T_n} \text{Re} \left\{ \frac{\exp(ita)}{\varphi_Z(t)} \right\} f_Y(y) dy dt.$$

Under the assumption that $\varphi_Z(t) \neq 0$, $1/\varphi_Z(t)$ is a continuous function, hence it is bounded above on a compact set $[-T_n, T_n]$. In addition, $|e^{ita}|$ is bounded by 1. Hence the inner integrand in the above integration is absolutely integrable. So the order of integration can be changed based on Fubini’s theorem. Then, using the definition of the characteristic function of $Y$, and the relation between $\varphi_Z(\cdot)$, $\varphi_X(\cdot)$ and $\varphi_Y(\cdot)$, we have the following:

$$E(\hat{\beta}) = \frac{1}{2T_n} \text{Re} \int_{-T_n}^{T_n} \exp(-ita) \varphi_X(t) dt$$

$$= \frac{1}{2T_n} \int_{-T_n}^{T_n} \exp(-ita) \left\{ p \cdot \exp(ita) + (1 - p) \varphi_c(t) \right\} dt$$

$$= p + \frac{1 - p}{2T_n} \int_{-T_n}^{T_n} \exp(-ita) \varphi_c(t) dt.$$

This completes the proof. \hfill \Box

Proof of Lemma 2.3.2  When a random variable is symmetric about 0, its characteristic function is a real valued function, and symmetric about 0. So $\varphi_Z(\cdot)$ is a real valued even function. Then we can get

$$\text{Var}(\hat{\beta}) = \frac{1}{4nT_n^2} \text{Var} \left( \int_{-T_n}^{T_n} \frac{\cos t(Y - a)}{\varphi_Z(t)} dt \right)$$

$$= \frac{1}{nT_n^2} E \left( \int_{0}^{T_n} \frac{\cos tV - E(\cos tV)}{\varphi_Z(t)} dt \right)^2,$$

where $V = Y - a$. The second equality is possible from Fubini’s theorem and the fact
that \( \cos(\cdot) \) is also an even function. Recall the cosine product formula that
\[
2 \cos A \cos B = \cos(A + B) + \cos(A - B).
\]
Then the above equation becomes
\[
\Var(\hat{p}) = \frac{1}{2nT_n^2} \int_0^{T_n} \int_0^{T_n} \frac{E\{\cos(s + t)V + \cos(s - t)V\} - 2E(\cos sV)E(\cos tV)}{\varphi_Z(s)\varphi_Z(t)} ds \, dt
\]
\[
= \frac{1}{2nT_n^2} \int_0^{T_n} \int_0^{T_n} \frac{\Re\{\varphi_V(s + t) + \varphi_V(s - t)\} - 2\Re\{\varphi_V(s)\}\Re\{\varphi_V(t)\}}{\varphi_Z(s)\varphi_Z(t)} ds \, dt.
\]
This completes the proof. \( \square \)

**Proof of Theorem 2.3.3**  To show the consistency, we will show that both the bias and the variance of \( \hat{p} \) converges to 0. From Lemma 2.3.1,
\[
\text{bias}(\hat{p}) = \frac{1 - p}{2T_n} \int_{-T_n}^{T_n} \exp(-ita) \varphi_c(t) dt
\]
\[
= \frac{(1 - p) \pi}{T_n} \cdot \frac{1}{2\pi} \int_{-T_n}^{T_n} \exp(-ita) \varphi_c(t) dt.
\]
Clearly \( (1 - p)\pi/T_n \) converges to 0, and the latter part converges to \( f_c(a) \) because \( \varphi_c(t) \) is a characteristic function of a continuous random variable \( X_c \). Therefore the bias of \( \hat{p} \) converges to 0 as \( n \to \infty \).

Since \( |\varphi_V(t)| \leq 1 \) for any \( t \),
\[
|\Re\{\varphi_V(s + t) + \varphi_V(s - t)\} - 2\Re\varphi_V(s)\Re\varphi_V(t)| \leq 4.
\]
Then the variance of \( \hat{p} \) is bounded by
\[
\Var(\hat{p}) \leq \frac{2}{nT_n^2} \int_0^{T_n} \int_0^{T_n} \frac{1}{\varphi_Z(s)\varphi_Z(t)} ds \, dt = 2 \left( \frac{1}{n^{1/2} T_n} \int_0^{T_n} \frac{1}{\varphi_Z(t)} dt \right)^2.
\]
Hence it converges to 0 according to (2.7). \( \square \)
Proof of Theorem 2.3.4  First, we divide \( \hat{f}_c(x) \) into three parts:

\[
\hat{f}_c(x) = \frac{1}{2\pi(1-\hat{p})} \int_{-M_n}^{M_n} \Re \left\{ \frac{1}{n} \sum_{j=1}^{n} \exp \left( \frac{it(Y_j - x)}{\varphi_Z(t)} \right) - \hat{p} \cdot \exp \left( it(a - x) \right) \right\} dt
\]

\[
= \frac{1-p}{1-\hat{p}} \left[ \frac{Re}{2\pi(1-p)} \int_{-M_n}^{M_n} \left\{ \frac{1}{n} \sum_{j=1}^{n} \exp \left( \frac{it(Y_j - x)}{\varphi_Z(t)} \right) - p \cdot \exp \left( it(a - x) \right) \right\} dt \right.
\]

\[
+ \frac{Re}{2\pi(1-p)} \int_{-M_n}^{M_n} (p - \hat{p}) \exp \left( it(a - x) \right) dt \left. \right]
\]

\[
= T_1(T_2 + T_3),
\]

where

\[
T_1 = \frac{1-\hat{p}}{1-p}, \quad T_2 = \frac{1}{2\pi(1-p)} \Re \int_{-M_n}^{M_n} \left\{ \frac{1}{n} \sum_{j=1}^{n} \exp \left( \frac{it(Y_j - x)}{\varphi_Z(t)} \right) - p \cdot \exp \left( it(a - x) \right) \right\} dt,
\]

and \( T_3 = \frac{1}{2\pi(1-p)} \Re \int_{-M_n}^{M_n} (p - \hat{p}) \exp \left( it(a - x) \right) dt. \)

To show the consistency of \( \hat{f}_c(x) \), we will show that \( T_1 \to 1, T_2 \to f_c(x) \) and \( T_3 \to 0 \) in probability, as \( n \) goes to infinity.

Since \( \hat{p} \) converges to \( p \) in probability by Theorem 2.3.3, \( T_1 \) converges to 1 in probability. It is because \( \hat{p} \) is a consistent estimator of \( p \) and \( f(x) = 1/(1-x) \) is a continuous function of \( x \) except the case \( x = 1 \).

Now we show \( T_3 \) converges to 0. Since we only consider the case \( x \neq a \),

\[
T_3 = \frac{1}{2\pi(1-p)} (p - \hat{p}) \cdot \Re \int_{-M_n}^{M_n} e^{it(a-x)} dt
\]

\[
= \frac{p - \hat{p}}{2\pi(1-p)} \int_{-M_n}^{M_n} \cos t(a-x) dt = \frac{(p - \hat{p}) \sin M_n(a-x)}{\pi(1-p)(a-x)}. \tag{2.14}
\]

Here, \( |\sin M_n(a-x)| \) is uniformly bounded by 1, and we already showed \( \hat{p} \) converges to \( p \), i.e. \( \hat{p} - p \) converges to 0 in probability. Hence \( T_3 \) converges to 0 in probability.

For the last step, we will show that \( T_2 \) converges to \( f_c(x) \) in probability. For that, it suffices
to show that $E(T_2) \to f_c(x)$, and $\text{Var}(T_2) \to 0$ as $n$ goes to infinity. From the definition of $T_2$, 

$$E(T_2) = \frac{1}{2\pi(1-p)} \text{Re} \int_{-\infty}^{\infty} \int_{-M_n}^{M_n} \left\{ \frac{\exp(it(y-x))}{\varphi_z(x)} - p \cdot \exp(it(a-x)) \right\} f_Y(y) \, dt \, dy.$$ 

Since $\varphi_z(t) \neq 0$ for any $t$, the absolute value of the above integrand is bounded by an integrable function, i.e.

$$\left| \frac{\exp(it(y-x))}{\varphi_z(x)} - p \cdot \exp(it(a-x)) \right| f_Y(y) \leq \left( \frac{1}{\varphi_z(x)} + p \right) f_Y(y).$$ 

Then, by Fubini’s theorem, $E(T_2)$ is rewritten as

$$E(T_2) = \frac{1}{2\pi(1-p)} \text{Re} \int_{-M_n}^{M_n} \left\{ \varphi_Y(t) \frac{\exp(-itx)}{\varphi_z(t)} - p \cdot \exp(it(a-x)) \right\} dt$$ 

$$= \frac{1}{2\pi(1-p)} \text{Re} \int_{-M_n}^{M_n} \left\{ \varphi_Y(t) - p \cdot \exp(itx) \right\} \exp(-itx) \, dt.$$ 

The last equality comes from (2.5). Since $M_n$ goes to infinity as $n$ increases, 

$$E(T_2) \to \text{Re} \left( \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi_c(t) \exp(-itx) \, dt \right) = \text{Re} \left( f_c(x) \right) = f_c(x),$$

as $n$ goes to infinity.

The next part shows the calculation of the variance of $T_2$, which is very similar to the computation of $\text{Var}(\hat{p})$ in the proof of Lemma 2.3.2.

$$\text{Var}(T_2) = \text{Var} \left[ \frac{1}{2n\pi(1-p)} \sum_{j=1}^{n} \int_{-M_n}^{M_n} \left\{ \frac{\cos t(Y_j - x)}{\varphi_z(t)} - p \cdot \exp(it(a-x)) \right\} dt \right]$$ 

$$= \text{Var} \left[ \frac{1}{2n\pi(1-p)} \sum_{j=1}^{n} \int_{-M_n}^{M_n} \frac{\cos t(Y_j - x)}{\varphi_z(t)} dt \right]$$
By letting $V = Y - x$, we can get

$$
\text{Var}(T_2) = \frac{1}{n\pi^2(1-p)^2} E \left[ \int_0^{M_n} \frac{\cos tV - E \cos tV}{\varphi(t)} \right]^2
= \frac{1}{n\pi^2(1-p)^2} \int_0^{M_n} \int_0^{M_n} \left[ \frac{\text{Re} \{ \varphi V(s + t) + \varphi V(s - t) \}}{2} - \frac{\text{Re} \{ \varphi V(s) \} \text{Re} \{ \varphi V(t) \}}{2} \right] \times \frac{1}{\varphi Z(s) \varphi Z(t)} ds \ dt
\leq \frac{2}{\pi^2(1-p)^2} \left( n^{-1/2} \int_0^{M_n} \frac{1}{|\varphi Z(t)|} dt \right)^2.
$$

(2.15)

From (2.8), the above variance converges to 0 as $n$ goes to infinity, hence $T_2$ converges to $f_c(x)$ in probability.

Therefore $\hat{f}_c(x) = T_1(T_2 + T_3)$ converges to $f_c(x)$ in probability, i.e. $\hat{f}_c(x)$ is a consistent estimator of $f_c(x)$. □

**Proof of Theorem 2.3.5** The proof is the same as the proof of Theorem 2.3.4, except the convergence of $T_3$ in (2.13).

Since $M_n \to \infty$ and $M_n = o(T_n)$, $T_n$ also goes to infinity as $n \to \infty$. In addition, by (2.9),

$$
\frac{1}{n^{1/2} T_n} \int_0^{T_n} \frac{1}{\varphi Z(t)} dt = \frac{1}{T_n} \cdot n^{-1/2} \int_0^{T_n} \int_0^{T_n} \frac{1}{\varphi Z(t)} dt = \frac{1}{T_n} \cdot O(1) \to 0.
$$

This means that all conditions in Theorem 2.3.3 are satisfied, so $\hat{p}$ converges to $p$ in probability. Then $T_1$ and $T_2$ in the proof of Theorem 2.3.4 converge to 1 and $\hat{f}_c(x)$, respectively. The proof of these parts is exactly the same as that in the proof of Theorem 2.3.4.

The difficulty of providing the convergence of $T_3$ comes from the fact that the integration in (2.13) is not bounded when $x = a$. Since (2.13) is the same as $M_n(\hat{p} - p)/(\pi(1 - p))$, it suffices to show that $M_n(\hat{p} - p)$ converges to 0 in probability, in order to show the convergence of $T_3$. When the condition (2.9) is satisfied,

$$
E(M_n(\hat{p} - p)) = (1 - p) \cdot \frac{1}{T_n} \cdot 2\pi \int_{-T_n}^{T_n} \exp(-ita) \varphi_c(t) dt
\to (1 - p) \cdot 0 \cdot f_c(a) = 0.
$$

31
In addition,

\[
\text{Var}(M_n(\hat{p} - p)) = \frac{M_n^2}{2nT_n^2} \int_0^{T_n} \int_0^{T_n} \frac{Re\{\varphi_V(s + t) + \varphi_V(s - t)\} - 2Re\{\varphi_V(s)\}Re\{\varphi_V(t)\}}{\varphi_Z(s)\varphi_Z(t)} \, ds \, dt \\
\leq 2 \left( \frac{M_n}{n^{1/2}T_n} \int_0^{T_n} \frac{1}{\varphi_Z(t)} \, dt \right)^2 \to 0.
\]

This implies \( M_n(\hat{p} - p) \) converges to 0, so does \( T_3 \). Hence \( \hat{f}_c(a) = T_1(T_2 + T_3) \) converges to \( f_c(a) \) in probability.

**Proof of Lemma 2.3.7** Note that

\[
\text{Bias}(\hat{f}_X(x)) = \text{Bias}(\hat{p}) + (1 - p)\{E(T_2) - f_c(x)\} + (1 - p)E(T_3),
\]

where \( T_2 \) and \( T_3 \) are the same as those in the proof of Theorem 2.3.4.

From Lemma 2.3.1, the bias of \( \hat{p} \) is given by

\[
\text{Bias}(\hat{p}) = \frac{(1 - p)\pi}{T_n} \cdot \frac{1}{2\pi} \int_{-T_n}^{T_n} \varphi_c(t) \exp(-ita) \, dt.
\]

Since \( f_c \) is continuous, \( \varphi_c(\cdot) \) is integrable. In addition, \( T_n \) goes to infinite as \( n \to \infty \). Hence, from the Inverse-Fourier transformation,

\[
\frac{1}{2\pi} \int_{-T_n}^{T_n} \varphi_c(t) \exp(-ita) \, dt \to f_c(a), \quad \text{as} \quad n \to \infty.
\]

I.e. regardless of \( f_c \), the bias of \( \hat{p} \) has the order \( O(T_n^{-1}) \), as \( n \to \infty \). In addition, from (2.14), we get

\[
E(T_3) = O(\text{Bias}(\hat{p})) = O(T_n^{-1}).
\]

The order of \( E(T_2) - f_c(x) \) is determined by the tail property of \( \varphi_c \); Again from the
Inverse-Fourier transformation,

\[ |E(T_2) - f_c(x)| = \left| \frac{1}{2\pi} \int_{-M_n}^{M_n} \varphi_c(t) \exp(-itx) dt - \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi_c(t) \exp(-itx) dt \right| \]
\[ \leq \frac{1}{2\pi} \int_{-\infty}^{\infty} |\varphi_c(t)| dt + \frac{1}{2\pi} \int_{M_n}^{\infty} |\varphi_c(t)| dt = I_1 + I_2 \]

First, consider the case (B1). Then, for sufficiently large \( n \),

\[ I_2 = \int_{M_n}^{\infty} |\varphi_c(t)| dt \leq \int_{M_n}^{\infty} d_1 t^{-\beta_1} dt = O(M_n^{1-\beta_1}), \]

and by the similar computation, \( I_1 \) can be easily shown to have the same order as \( I_2 \).

Now, suppose that (B2) holds. When \( n \) is large enough,

\[ I_2 = \int_{M_n}^{\infty} d_1 \exp \left( -\frac{t^{\beta_1}}{\gamma_1} \right) dt \]
\[ \leq d_1 \int_{M_n}^{\infty} \left( \frac{t}{M_n} \right)^{\beta_1-1} \exp \left( -\frac{t^{\beta_1}}{\gamma_1} \right) dt = O(M_n^{1-\beta_1} \exp \left( -\frac{M_n^{\beta_1}}{\gamma_1} \right)), \]

and \( I_2 \) has the same order as \( I_1 \). This completes the proof. \( \square \)

**Proof of Lemma 2.3.8**  From (2.10), \( \hat{f}_X(x) \) can be expressed as

\[ \hat{f}_X(x) = \hat{p} \delta_a(x) + (1 - p)(T_2 + T_3), \]

where \( T_2 \) and \( T_3 \) are defined in the proof of Theorem 2.3.4. Hence the variance of \( \hat{f}_X(x) \) has the form

\[ \text{Var}(\hat{f}_X(x)) = (1 - p)^2 \{ \text{Var}(T_2) + \text{Var}(T_3) + 2\text{Cov}(T_2, T_3) \} \]
\[ + \{ \text{Var}(\hat{p}) + 2(1 - p)\text{Cov}(\hat{p}, T_3) + 2(1 - p)\text{Cov}(\hat{p}, T_2) \} \delta_a(x). \]

Note that \( \text{Var}(T_3) = O(\text{Var}(\hat{p})) \) and \( \text{Cov}(\hat{p}, T_3) = O(\text{Var}(\hat{p})) \) from (2.14). In addition, by
Cauchy-Schwartz inequality,

\[ |\text{Var}(T_2) + \text{Var}(T_3) + 2\text{Cov}(T_2, T_3)| \leq 2\{\text{Var}(T_2) + \text{Var}(T_3)\}, \]

\[ |\text{Cov}(\hat{p}, T_2)| \leq \sqrt{\text{Var}(\hat{p})}\sqrt{\text{Var}(T_2)}. \]

Hence we get \( \text{Var}(\hat{f}_X(x)) = O(\text{Var}(\hat{p})) + O(\text{Var}(T_2)) \), as \( n \to \infty \).

Suppose that (V1) is satisfied. Then, there exists a constant \( c > 1 \) such that \( 1/\|\varphi_Z(t)\| \leq (1/d_2)d^{\beta_2} \) for any \( t \geq c \). Note that \( \varphi_Z(t) \) is continuous and \( \varphi_Z(t) \neq 0, 1/\|\varphi_Z(t)\|^2 \) is also a continuous function, hence is integrable on a compact interval \([0, c]\). From (2.12) and Jensen’s inequality,

\[
\text{Var}(\hat{p}) \leq \frac{2}{nT_n^2} \int_0^{T_n} \frac{1}{\|\varphi_Z(t)\|^2} dt = \frac{2}{nT_n^2} \left[ \int_0^c \frac{1}{\|\varphi_Z(t)\|^2} dt + \int_c^{T_n} \frac{1}{\|\varphi_Z(t)\|^2} dt \right] \leq \frac{2}{nT_n^2} \left[ \int_0^c \frac{1}{\|\varphi_Z(t)\|^2} dt + \int_c^{T_n} \frac{1}{d_2^2} t^{2\beta_2} dt \right] = O\left(n^{-1}T_n^{2\beta_2-1}\right),
\]

and similarly by (2.15), \( \text{Var}(T_2) = O(n^{-1}M_n^{1+2\beta_2}) \). Finally, we get

\[
\text{Var}(\hat{f}_X(x)) = O(n^{-1}M_n^{1+2\beta_2}) + O(n^{-1}T_n^{2\beta_2-1}).
\]

Now, consider the case (V2). We can also find a constant \( c \) such that \( \|\varphi_Z(t)\|\exp(t^{\beta_2}/\gamma_2) > d_2 \) for any \( t \geq c \). In this case,

\[
\int_c^{T_n} \frac{1}{\|\varphi_Z(t)\|^2} dt \leq \int_c^{T_n} \frac{1}{d_2^2} \exp\left(\frac{2t^{\beta_2}}{\gamma_2}\right) dt = O\left(T_n \exp\left(\frac{2T_n^{\beta_2}}{\gamma_2}\right)\right),
\]

and hence

\[
\text{Var}(\hat{p}) = O\left(\frac{1}{nT_n} \exp\left(\frac{2T_n^{\beta_2}}{\gamma_2}\right)\right) \quad \text{and} \quad \text{Var}(T_2) = O\left(\frac{M_n}{n} \exp\left(\frac{2M_n^{\beta_2}}{\gamma_2}\right)\right).
\]

In particular, when \( \beta_2 \geq 1 \), we can get a better upper bound of the variance; From the
fact that \( t^{\beta_2} \leq t \cdot T_n^{\beta_2-1} \) for any \( 0 < t \leq T_n \), therefore

\[
\int_c^{T_n} \frac{1}{\varphi_Z(t)} dt \leq \int_c^{T_n} \frac{1}{d_2} \exp \left( \frac{t^{\beta_2}}{\gamma_2} \right) dt \\
\leq \int_c^{T_n} \frac{1}{d_2} \exp \left( \frac{t \cdot T_n^{\beta_2-1}}{\gamma_2} \right) dt = O \left( T_n^{1-\beta_2} \exp \left( \frac{T_n^{\beta_2}}{\gamma} \right) \right).
\]

Combined with (2.12), it gives

\[
\text{Var}(\hat{\beta}) = O \left( \frac{1}{n T_n^{2\beta_2}} \exp \left( \frac{2T_n^{\beta_2}}{\gamma} \right) \right), \quad \text{and Var}(T_2) = O \left( \frac{M_n^{2(1-\beta_2)} \exp \left( \frac{2M_n^{\beta_2}}{\gamma} \right)}{n} \right),
\]

which completes the proof. □

**Proof of Theorem 2.3.9** Define \( \alpha = \min(\beta_1 - 1, 1) \). Let \( M_n = (\gamma_2/4)^{1/\beta_2} (\log n)^{1/\beta_2} \) and let \( T_n = (\gamma_2/4)^{1/\beta_2} (\log n)^{\alpha/\beta_2} \). By Lemma 2.3.7 and the fact \( 0 < \alpha \leq \beta_1 - 1 \), we can get

\[
\text{Bias}(\hat{f}_X(x)) = O \left( (\log n)^{-2\alpha/\beta_2} \right) + O \left( (\log n)^{-2(\beta_1 - 1)/\beta_2} \right) = O \left( (\log n)^{-2\alpha/\beta_2} \right),
\]

as \( n \to \infty \).

Since \( \alpha \) is at most 1, \( (2/\gamma_2)T_n^{\beta_2} = (1/2)(\log n)^{\alpha} \leq (1/2) \log n, \) for any \( n \geq 3 \). Then, from Lemma 2.3.8, we can get

\[
\text{Var}(\hat{f}_X(x)) = O \left( n^{-1} T_n^{-2\beta_2} \exp \left( \frac{2T_n^{\beta_2}}{\gamma_2} \right) \right) + O \left( n^{-1} M_n^{-2(\beta_2-1)} \exp \left( \frac{2M_n^{\beta_2}}{\gamma_2} \right) \right) \\
\leq O \left( n^{-1/2} (\log n)^{-2\alpha} \right) + O \left( n^{-1/2} (\log n)^{-2(\beta_2-1)/\beta_2} \right) \\
= o(n^{-1/3}) = o((\log n)^{-2\alpha/\beta_2}), \quad \text{as } n \to \infty.
\]

Here, we can see the bias term dominates the variance. Finally we can get

\[
E(\hat{f}_X(x) - f_X(x))^2 = \text{Bias}^2(\hat{f}_X(x)) + \text{Var}(\hat{f}_X(x)) = O \left( (\log n)^{-2\alpha/\beta_2} \right),
\]

as \( n \to \infty \). This completes the proof. □.
Chapter 3

Sieve Type Deconvolution Based on Maximum Likelihood

3.1 Introduction

In Chapter 2, we proposed the direct deconvolution (DC) estimator for the special mixture distributions in (2.1) based on Fourier deconvolution. However, the proposed estimator gives poor performance at the known boundary; the DC estimator performs well when the target distribution is smooth, e.g. normal distributions, but has boundary effects when the target distribution has non-smooth boundaries. For example, in the exponential simulation in Section 2.4.3, the direct deconvolution density estimators put some probability mass on the negative half-line, even the true density is only supported on the positive real line. Our motivating application involves a single atom at zero and a continuous distribution whose density is assumed to be supported only on the positive half-line. However it is not clear how to use this known boundary information in implementing the direct deconvolution estimators, hence it suffers from the boundary problem.

Here, we more focus on right handling of the known boundary information in estimating distributions in measurement error models. In addition, we consider more general distribution of $X$ which can contain finitely many single atoms, i.e. $0 \leq \nu < \infty$ in (1.2).

Parametric fit can easily reflect the boundary by choosing distributions with bounded support. However, it is extremely hard to check the validity of the assumed distribution.
Pensky (2002), Hall and Qiu (2005), Meister (2007) and Zhang and Karunamuni (2008) propose nonparametric deconvolution estimators which reduce the boundary effects; however they only consider the case where $X$ is purely continuous. It might be possible to extend their algorithms to our case, but it is not very straightforward. Ruppert et al. (2007) propose a sieve type density estimator when the boundary is known. However, they consider some special continuous mixture distributions, and do not incorporate measurement errors.

In this chapter, we approach the estimation problem via three main ideas: discretization, maximum likelihood and penalization. First, we approximate the distribution of $X$ using discretization, which gives a sieve of the distribution family. Then, we estimate the distribution based on the maximum likelihood method in each sieve. The mixture structure and the known boundary information are reflected in the construction of the sieve. The measurement error problem is then solved via the computation of the likelihood function. In addition, when the smoothness of the target distribution is assumed, we improve the proposed basic estimator by using a roughness penalty function.

The remainder of this chapter is organized as follows. In Section 3.2, we explicitly state the problem of our interest and the model, and then propose two sieve estimators along with some estimation algorithms. Section 3.3 studies some minimal conditions for the consistency of the proposed distribution estimators. Section 3.4 shows the performance of the proposed estimators via a simulation study, and in Section 3.5, we apply the proposed estimators to the virus-lineage data given by Burch et al. (2007), and validate the traditional exponential assumption. Some technical details and proofs of the theorems appear in the Section 3.6.

### 3.2 Model and Methodology

In this section, first we describe the model which we are interested in, and then develop two sieve estimation procedures. In particular, Section 3.2.2 provides the basic standard sieve estimator, and Section 3.2.4 improves the proposed method by introducing a roughness penalization on the estimator. Theoretical properties of the two estimators are investigated in Section 3.3.
3.2.1 Model

Suppose we observe independent and identically distributed data \( Y_1, \ldots, Y_n \), generated as
\[ Y = X + Z, \]
here \( X \) and \( Z \) are independent, \( Z \) has a known density \( f_Z \), and we wish to estimate the distribution of \( X \). We assume that, for an integer \( \nu \geq -1 \),
\[ F_X(x) = P(X \leq x) = p_1 I(a_1 \leq x) + \cdots + p_\nu I(a_\nu \leq x) + p_{\nu+1} P(X_c \leq x), \tag{3.1} \]
where the nonnegative quantities \( p_1, \ldots, p_{\nu+1} \) add up to 1, and the random variable \( X_c \) has a continuous distribution. Without loss of generality, we can assume \( a_l < a_{l+1} \) for any \( l \). The case \( \nu = 0 \) corresponds to \( X \) being continuous, or, in effect, \( X = X_c \). It will be assumed that \( \nu \) and the atoms \( a_1, \ldots, a_\nu \), although not their masses \( p_1, \ldots, p_\nu \), are known, which is true for our virus lineage application. We wish to estimate \( p_1, \ldots, p_{\nu+1} \) and the distribution of \( X_c \).

The practical problem discussed in Section 3.5, and which motivated our work, involved \( \nu = 1 \) and \( a_1 = 0 \), and required \( X_c \) to be distributed on the positive half-line. Physical considerations indicated that the density of \( X_c \) would likely have a jump discontinuity at the origin, but be continuous on \((0, \infty)\). Even if there were no atom at zero, a conventional method based on Fourier deconvolution, (i) would not respond well to the presence of a jump, (ii) would be difficult to modify so as to incorporate that information, and (iii) would produce a density estimator that took positive values on the negative half-line and was oscillatory, and sometimes negative, in the tails. For more details, see Chapter 2.

The sieve techniques developed below do not suffer from these difficulties. Even in cases where there is no atom and the density of \( X_c \) is smooth on the real line and without discontinuities, our methodology is attractive, because it produces distribution and density estimators which are bona fide distribution and density functions, respectively. This is of substantial value in many practical problems. In particular, if one of our aims in estimating the distribution of \( X \) is to enable bootstrap simulation from the estimated distribution, then it is essential that the estimator is a proper distribution function. Another advantage of our method is the reduction of computational cost; the computation time is much shorter than in the case of the direct deconvolution estimator in Chapter 2.
3.2.2 Basic Description of the Estimation Method

We first consider the case where the distribution of $X$ has a single atom ($\nu = 1$), located at zero ($a_1 = 0$), and the distribution of $X_c$ is supported on $(0, \infty)$. Our proposed method is of the sieve type (Grenander, 1981): we first consider a sieve, which is a sequence of classes of specific distributions, and restrict the problem to the estimation in the sieve. Then the sieve is extended to the entire distribution space as the sample size increases, and hence we obtain a solution of the original estimation problem.

In this setting we establish a lattice on the positive half-line, taking it to be the sequence of points $(j - 0.5)h$ for $j \geq 1$, where $h > 0$ plays a role not unlike that of a bandwidth, or of a binwidth in histogram estimation. The distribution of $X_c$ in (3.1) is then approximated by the distribution of a random variable $\tilde{X}_c$ with potential atoms at each of the points $x_j = (j - 0.5)h$:

$$P(\tilde{X}_c = x_j) = \theta_j \quad \text{for } j = 1, \ldots, r.$$  \hspace{1cm} (3.2)

Here, each $\theta_j$ is nonnegative, and $\sum_j \theta_j = 1$, i.e. $\theta = (\theta_1, \ldots, \theta_r)^T$ determines an a discrete distribution with $r$ possible values. Having computed estimators $(\hat{p}_1, \hat{\theta})$ of $(p_1, \theta)$, see Section 3.2.3 below, we take $(\hat{p}_1, (1 - \hat{p}_1)\hat{\theta}^T)^T$ to be our initial estimator of the distribution of $X$, approximated with atoms at zero and $x_j$ for $j = 1, \ldots, r$. We describe at the end of Section 3.2.3 how to construct smooth estimators of the distribution and density functions of $X_c$.

We considered a model where there was an atom at zero and the distribution of $X_c$ was approximated by a histogram; in particular, $X_c$ was taken to be uniformly distributed on $[(j - 1)h, jh]$, instead of having a mass at $x_j = (j - 0.5)h$ for $j \geq 1$. However, this approach did not perform as well as the method suggested by (3.2).

The proposed method can easily be extended to the general case where $1 < \nu < \infty$ and $X_c$ lies in a finite interval $[a, b]$. The only difference is choices of the lattice points $x_1, \ldots, x_r$ in discretizing $X_c$. When every $a_i$ lies outside $[a, b]$, it is simple to find $x_j$. One of the simplest choice is $x_j = a + (j - 0.5)h$ for any $j$, which gives equally spaced grid points covering the range of $X_c$.

A problem arises when some atoms of $X$ lie on $[a, b]$. In order to make the estimation
identifiable, each \( x_j \) should be different from \( a_l \), for any \( l = 1, \ldots, \nu \). If there are just \( \nu = 2 \) atoms in the interval \([a, b]\) with known locations \( a_1 < a_2 \), we take \( h \) to equal \((a_2 - a_1)/m\) for an integer \( m \), and take \( x_j \)s covering the support of \( f_c \) and satisfying

\[
x_{j+1} = x_j + h \quad \forall \ j, \quad \text{and} \quad \min_j |a_1 - x_j| = \min_j |a_2 - x_j| = \frac{h}{2}.
\]

If there are \( \nu > 2 \) atoms then we use potentially different values of \( h \) between adjacent atoms, to ensure that the pseudo atoms of \( X_c \) are equally spaced there. We use the term pseudo atoms since the majority of the \( \theta_j \)s are artifacts of our method for approximating the distribution of \( X_c \); they have no direct counterpart in the true distribution.

### 3.2.3 A Standard Sieve Estimator

Suppose that the measurement error \( Z \) has a known density function \( f_Z \), and let \( f_Y \) be the density of \( Y \). In view of (1.1), our approximation, based on (3.2), to \( f_Y \) is given by

\[
f_Y(y|\mathbf{p}, \mathbf{\theta}) = \sum_{l=1}^{\nu} p_l f_Z(y - a_l) + p_{\nu+1} \sum_{j=1}^{r} \theta_j f_Z(y - x_j),
\]

where \( \mathbf{p} = (p_1, \ldots, p_{\nu+1})^T \). There is a variety of different ways of estimating the parameters \( \mathbf{p} \) and \( \mathbf{\theta} \). We shall suggest one of them below, namely maximum likelihood. The theoretical properties of the estimation procedure will be investigated in Section 3.3.

According to (3.3), the log-likelihood function of \((\mathbf{p}, \mathbf{\theta})\) can be written as

\[
l(\mathbf{p}, \mathbf{\theta}) = \sum_{i=1}^{n} \log \left[ \sum_{l=1}^{\nu} p_l f_Z(Y_i - a_l) + p_{\nu+1} \sum_{j=1}^{r} \theta_j f_Z(Y_i - x_j) \right].
\]

We propose the standard sieve (SS) estimator for \((\mathbf{p}, \mathbf{\theta})\) as the maximizer of the above log-likelihood function, i.e.,

\[
(\hat{\mathbf{p}}^{\text{SS}}, \hat{\mathbf{\theta}}^{\text{SS}}) = \arg\max_{\mathbf{p}, \mathbf{\theta}} \left\{ l(\mathbf{p}, \mathbf{\theta}) : p_l \geq 0, \ \theta_j \geq 0, \ \sum_{l=1}^{\nu+1} p_l = 1, \ \text{and} \ \sum_{j=1}^{r} \theta_j = 1 \right\}.
\]

As it turns out, it is nontrivial to solve the above constrained maximization problem.
Below we describe an iterative algorithm to maximize the log-likelihood in (3.4).

**An Iterative Maximization Algorithm**

**Step 1.** Initialize: set \( p = (1, \ldots, 1)^T/(\nu + 1) \) and \( \theta = (1, \ldots, 1)^T/r \), and denote them as \( p^{(0)} \) and \( \theta^{(0)} \), respectively;

**Step 2.** Update:

(a) \( \theta^{(1)} = \arg\max_{\theta} \left\{ l(\theta|p^{(0)}) : \theta_j \geq 0 \text{ and } \sum \theta_j = 1 \right\} \);
(b) \( p^{(1)} = \arg\max_{p} \left\{ l(p|\theta^{(1)}) : p_l \geq 0 \text{ and } \sum p_l = 1 \right\} \);

**Step 3.** Repeat Step 2 by setting \( p^{(0)} = p^{(1)} \) and \( \theta^{(0)} = \theta^{(1)} \) until convergence.

Note that in Step 2 of the above algorithm, \( l(\theta|p^{(0)}) \) denotes the conditional log-likelihood of \( \theta \) given that \( p = p^{(0)} \), while \( l(p|\theta^{(1)}) \) is similarly defined. In Step 2(a), due to the constraint that \( \theta \) determines a probability distribution, the maximization of the conditional log-likelihood \( l(\theta|p^{(0)}) \) is not simple. This constrained maximization can be carefully solved using a combination of the Karush-Kuhn-Tucker (KKT) condition in optimization (Bertsekas, 2005) and Newton’s iteration algorithm. Technical details of this optimization algorithm can be found in Section 3.6. Since \( p \) also determines a discrete distribution, the maximization in Step 2(b) is established as in Step 2(a).

Once we obtain the estimator \( (\hat{p}^{SS}, \hat{\theta}^{SS}) \), we take \( \hat{p}^{SS} \) to be our estimator of \( p = (p_1, \ldots, p_{\nu+1})^T \). In addition, as a density estimator of \( X_c \), we construct a continuous function on \((x_1-0.5h, x_r+0.5h)\) by interpolating \((x_j, \hat{\theta}_j)\)s. For example, we can use linear interpolation as follows. Notice that

\[
\begin{align*}
\theta_j & = P(\tilde{X}_c = x_j) = P(x_j - 0.5h \leq \tilde{X}_c < x_j + 0.5h) \\
& \approx P(x_j - 0.5h \leq X_c < x_j + 0.5h) \approx h f_c(x_j).
\end{align*}
\]

For sufficiently small \( h \) and a smooth function \( f_c \), the interpolated estimator can be con-
constructed as
\[
\tilde{f}^{SS}_c(x) = \frac{\hat{\theta}_{j-1}}{h} + \frac{\hat{\theta}_j - \hat{\theta}_{j-1}}{h(x_j - x_{j-1})} (x - x_j)
\] for \( x \in [x_{j-1}, x_j) \) for any \( 1 \leq j \leq r + 1 \), and 0 otherwise. We set \( x_0 = x_1 - 0.5h, x_{r+1} = x_r + 0.5h, \hat{\theta}_0 = \hat{\theta}_1 \) and \( \hat{\theta}_{r+1} = \hat{\theta}_r \). Then \( \tilde{f}^{SS}_c \) is a continuous density function which is supported in \( [x_0, x_0 + rh) \), and has the value 0 outside the interval. The distribution estimator \( \tilde{F}^{SS}_c \) can be easily obtained from \( \tilde{f}^{SS}_c \).

3.2.4 A Penalized Sieve Estimator

The standard sieve estimation described above in Section 3.2.3 involves discretizing the distribution of the continuous component \( X_c \). As a result of this discretization, the resulting density estimator \( \hat{f}^{SS}_c \) can be rather rough for finite samples. To reduce this problem, in this section, we introduce into the estimation procedure a roughness penalty on \( \theta \).

More specifically, we consider the following penalized log-likelihood function,

\[
l_{\lambda}(\mathbf{p}, \mathbf{\theta}) = l(\mathbf{p}, \mathbf{\theta}) - \lambda P(\mathbf{\theta}),
\]

where \( l(\mathbf{p}, \mathbf{\theta}) \) is the log-likelihood in (3.4), \( P(\cdot) \) is some roughness penalty on \( \mathbf{\theta} \), and \( \lambda \) is a penalty parameter that balances the effects of the log-likelihood and the penalty term. We then propose the penalized sieve (PS) estimator for \((\mathbf{p}, \mathbf{\theta})\) as the maximizer of the above penalized log-likelihood (3.6), i.e.,

\[
(\hat{\mathbf{p}}^{PS}, \hat{\mathbf{\theta}}^{PS}) = \arg\max_{\mathbf{p}, \mathbf{\theta}} \left\{ l_{\lambda}(\mathbf{p}, \mathbf{\theta}) : p_l \geq 0, \theta_j \geq 0, \sum_{l=1}^{r+1} p_l = 1, \text{ and } \sum_{j=1}^{r} \theta_j = 1 \right\}.
\]

Considering again the penalized criterion (3.6), the roughness penalty \( P(\cdot) \) can be any function that decreases when \( \mathbf{\theta} \) gets smoother. For example, we can choose the sum of first order squared differences, i.e. \( P(\mathbf{\theta}) = \sum_{j=2}^{r} (\theta_j - \theta_{j-1})^2 \). This function has the minimum value 0 when all \( \theta_j \)s are the same. Another possible choice is the sum of second order squared differences, \( P(\mathbf{\theta}) = \sum_{j=2}^{r-1} (\theta_{j+1} - 2\theta_j + \theta_{j-1})^2 \), which is minimized when the \((x_j, \theta_j)\) form a straight line.
The penalty parameter $\lambda$ plays the role of a smoothing parameter. When $\lambda = 0$, the penalty term disappears, and $\hat{\theta}^{PS}$ is exactly the same as $\hat{\theta}^{SS}$, the standard sieve estimator. As we pointed out earlier, the density estimator may not be smooth in this case. On the other hand, as $\lambda$ goes to infinity, the penalty term dominates the log-likelihood. As a result, the estimator is flat in the limit if the first order squared difference is used as the penalty function. We illustrate this trade-off via a simulation study in Section 3.4.3.

An alternative approach is to combine the two sieve estimation procedures. First, we estimate $p$ by the standard sieve estimator $\hat{p}^{SS}$. After that, we apply the penalized sieve estimation idea to get a smooth density estimator for $f_c$. This means that we define $\hat{\theta}^{PS}$ as the maximizer of the following penalized conditional log-likelihood function,

$\hat{\theta}^{PS} = \arg\max_\theta \left\{ l(\theta | \hat{p}^{SS}) - \lambda P(\theta) : \theta_j \geq 0 \text{ and } \sum \theta_j = 1 \right\}.$

After obtaining $\hat{\theta}^{PS}$, we can use linear interpolation in the same manner as discussed in Section 3.2.4. This hybrid method results in a discrete distribution estimator $\hat{p}$ with small bias, as well as a smooth density estimator.

### 3.3 Consistency of the Proposed Estimators

In this section, our aim is to show that a consistent estimator of the distribution can be obtained under the minimal assumption that the bandwidth $h$ is of larger order than $n^{-1}$. That is, under some regularity conditions, $\hat{F}_X(x)$ converges to $F_X(x)$ with probability 1, where

$$\hat{F}_X(x) = \hat{p}_1 I(a_1 \leq x) + \cdots + \hat{p}_\nu I(a_\nu \leq x) + \hat{p}_{\nu+1} \sum_{j=1}^\nu \hat{\theta}_j I(x_j \leq x), \quad (3.7)$$

and $F_X$ is the true distribution of $X$ as defined in (3.1). (It can be proved that if $h$ is $O(n^{-1})$, then consistency is generally not possible.) We focus less on consistency of the density estimator, since our conditions on $f_c$ do not require continuity of that function at any point. If continuity of $f_c$ is assumed, then it can be proved that the interpolated estimator in (3.5) is a consistent density estimator.
In order to make our results simple to frame, we assume that the distribution of \( X_c \) has a bounded density \( f_c \) supported in a compact interval \([a, b]\). In general, it is not required that the exact values of \( a \) and \( b \) are known. We construct the distribution estimator in a potentially larger interval \([a^*, b^*]\), where \(-\infty < a^* \leq a < b \leq b^* < \infty\). Let \( c = \min\{a_1, a^*\} \) and let \( d = \max\{a_\nu, b^*\} \). Define

\[
g(y) = g(y|c,d) = \sup_{c \leq u_1, u_2 \leq d} \frac{f_Z(y-u_1)}{f_Z(y-u_2)}, \tag{3.8}
\]

and write \( \mathcal{F} \) for the set of all distributions which have a mixture structure with atoms at \( a_1, \ldots, a_\nu \) and a continuous distribution supported in \([a, b]\).

In addition, we assume that for some \( x_0 \in [c, d] \), and a constant \( B > 0 \) not depending on \( s \),

\[
E \left| \log g(Y)^s \right| + E \left| \log f_Z(Y - x_0)^s \right| \leq s! B^s, \tag{3.9}
\]

\[
\inf_{F \in F} \text{Var}\left[ \log \left\{ \int f_Z(Y - x) \, dF(x) \right\} \right] > 0. \tag{3.10}
\]

Property (3.9) can be verified by direct calculation for a large class of densities \( f_Z \). For example, the normal and Laplace distributions are both particular cases of the set of Subbotin distributions \( \text{SD}_\gamma(\mu, \sigma) \), with probability density

\[
f(x|\gamma, \mu, \sigma) = \frac{C_\gamma}{\sigma} \exp\left( -\frac{|x - \mu|^{\gamma}}{\gamma \sigma^{\gamma}} \right),
\]

where \( \gamma, \sigma > 0, -\infty < \mu < \infty \) and \( C_\gamma^{-1} = 2 \Gamma(1/\gamma) \gamma^{(1/\gamma)-1} \). See Donoho and Jin (2004) for discussion of this class. Property (3.9) holds for these and many other distributions, for example distributions for which \( f_Z \) has regularly varying tails, as in the case of Student’s \( t \)-distribution. Property (3.10) can be established by contradiction, noting that if it fails then there is an \( F \in \mathcal{F} \) for which the variance equals zero, and that this is not possible for the classes of error distributions just mentioned.
Next we list regularity conditions, in which we take

\[ f_X(x) = \sum_{l=1}^{\nu} p_l \delta_{a_l}(x) + p_{\nu+1} f_c(x) \]  

(3.11)

to be the generalized density (Cuevas and Walter, 1992) of \( X \),

(R1) \( f_c \) is supported in the interval \([a, b]\) and bounded above by a constant \( C > 0 \);

(R2) we construct the histogram approximation \( \theta = (\theta_1, \ldots, \theta_r)^T \) to \( f_c \) in the interval \([a^*, b^*]\),

where \(-\infty < a^* \leq a < b \leq b^* < \infty\), in which case \( rh \leq b^* - a^* \);

(R3) the error density \( f_Z \) satisfies the conditions (3.9) and (3.10);

(R4) we restrict each \( \theta_j \) by insisting that \( \theta_j \leq C_1 h \), where \( C_1 \geq C \), the latter constant is as

in (R1), and \( C_1 \) is arbitrarily large;

(R5) \( r = o(n) \);

(R6) the distribution of \( X \) is uniquely identifiable from its convolution with the distribution

of \( Z \).

Assumptions (R1)–(R3) merely formalize constraints discussed earlier; (R4) requires that
our construction of the estimator reflects the boundedness assumption, but permits our prior
impression of the bound to be arbitrarily large; (R5) is the key assumption, and, since it
requires only \( r = o(n) \), or equivalently that the bandwidth \( h \) be of larger order than \( n^{-1} \), it
is the weakest possible condition of this type; and (R6) is a necessary condition, and holds if
(for example) the characteristic function of \( Z \) vanishes only on a set of measure zero. When
\( X_c \) has a smooth density \( f_c \), the constraint (R4) is not necessary in practice, since estimators
of the masses \( \theta_j \) are only very rarely much larger than the probabilities associated with the
corresponding histogram blocks.

**Theorem 3.3.1.** Suppose that (R1)–(R6) hold. Then, with probability 1, the distribution
estimator given in (3.7), where \( \hat{p} = \hat{p}^{SS} \) and \( \hat{\theta} = \hat{\theta}^{SS} \) are obtained by the standard sieve
estimation, converges to the true distribution of $X$.

The next theorem establishes the parallel consistency result for the penalized sieve distribution estimator. For that, we need some extra conditions on the roughness penalty function and the penalty parameter.

**Theorem 3.3.2.** In addition to (R1)–(R6), suppose that the roughness penalty $P(\cdot)$ is asymptotically bounded. If the penalty parameter $\lambda$ increases slower than $n$, i.e. $\lambda = o(n)$ as $n \to \infty$, then $\hat{F}_X(x)$ given by the penalized sieve estimation converges to the true distribution of $X$ with probability 1.

The detailed proofs of the above two theorems are given in the Section 3.6.1.

### 3.4 A Simulation Study

In this section, we perform a simulation study to investigate the performance of the proposed estimators.

#### 3.4.1 Simulation Description

To make the simulation scientifically relevant, our simulation scheme reflects the biological context of the mutation effect distribution analyzed in Section 3.5. In biological experiments when estimating the distribution of mutation effects on fitness ($S$), the most common approach is parametric, which fits an exponential (or gamma) distribution. In practice, the mutation effect itself cannot be measured directly; hence evolutionary biologists measure the mutation effect through the change in a fitness-related characteristic over time. For our motivating application, Burch and Chao (2004) measured virus plaque size reduction ($X$), and related it to $S$ using the experimentally determined parametric relationship (2.11).
For our simulation, we first generate $S$ from a mixture of a pointmass at 0 and an exponential distribution such that

$$S \sim \begin{cases} 
\text{Exp}(0.12) \text{ with probability } .1, \\
0 \text{ with probability } .9, 
\end{cases}$$

where \(\text{Exp}(0.12)\) stands for an exponential distribution with mean 0.12. We then define $X$ according to (2.11). Hence, the distribution of $X$ is a mixture of the pointmass at 0 with probability 0.9, and a continuous distribution with density $f_c$. In addition, $f_c$ is supported only on the positive real line. This means that the distribution of $X$ has a jump discontinuity at 0, but is continuous in the rest of the support.

We consider $X$ as the unobservable variable in whose distribution we are interested. What we actually observe are the measurement error perturbed observations $Y = X + Z$, where the measurement error $Z$ has a normal distribution with mean 0 and standard deviation 0.48. This value is suggested by analyzing the virus lineage data. We then simulate 100 random samples of size $n = 350$ following the above simulation scheme. If interested as in the real application of Section 3.5, the density estimate of $S$ can be easily obtained from (2.11) and simple change of variables.

### 3.4.2 Pointmass Estimation

For the pointmass estimation, we compare our standard sieve estimator with the direct deconvolution (DC) estimator proposed in Chapter 2. Note that the DC estimator makes use of Fourier deconvolution, which has difficulty incorporating boundary information.

The two panels of Figure 3.1 plot the 100 estimates given by the two estimators, respectively. In each panel, the gray dots correspond to the individual estimates plotted at random jittered heights to separate them out; we also superimpose a kernel density estimate (the black solid curve); the dotted and dashed vertical lines indicate the location of the true pointmass and the average of the 100 estimates. To make the comparison clear, the two panels are plotted on the same horizontal range.

As one can see, the standard sieve (SS) estimator has a smaller bias and a smaller variance.
Figure 3.1: Comparison of the pointmass estimation between (a) the direct deconvolution estimator and (b) the standard sieve estimator. The dots are the 100 individual estimates with jittered heights, superimposed on a kernel density estimate. The dotted vertical line is the true pointmass. The dashed vertical line is the average estimate.

than the DC estimator. In fact, the average of the SS estimator is 0.898, which is very close to the true value 0.9, while the average of the DC estimator is 0.884. In addition, the DC estimator produces several extremely small estimates, which lead to a much larger standard deviation.

3.4.3 Continuous Distribution Estimation

For density estimation, we compare the direct deconvolution estimator, the standard sieve estimator and the penalized sieve (PS) estimator with $\lambda = 100$ and 10,000 in Figure 3.2.

For our sieve estimators, we use $r = 10$ lattice points, and the first order squared difference as the smoothness penalty. The hybrid method described in Section 3.2.4 is used to derive the PS estimators: First we estimate the pointmass by the standard sieve method (i.e. $\lambda = 0$), and then estimate $\theta$ by the maximizer of the penalized log-likelihood. Since the target density $f_c$ is continuous, we display the interpolated form of the estimators as in (3.5) to make the comparison easier.

In each panel of Figure 3.2, a density-envelope, formed by a bundle of light gray curves, is superimposed to represent the natural variation of the various estimators. Each light gray curve is an individual estimator obtained for one of the 100 simulated data sets. The black
Figure 3.2: Comparison of the density estimation between (a) the deconvolution estimator, (b) the standard sieve estimator, (c) the penalized sieve estimator with $\lambda = 100$ and (d) $\lambda = 10,000$. In each panel, the black dashed curve shows the true density, each light gray curve shows an individual density estimator, and the dark gray solid curve is an average of the 100 estimators.

dashed curve and the dark gray solid curve show the true density and the average of the 100 individual estimators, respectively.

In Figure 3.2(a) we can see the serious boundary problem of the DC estimator near $x = 0$, where it is seriously biased and it puts some positive probability mass even on the negative half-line. Moreover, the density estimator takes negative values, which contradicts the basic nonnegative property of probability density functions. In obtaining the DC estimator, we used a suitable integration range parameter as suggested in Chapter 2.

As a comparison, the standard sieve (SS) estimator in Figure 3.2(b) is almost unbiased. However, the estimation variance is large, as shown by the density envelope. As pointed out in Section 3.2.4, the SS estimator is under-smoothed as it corresponds to the penalized sieve
(PS) estimator with $\lambda = 0$. On the other hand, Figure 3.2(d) shows the PS estimator with $\lambda = 10,000$, which is clearly over-smoothed: the bias of the estimator is too large, but the estimation variance is very small.

Figure 3.2(c) plots the PS estimator for $\lambda = 100$, which performs the best among the various estimators. It appropriately addresses the boundary effects exhibited by the DC estimator. In addition, the individual estimates are very smooth, and the estimation variance is much smaller than the standard sieve estimator. Furthermore, the average estimator is very close to the true density curve, which means the bias is small. The optimal value of $\lambda$ is chosen by minimizing the integrated mean square error of the estimation, which is discussed later in Section 3.5.2.

Comparing Figures 3.2(b)–(d), we can study the effect of the penalty parameter $\lambda$ in the sieve estimation. As $\lambda$ increases, the estimator gets smoother, so a larger $\lambda$ is preferred when the target density is known to be smooth. In addition, the estimation bias gets larger, while the variance gets smaller as $\lambda$ increases, as illustrated later in Figure 3.5.

3.5 Application to the Virus-lineage Data

In this section, we apply the proposed estimators to the virus-lineage data analyzed by Burch et al. (2007). Our primary goal is to estimate the distribution of the mutation effect on fitness. In addition, we use the density-envelope plot to provide a graphical validation of the exponential assumption on the mutation effect distribution. Such an assumption is widely made in the evolutionary biology literature without much investigation into its validity. Details on the data set can be found in Section 2.5.

3.5.1 Estimation Results

We now apply our estimators to the virus data. As for the pointmass, the standard sieve estimator suggests that the proportion of no or silent mutations is $\hat{p}_1^{SS} = 0.9095$. This estimate is similar to the pointmass estimate given by Burch et al. (2007), 0.9027, both of which are smaller than 0.9363, the direct deconvolution (DC) estimate given in Chapter 2.
As for the estimation of the continuous component, Figure 3.3 shows the various density estimators of deleterious mutation effects. We compare our estimator with the classical parametric exponential fit, and the direct deconvolution estimator proposed in Chapter 2. For the exponential fit, the method of moments is used to estimate the mean parameter. Note that the direct deconvolution estimator (the gray dash-dotted curve) puts positive density on the negative real line, which shows the seriousness of the boundary problem. As a comparison, the standard sieve estimator (the gray solid curve) has the correct support but is rather noisy, while the penalized sieve estimator (the black solid curve) with $\lambda = 50$ works very well. In addition, the penalized sieve estimator is much smoother than the standard sieve estimator, and visually similar to the fitted exponential density (the black dash-dotted curve). The value $\lambda = 50$ is chosen as discussed in Section 3.5.2. In deriving the sieve estimators, we consider $r = 10$ lattice points, and use the first order squared difference as the penalty function.

![Figure 3.3: The virus-lineage application. This plot shows various density estimators of the deleterious mutation effects.](image)

### 3.5.2 Validation of the Exponential Assumption

One more statistical issue we want to address is the validation of the exponential fit to the mutation effect distribution. Exponential distributions are currently used to fit the mutation effect distribution, but no serious work has been done to validate this parametric assumption. Figure 3.3 shows that the entire trend of the nonparametric penalized sieve estimator is similar...
to that of an exponential distribution, but this is not enough to validate the exponential assumption.

![Graphs showing density envelopes for different values of λ](image)

**Figure 3.4:** Validation of the exponential assumption: the density-envelope plots for various values of the roughness penalty parameter $\lambda$. The light gray curves form the envelope to show the natural estimation variation, the black dash-dotted curve shows the exponential density which generates the simulation samples, the black solid curve is the penalized sieve density estimator obtained from the data, and the dark gray solid curve is the average of the 100 gray curves.

To formally check the validity of the exponential assumption, instead of those numerical measures, we propose to use the *density-envelope plot*, shown in Figure 3.4. To obtain the envelope, we first fit an exponential distribution to our data, then generate 100 random samples from the mixture of a pointmass at 0, and the fitted exponential distribution. The simulated samples are of the same size as our original data. We then obtain the penalized sieve density estimators for each sample. Finally these 100 estimators are plotted to form the envelope. In addition to the envelope, the black dash-dotted curve shows the exponential
density which generates the simulation samples, the black solid curve is the PS estimator obtained from the data, and the dark gray solid curve is the average of the 100 light gray curves.

The underlying idea of the density-envelope plot is that the samples generated from the same population may give similar distribution estimators, when the same estimation technique is used. We expect that the density estimates obtained from the data would close to the density estimates based on the simulated samples, if the exponential assumption is correct. The width of the formed envelope corresponds to the natural variation of the estimator, so it can be understood as an acceptable range of the density estimates under the assumption that the true distribution is the fitted exponential.

According to Figure 3.4, the PS estimator obtained from the data lies in the density envelope in each panel. In particular, in Figures 3.4(c) and (d), the PS estimator almost overlaps the average estimator. This implies that the difference between the PS estimator and the fitted exponential density can be explained as natural variation. Equivalently, we can state that an exponential distribution is a reasonable model for deleterious mutation effects.

The density-envelope plot also provides a tool for selecting the penalty parameter $\lambda$. From these 100 envelope curves and the assumed true (exponential) distribution, we can compute the empirical bias and variance of the penalized sieve estimator for each $\lambda$. Figure 3.5 (a) plots the integrated squared bias, variance and mean squared error (MSE) of the PS estimator for a range of $\lambda$ on the log scale. The integrated MSE is minimized when $\lambda = 50$; hence we choose this value to obtain our final estimator, which is plotted in Figure 3.5(b), along with several other estimators.

### 3.6 Proofs and Technical Details

In this section, we provide technical proofs for the theorems in Section 3.3, and the detailed implementation of the iterative algorithm in 3.2.3.
Figure 3.5: Bias and variance tradeoff for the PS estimator. In panel (a), the horizontal axis shows the value of $\lambda$ on the log-scale. As $\lambda$ increases, the estimation bias increases and the variance decreases. In addition, the MSE is minimized when $\lambda = 50$. Panel (b) shows the penalized density estimate with $\lambda = 50$, along with the corresponding density envelope plot.

3.6.1 Proof of the Theorems

Proof of Theorem 3.3.1  Let $m = m(n) \geq 2$ be an integer. We shall choose $m$ to diverge to infinity as $n$ increases, and such that

$$\frac{r}{m} \to 0. \tag{3.12}$$

Details will be given below (3.17). Consider a lattice of values of $\phi_1, \ldots, \phi_r$, where each $\phi_j$ is expressed as $m_j/m$, each $m_j$ is a nonnegative integer, and $m_1 + \cdots + m_r = m$. And suppose that each $\phi_j \leq C_1 h$, where $C_1$ is as in Assumption (R4). Then $\phi_j \leq C_2/r$ for some constant $C_2$ from (R1). Therefore, these $\phi_j$s can assume at most $C_3 m/r$ different values for some $C_3 \geq 1$.

In addition, consider $q = (q_1, \ldots, q_{\nu+1})^T$, where each $q_l = n_l/[m/r]$, where $[x]$ denotes the integer part of $x$, $n_l$ is a nonnegative integer, and $n_1 + \cdots + n_{\nu+1} = [m/r]$. Then each $q_l$ can have at most $m/r$ different values. Write $Q$ for the class of all $\tau$’s which have the form $\tau = (q_1, \ldots, q_{\nu+1})\phi^T$ arising in this way. Then

$$\text{card}(Q) \leq \left( \frac{C_3 m}{r} \right)^{r+\nu-1}. \tag{3.13}$$
Define \( d(y|\tau) = \log \left\{ \sum_{l=1}^{r} q_l f_Z(y - a_l) + q_{r+1} \sum_{j=1}^{r} \phi_j f_Z(y - x_j) \right\} \), and let \( g \) be as defined in (3.8). Note that if \( x_0 \in [c, d] \) then, for all \( y \),

\[
\frac{1}{g(y)} \leq \frac{\exp\{d(y|\tau)\}}{f_Z(y - x_0)} \leq g(y).
\]

Therefore, \(|d(y|\tau)| \leq 2 \max\{\log g(y), |\log f_Z(y - x_0)|\} \}, from which it follows that

\[
|d(y|\tau)|^s \leq 2^s \left\{ |\log g(y)|^s + |\log f_Z(y - x_0)|^s \right\}.
\] (3.14)

This result, and the assumptions (3.9) and (3.10), imply that there exists a constant \( B_1 > 0 \) such that

for all integers \( s \geq 3 \), \( E\{|d(Y_i|\tau)|^s\} \leq \frac{s!}{2} \text{Var}\{d(Y|\tau)\}^2 B_1^{s-2} \), \( (3.15) \)

where \( B_1 \) does not depend on \( \tau \) or \( s \). In view of (3.15), Bernstein-type bounds (Theorem 1.1 of De La Peña (1999)) imply that if \( \kappa^2 \) denotes an upper bound to \( \text{Var}\{d(Y|\tau)\} \} \) for all \( \tau \). If \( 0 < \eta < \kappa^2/B_1 \), then

\[
P\left[ \left| \frac{1}{n} \sum_{i=1}^{n} \{d(Y_i|\tau) - Ed(Y_i|\tau)\} \right| > \eta \right] \leq 2 \exp\left( -\frac{n \eta^2}{2 \text{Var}\{d(Y|\tau)\} + B_1 \eta} \right)
\]

\[
\leq 2 \exp\left( -\frac{n \eta^2}{4\kappa^2} \right). \] (3.16)

Note that Condition (3.14) implies that the upper bound \( \kappa^2 \) exists and is finite.

Defining \( B_2 = (4\kappa^2)^{-1} \) and

\[
\delta(\tau) = \frac{1}{n} \sum_{i=1}^{n} \{d(Y_i|\tau) - Ed(Y_i|\tau)\} ,
\]
we deduce from (3.13) and (3.16) that
\[
P\left\{ \sup_{\tau \in Q} |\delta(\tau)| > \eta \right\} \leq \sum_{\tau \in Q} P\{ |\delta(\tau)| > \eta \}
\leq \text{card}(Q) \cdot \sup_{\tau \in Q} P\{ |\delta(\tau)| > \eta \}
\leq 2 \left( \frac{C_3 m}{r} \right)^{r+\nu-1} \exp \left( -B_2 n \eta^2 \right). \quad (3.17)
\]

Since \( r = o(n) \) by (R4) and \( \nu \) is finite, we may write \( \xi = \xi(n) = n/(r+\nu-1) \), which diverges to infinity as \( n \) goes to infinity. It suffices to treat the case where \( r \) is relatively large, and indeed we may assume without loss of generality that \( \xi \leq \log n \). Let \( m \) equal the integer part of \( rC_3^{-1} \exp(B_2 \xi^{1/2}/2) \) and \( \eta = \xi^{-1/4} \). Then \( r/m \to 0 \) and
\[
\log \left( \left( \frac{C_3 m}{r} \right)^{r+\nu-1} \exp \left( -B_2 n \eta^2 \right) \right) \leq \left( \frac{r+\nu-1}{2} \right) B_2 \xi^{1/2} - B_2 n \xi^{-1/2}
= -\frac{B_2 n}{2} \xi^{-1/2}.
\]
Hence, by (3.17),
\[
P\left\{ \sup_{\tau \in Q} |\delta(\tau)| > \eta \right\} \leq 2 \exp \left( -\frac{B_2 n}{2} \xi^{-1/2} \right),
\]
which converges to zero faster than any power of \( n^{-1} \). Hence, by the Borel-Cantelli Lemma,
\[
\sup_{\tau \in Q} |\delta(\tau)| \to 0 \quad \text{with probability 1.} \quad (3.18)
\]

Let \( \mathcal{P} \) be the class of all \( \pi = (p_1, \ldots, p_\nu, p_{\nu+1} \theta^T)^T \), where each \( p_i \geq 0, \theta_j \geq 0, \) and \( \sum_l p_l = \sum \theta_j = 1 \). In fact, \( \mathcal{P} \) is the class of all \( (r+\nu) \)-variate discrete distributions. In addition, each \( \theta_j \leq C_1 h \), and given \( \pi \in \mathcal{P} \), let \( \tau = \tau(\pi) \) be the best approximation of \( \tau \) in the sense that it minimizes the \( L_1 \) distance between \( \pi \) and \( \tau \), i.e. \( ||\pi - \tau||_1 = \sum_{l=1}^{\nu} |p_l - q_l| + \sum_{j=1}^{r} |p_{\nu+1} \theta_j - q_{\nu+1} \phi_j| \) over all \( \tau \in \mathcal{Q} \). By construction of \( \mathcal{Q} \),
\[
||\pi - \tau||_1 \leq \frac{r(\nu + 2)}{m}, \quad (3.19)
\]

56
for sufficient large $n$ from (3.12). Define $D(y|\tau) = \sum_{l=1}^{\nu} q_l f_Z(y-a_l) + q_{\nu+1} \sum_j \phi_j f_Z(y-x_j)$ and define $D(y|\pi)$ in the same way; and put

$$D(y|\tau, \pi) = \frac{|D(y|\tau) - D(y|\pi)|}{\min\{D(y|\tau), D(y|\pi)\}} \geq 0.$$  

Property (3.19) implies that

$$|D(y|\tau) - D(y|\pi)| \leq \frac{r(\nu + 2)}{m} \cdot \sup_{c \leq u \leq d} f_Z(y-u),$$

from which it follows that $D(y|\tau, \pi) \leq r(\nu+2)g(y)/m$. Hence, if $\delta \in (0,1]$ and $r < m/(\nu+2)$,

$$| \log D(y|\tau) - \log D(y|\pi) | = \log \{1 + D(y|\tau, \pi)\} \leq \log \left\{ 1 + \frac{r(\nu + 2)g(y)}{m} \right\} \leq \delta + \log \{1 + g(y)\} \left\{ \frac{r(\nu + 2)g(y)}{m} > \delta \right\} \leq \delta + \{\log 2 + \log g(y)\} I \left\{ \frac{r(\nu + 2)g(y)}{m} > \delta \right\},$$

where the last inequality follows from the fact that $g(\cdot) \geq 1$. Therefore,

$$\sup_{\pi \in \mathcal{P}} \left| \frac{1}{n} \sum_{i=1}^{n} \{d(Y_i|\tau) - d(Y_i|\pi)\} \right| = \left| \frac{1}{n} \sum_{i=1}^{n} \{ \log D(Y_i|\tau) - \log D(Y_i|\pi) \} \right| \leq \delta + \frac{1}{n} \sum_{i=1}^{n} \{ \log 2 + \log g(Y_i) \} I \left\{ g(Y_i) > \frac{m\delta}{r(\nu + 2)} \right\}. \quad (3.20)$$

In view of (3.12), the right-hand side of (3.20) converges to $\delta$, with probability 1, as $n \to \infty$. Since this is true for each $\delta > 0$ then the left-hand side converges to zero:

$$\sup_{\pi \in \mathcal{P}} \left| \frac{1}{n} \sum_{i=1}^{n} \{d(Y_i|\tau) - d(Y_i|\pi)\} \right| \to 0 \ \text{with probability 1.} \quad (3.21)$$

57
A similar argument shows that, as $n \to \infty$,
\[
\sup_{\pi \in \mathcal{P}} \left| E\{d(Y_i|\pi) - d(Y_i|\pi)\} \right| \to 0. \tag{3.22}
\]

Results (3.21), (3.22) and (3.18) imply that
\[
\sup_{\pi \in \mathcal{P}} |\delta(\pi)| \to 0 \text{ with probability 1}. \tag{3.23}
\]

Let $F_X$ denote the true distribution of $X$. For any $p \in \mathcal{P}$, let $f_Y(y|\pi) = \sum_{l=1}^{\nu} p_l f_Z(y - a_l) + p_{\nu+1} \sum_{j=1}^{r} \theta_j f_Z(y - x_j)$ denote the approximation to the density $f_Y$ of $Y$, which is obtained by approximating $F_X$ by the distribution with atoms $a_1, \ldots, a_{\nu}, x_1, \ldots, x_r$ having probability $p_1, \ldots, p_{\nu}, p_{\nu+1} \theta_1, \ldots, p_{\nu+1} \theta_r$, respectively. Define the negative entropy, $e_Y$, of the distribution of $Y$ to be $e_Y = \int (\log f_Y) f_Y$. Now,
\[
E\{d(Y|\pi)\} = \int \log \{f_Y(y|\pi)\} f_Y(y) dy,
\]
and the Kullback-Leibler divergence of the distribution with density $f_Y(\cdot|\pi)$ from that with density $f_Y$ is given by
\[
d\{f_Y, f_Y(\cdot|\pi)\} = \int \log \left\{ \frac{f_Y}{f_Y(\cdot|\pi)} \right\} f_Y \geq 0.
\]
Hence, $e_Y - E\{d(Y|\pi)\} = d\{f_Y, f_Y(\cdot|\pi)\}$, and so by (3.23), and with probability 1,
\[
e_Y - \frac{1}{n} \sum_{i=1}^{n} d(Y_i|\pi) = d\{f_Y, f_Y(\cdot|\pi)\} + o(1), \text{ uniformly in } \pi \in \mathcal{P}. \tag{3.24}
\]

Let $\hat{\pi}$ denote a random element of $\mathcal{P}$ which gives a global maximum of $\sum_i d(Y_i|\pi)$ over $\pi \in \mathcal{P}$. Knowing the distribution $f_X$ it is straightforward to construct a vector $\hat{\pi} = (\tilde{p}_1, \ldots, \tilde{p}_{\nu}, \tilde{p}_{\nu+1}\tilde{\theta}_1, \ldots, \tilde{p}_{\nu+1}\tilde{\theta}_r)^T \in \mathcal{P}$ for which
\[
d\{f_Y, f_Y(\cdot|\hat{\pi})\} \to d(f_Y, f_Y) = 0. \tag{3.25}
\]
One can simply choose $\tilde{p}_l = P(X = a_l)$ for $1 \leq l \leq \nu$, $\tilde{p}_{\nu+1} = 1 - \sum_{l=1}^{\nu} \tilde{p}_l$, and $\hat{\theta}_j$ to equal the probability that $X_c \in [(j-1)h, jh)$ for $j \geq 1$. By definition of $\hat{\pi}$, $\sum_i d(Y_i|\pi) \geq \sum_i d(Y_i|\tilde{\pi})$.

Therefore, by (3.24) and (3.25), and with probability 1,

$$d\{f_Y, f_Y(\cdot|\tilde{\pi})\} + o(1) = e_Y - \frac{1}{n} \sum_{i=1}^{n} d(Y_i|\tilde{\pi})$$

$$\leq e_Y - \frac{1}{n} \sum_{i=1}^{n} d(Y_i|\tilde{\pi})$$

$$= d\{f_Y, f_Y(\cdot|\tilde{\pi})\} + o(1) \rightarrow 0. \quad (3.26)$$

However $d\{f_Y, f_Y(\cdot|\tilde{\pi})\}$, being a Kullback-Leibler divergence, is nonnegative, and therefore (3.26) implies that $d\{f_Y, f_Y(\cdot|\tilde{\pi})\} \rightarrow 0$ with probability 1. Hence, with probability 1, the distribution with density $f_Y(\cdot|\tilde{\pi})$ converges, almost surely, to the distribution with density $f_Y$. The theorem then follows from this result and the assumption (R6).

**Proof of Theorem 3.3.2** The proof is very similar with the proof of Theorem 3.3.1. Define $\hat{\pi}^{PS}$ as a global maximum of the penalized log-likelihood $\sum_i d(Y_i|\pi) - \lambda P(\pi)$ over $\pi \in \mathcal{P}$.

By the definition of $\hat{\pi}^{PS}$, it is clear that

$$\sum_{i=1}^{n} d(Y_i|\pi^{PS}) \geq \sum_{i=1}^{n} d(Y_i|\tilde{\pi}) + \lambda \left\{ P(\hat{\pi}^{PS}) - P(\tilde{\pi}) \right\},$$

where $\hat{\pi}$ is chosen in the same way as in the previous proof. Then, from (3.24) and (3.25), and with probability 1,

$$d\{f_Y, f_Y(\cdot|\pi^{PS})\} + o(1) = e_Y - \frac{1}{n} \sum_{i=1}^{n} d(Y_i|\pi^{PS})$$

$$\leq e_Y - \frac{1}{n} \sum_{i=1}^{n} d(Y_i|\tilde{\pi}) - \frac{\lambda}{n} \left\{ P(\hat{\pi}^{PS}) - P(\tilde{\pi}) \right\}$$

$$= d\{f_Y, f_Y(\cdot|\tilde{\pi})\} + o(1) + O\left(\frac{1}{n}\right) \rightarrow 0, \quad (3.27)$$

since $P(\cdot)$ is asymptotically bounded and $\lambda = o(n)$. Because $d\{f_Y, f_Y(\cdot|\pi^{PS})\}$ is nonnegative, it converges to 0 with probability 1 by (3.27). Hence the distribution with density $f_Y(\cdot|\pi^{PS})$
converges to the true distribution $f_Y$ almost everywhere, with probability 1. Hence, from the fact (R6), we conclude that the distribution estimator of $X$ characterized by $\pi^{PS}$ converges to the true distribution with probability 1.

\[\square\]

### 3.6.2 Optimization details in the Sieve ML Estimation

Below we provide some technical details for the optimization algorithm involved in the sieve estimation. For additional details, see Bertsekas (2005).

For the nonlinear constrained optimization, the most popular method uses Lagrange multipliers. Suppose that we want to maximize (or minimize) some function defined on the $r$-dimensional space with $k$ constraints. By introducing a new unknown variable, called the Lagrange multiplier, for each constraint, this method replaces the problem by one of maximizing (or minimizing) an unconstrained function, the Lagrangian, in $r + k$ variables.

The Karush-Kuhn-Tucker (KKT) condition is simply the generalization of the Lagrange multiplier theorem. Compared to the fact that the Lagrange multiplier theorem contains only the equality constraints, the KKT condition provides the analogous conditions in the optimization problem with both the equality and inequality constraints.

Consider the following problem:

\[
\text{minimize } \left\{ f(\theta) : \theta \in \mathbb{R}^n \right\} \quad \text{subject to } \quad h(\theta) = 0, \quad g(\theta) \leq 0 \quad (3.28)
\]

with a strictly convex function $f(\cdot)$, a $p$-dimensional vector $h(\cdot)$ and an $m$-dimensional vector $g(\cdot)$. From the KKT condition, the necessary and sufficient conditions for a unique minimum point, say $\theta^*$, are that there exist multipliers $\lambda^* \in \mathbb{R}^p$ and $s^* \in \mathbb{R}^m$ such that

\[
\nabla f(\theta^*) + \sum_{j=1}^{p} \lambda^*_j \nabla h_j(\theta^*) + \sum_{j=1}^{m} s^*_j \nabla g_j(\theta^*) = 0,
\]
where
\[
  h(\theta^*) = (0, 0, \ldots, 0)^T, \quad g_j(\theta^*) s^*_j = 0 \quad \text{for any } j,
\]
\[
  g(\theta^*) \leq (0, 0, \ldots, 0)^T \quad \text{and} \quad s^* \geq 0.
\]

Here, \(\nabla\) represents the derivative of a function.

In our case, for example Step 2(a) in the algorithm of Section 3.2.3, \(f(\theta)\) is taken to be the conditional negative log-likelihood of \(\theta\), i.e. \(f(\theta) = -l(\theta|p) = -l(p, \theta)\) for some known \(p\), and \(l\) is the log-likelihood in (3.4). According to the form of the log-likelihood, \(f(\theta)\) can be expressed as \(f(\theta) = -\sum_{i=1}^{n} \log(a_i + c_i^T \theta)\), where \(n\) is the sample size, and \(a_i\) and \(c_i\) are sample based constants such as

\[
  a_i = \sum_{l=1}^{\nu} p_l f_Z(Y_i - a_l) \\
  \text{and} \quad c_i = (c_i1, \ldots, c_ir)^T = p_{\nu+1}(f_Z(Y_i - u_1), \ldots, f_Z(Y_i - u_r))^T.
\]

So \(f\) is a strictly convex function. In addition, \(h(\theta) = \sum_{j=1}^{r} \theta_j - 1 = e^T \theta - 1\), where \(e\) is a \(r\)-dimensional vector with entries 1, and \(g(\theta) = -\theta\).

The standard approach for solving (3.28) is to consider the following penalized problem:

\[
\begin{align*}
\text{minimize} & \{ f(\theta) - \mu \sum_{j=1}^{r} \log \theta_j : \theta \in \mathbb{R}^r \} \\
\text{subject to} & \quad h(\theta) = 0, \quad \text{and} \quad g(\theta) \leq 0,
\end{align*}
\]

for small positive \(\mu\). Note that \(\theta_j\) should be greater than zero for this problem to make sense. Then the necessary and sufficient conditions for this problem are

\[
-\sum_{i=1}^{n} \frac{c_i}{a_i + c_i^T \theta} + \lambda e - \mu \theta^{-1} = 0, \quad \text{and} \quad e^T \theta = 1,
\]  

(3.29)

where \(\theta^{-1} = (\theta_1^{-1}, \ldots, \theta_r^{-1})\). We can get the solution of (3.29) using Newton’s iteration.
algorithm. If the solutions are $\theta(\mu)$ and $\lambda(\mu)$, then by letting $\mu \to 0$, we can obtain

$$
\theta^* = \lim_{\mu \to 0} \theta(\mu), \quad \text{and} \quad \lambda^* = \lim_{\mu \to 0} \lambda(\mu).
$$

This $\theta^*$ corresponds to the standard sieve estimation $\hat{\theta}^{SS}$ in Section 3.2.3. Similar results can be derived for the penalized sieve estimator in Section 3.2.4.
Chapter 4

Sieve Type Deconvolution Based on Least Squares

4.1 Introduction

Sieve type estimators of a generalized density, which are based on the maximum likelihood idea, are proposed in Chapter 3 under the model described in (1.1) and (1.2). In this chapter, followed to that, we propose alternative sieve type estimation algorithm. We consider the same mixture target distribution with bounded support, and the proposed method uses the same type of sieve with the estimators in Chapter 3. The main difference is that the least squares (LS) idea is used instead of ML in the actual estimation step.

Ruppert et al. (2007) is relevant to our study in a sense that they combine a sieve type density estimator with LS idea, and consider the target distribution is a kind of mixture distribution with the boundary problems. However, they only consider mixtures of a uniform distribution and other continuous distributions. In addition, they do not incorporate measurement errors.

Previously, Mendelsohn and Rice (1982) also approach the deconvolution problem via LS: they first model the target density using a linear combination of B-splines, and estimate the spline coefficients by LS. Our proposal is different from their method in the following two points:

- pointmasses are used as basis functions to expand (continuous) density functions;
we define a *distance* between two probability distributions in terms of distribution functions (or alternatively characteristic functions), instead of density functions.

In many contexts, LS methods require fewer assumptions, and are faster to compute than ML methods. These properties are also found in our estimators. Compared to the ML-based estimator in Chapter 3, our proposal requires shorter computation time, which is explored in Section 4.4 along with other small sample properties via simulation studies. In addition, Section 4.3 shows that the strong consistency of the proposed estimator can be established under weaker conditions than assumptions for the ML-based estimator.

Section 4.5 shows how our methods can be applied in a real data situation; we apply our methods to virus-lineage data in Burch et al. (2007), and compare with other estimators. Technical details in implementing the proposed estimation algorithms and proofs of the consistency theorems in Section 4.3 are given in Section 4.6.

### 4.2 The Proposed Estimators

Let $f_X$ be the generalized density of $X$. From the mixture structure of $X$ given in (1.2), $f_X$ can be expressed as

$$f_X(x) = \sum_{j=1}^{\nu} p_j \delta_{a_j}(x) + p_{\nu+1} f_c(x). \tag{4.1}$$

where $f_c$ is the density of $X_c$, and $a_1, \ldots, a_\nu$ are known. Hence the estimation of $f_X$ is equivalent to the estimation of both a density $f_c$ and a finite dimensional parameter $p = (p_1, \ldots, p_{\nu+1})^T$. Since $p$ determines a probability distribution, $p_j$ are nonnegative, and sum up to 1.

We first approximate $X_c$ by a discrete random variable $\tilde{X}_c$ which has values $x_1, \ldots, x_r$. In practice, $x_j$s are chosen to be equally spaced, i.e. $x_{j+1} - x_j = h$ for some $h > 0$, and $\tilde{X}_c$ is constructed to satisfy

$$\tilde{X}_c = x_j \quad \text{if and only if} \quad X_c \in [x_j - 0.5h, x_j + 0.5h).$$

In this paper, we focus on the case that $f_c$ is supported on a finite interval $[a, b]$, and has jumps
at its boundaries. This boundary information can be easily reflected by choosing $x_j$s only in the known support of $f_c$. For example, when $X_c$ is a nonnegative random variable, we can choose $x_j = (j - 0.5)h$ for some $h > 0$. This $h$ plays a role similar to the smoothing parameter in kernel density estimation, and the same as the binwidth in histogram estimation.

The proposed estimation algorithm consists of three steps: to approximate the density $f_c$ by some finite dimensional parameter, and estimate the parameter. The last step is the reconstruction of the density function form the parameter estimates.

Let $\theta = (\theta_1, \ldots, \theta_r)^T$ be the probability distribution of $\tilde{X}_c$, i.e.

$$\theta_j = P(\tilde{X}_c = x_j) \quad \text{for each } j = 1, \ldots, r,$$

where $\theta_j \geq 0$ and $\sum \theta_j = 1$. Roughly, $\theta_j$ can be understood as the approximated probability that $X_c$ lies in the interval $[x_j - 0.5h, x_j + 0.5h)$, from the construction of $\tilde{X}_c$. By replacing $X_c$ by $\tilde{X}_c$, the generalized density $f_X$ is approximated by a linear combination of atoms such as

$$\tilde{f}_X(x|\hat{p}, \hat{\theta}) = \nu \sum_{l=1}^{\nu} \hat{p}_l \delta_{a_l}(x) + \hat{p}_{\nu+1} \sum_{j=1}^r \hat{\theta}_j \delta_{x_j}(x).$$

Then our estimation problem turns into the estimation of finite dimensional parameters $\theta$ and $p$.

The next step is the estimation of $\theta$ and $p$. We used maximum likelihood (ML) to estimate these parameters in Chapter 3. Here, we use the least squares (LS) ideas, which improves the ML method. Details on the estimation are given below Sections 4.2.1 and 4.2.2.

After obtaining estimators $\hat{p}$ and $\hat{\theta}$, we get a naive estimator of $f_X(x)$ as

$$\tilde{f}_X(x|\hat{p}, \hat{\theta}) = \nu \sum_{l=1}^{\nu} \hat{p}_l \delta_{a_l}(x) + \hat{p}_{\nu+1} \tilde{f}_c(x|\hat{\theta}).$$

It can be improved by using a linear interpolation such as

$$\hat{f}_X(x|\hat{p}, \hat{\theta}) = \nu \sum_{l=1}^{\nu} \hat{p}_l \delta_{a_l}(x) + \hat{p}_{\nu+1} \hat{f}_c(x|\hat{\theta}),$$

(4.2)
where

\[
\hat{f}_c(x|\hat{\theta}) = \begin{cases} 
\frac{\hat{\theta}_{j-1}}{h} + \frac{\hat{\theta}_j - \hat{\theta}_{j-1}}{h(x_j - x_{j-1})}(x - x_{j-1}), & \text{for } x \in [x_{j-1}, x_j), \ 1 \leq j \leq r+1, \\
0, & \text{otherwise}, 
\end{cases}
\]  

(4.3)

which is a linear interpolation of \((x_j, \hat{\theta}_j)s\). To get a proper density function, we use \(x_0 = x_1 - 0.5h, \ x_{r+1} = x_r + 0.5h, \ \hat{\theta}_0 = \hat{\theta}_1 \text{ and } \hat{\theta}_{r+1} = \hat{\theta}_r\). The estimator in (4.3) is attractive especially when \(f_c\) is known to be continuous because it gives a continuous density estimator.

### 4.2.1 LS on Cumulative Distribution Functions

To estimate \(p\) and \(\theta\), first we use cumulative distribution functions (cdf). A naive idea is to minimize the distance between the true distribution of \(Y\) and the approximated distribution function, which has the form

\[
F^{SE}_Y(y|p, \theta) = \sum_{l=1}^{\nu} \nu l F_Z(y - a_l) + p_{\nu+1} \sum_{j=1}^{r} \theta_j F_Z(y - x_j),
\]  

(4.4)

where \(F_Z\) is a distribution function of \(Z\). This approach is reasonable, but the problem is the true distribution of \(Y\) is unknown. As an alternative, we use the empirical distribution function. A justification is that the empirical distribution function converges to the true distribution as the sample size goes to infinity.

Hence we estimate \(p\) and \(\theta\) by minimizing the distance between two distribution functions, i.e.

\[
(\hat{p}, \hat{\theta}) = \arg\min_{p, \theta} S(p, \theta) = \arg\min_{p, \theta} \int \left| \hat{F}_n(y) - F^{SE}_Y(y|p, \theta) \right|^2 w(y)dy, \quad (4.5)
\]

where \(\hat{F}_n(\cdot) = (1/n) \sum_k I(Y_k \leq \cdot)\) is the empirical distribution, and \(w(\cdot)\) is a nonnegative integrable weight function.

Note that \(S(\cdot, \cdot)\) is a quadratic function of both \(p\) and \(\theta\). In addition, these parameters are defined on compact subsets of Euclidean space. Hence there exists a unique minimizer of (4.5). A problem is that the minimizer does not have a closed form because of the constraints on \(p\) and \(\theta\). We use an iterative minimization to compute the unique minimizer. Details on
the iterative estimation algorithm are given in Section 4.6.2.

### 4.2.2 LS on Characteristic Functions

An alternative is to base the estimation on characteristic functions instead of distribution functions. Recall the convergence theorem (Chung, 2001) which says the convergence of the characteristic function implies the convergence of the corresponding distribution. In addition, the empirical characteristic function converges to the true function as the size of a random sample goes to infinity. Hence we expect the distance between the empirical characteristic function and the characteristic function corresponding to (4.4) might be short if the parameters are well chosen.

From that, we estimate \( p \) and \( \theta \) as follows:

\[
(\hat{p}, \hat{\theta}) = \arg \min_{p, \theta} S(p, \theta) = \arg \min_{p, \theta} \int |\hat{\varphi}_n(t) - \varphi_{SE}^{Y}(t|p, \theta)|^2 w(t) dt,
\]

(4.6)

where \( \hat{\varphi}_n(t) = (1/n) \sum_k \exp(itY_k) \) is the empirical characteristic function, and

\[
\varphi_{SE}^{Y}(t|p, \theta) = \sum_{l=1}^{\nu} p_l e^{it\alpha_l} \varphi_Z(t) + \nu \sum_{j=1}^{r} \theta_j e^{it\xi_j} \varphi_Z(t)
\]

is the characteristic function of (4.4). Here, \( \varphi_Z(t) = \int e^{itz} f_Z(z) dz \) is the known characteristic function of \( Z \), and \( w(\cdot) \) is a weight function which is nonnegative and integrable.

Similar to the cdf based estimation, \( S(\cdot, \cdot) \) is a quadratic function of \( p \) and \( \theta \), which are defined on the same compact sets; the only difference is the value of coefficients in the minimization problem. Hence we can find the minimizer using a similar iterative algorithm. Also, after obtaining the estimators of both \( p \) and \( \theta \), the distribution estimators can be established using the formula (4.3).

From now on, we will call this method LS-chf, for notational convenience. Similarly, we name the LS method based on the distribution function, discussed in Section 4.2.1, LS-cdf.
4.3 Theoretical Properties

In this section, we show that the strong consistency of the proposed distribution estimator is established under some minimal conditions.

Suppose that $f_c$, the density of $X_c$, is supported on a bounded interval $[a, b]$, and that the estimator of $F_X$ is constructed on a potentially larger interval $[a^*, b^*]$ which also cover all pointmasses $\{a_1, \ldots, a_\nu\}$. It is not required that $a$ and $b$ are known. Below, we list regularity conditions on $f_X$, $h$ and $w(\cdot)$ for the consistent estimators.

(R1) $f_c$ is supported in the interval $[a, b]$, and bounded above;

(R2) we construct the histogram approximation $\theta = (\theta_1, \ldots, \theta_r)^T$ to $f_c$ in the interval $[a^*, b^*]$, where $-\infty < a^* \leq a < b \leq b^* < \infty$, in which case $rh \leq b^* - a^*$;

(R3) $h = o(1)$ as $n \to \infty$;

(R4) the weight function $w$ is bounded, continuous, non-vanishing almost everywhere on the range of $Y$, and satisfies $\int w(t)dt = 1$;

(R5) the distribution of $X$ is uniquely identifiable from its convolution with the distribution of $Z$.

Assumptions (R1) and (R2) are the basis of our model. The key assumption is (R3); for the consistency of the distribution estimator, we only need $r \to \infty$, or equivalently the binwidth $h \to 0$ as the sample size $n$ goes to infinity. (R4) and (R5) are necessary conditions to get the unique $F_X$ in the given optimization problem.

These conditions are similar to those for ML type estimator in Chapter 3, but weaker: we only need $r \to \infty$, without additional condition $r = o(n)$. In addition, assumptions on the error distribution are disappeared. Indeed, since the empirical characteristic and distribution functions are uniformly bounded, strong consistency can be shown under very general assumptions.

Under these regularity conditions, the following Theorem 4.3.1 shows the consistency of the distribution function estimators, which are obtained by the estimation procedure described
Theorem 4.3.1. Suppose that (R1)–(R5) hold. Then, with probability 1, the general distribution estimator

$$\hat{F}_X(x|\hat{p}, \hat{\theta}) = \sum_{i=1}^{\nu} \hat{p}_i I(a_i \leq x) + \hat{p}_{\nu+1} \sum_{j=1}^{r} \hat{\theta}_j I(x_j \leq x),$$  \hspace{1cm} (4.7)

where $\hat{p} = \hat{p}^{LS}$ and $\hat{\theta} = \hat{\theta}^{LS}$ are obtained from the LS estimation (either LS-cdf or LS-chf), converges to the true distribution $F_X$ of $X$.

In Theorem 4.3.1, the weight function $w$ is fixed for all $n$. However, in practice, it is more common to consider a data-driven weight function. In the following Theorem 4.3.2, we consider a sequence of weight functions $\{w_n\}$ which is bounded by an integrable function, and converges to an nonnegative integrable weight function $w$.

Theorem 4.3.2. Suppose that (R1)–(R5) hold. Consider a sequence of weight functions $w_n(\cdot)$, which are nonnegative, integrable and satisfy,

$$\forall y, \ w_n(y) \leq w^*(y) \text{ and } w_n(y) \to w(y)$$  \hspace{1cm} (4.8)

for some integrable and bounded function $w^*$. Then, with probability 1, the distribution estimator (4.7) converges to the true distribution $F_X$ when $\hat{\theta}$ and $\hat{p}$ are obtained by

$$(\hat{p}, \hat{\theta}) = \arg\min_{p, \theta} \int |F_n(y) - F_{SE}^X(y|p, \theta)|^2 w_n(y) dy.$$  

4.4 Simulation Studies

In this section, we perform two simulation studies to investigate finite sample performance of the proposed estimators. In each simulation, $L = 100$ samples of size $n = 329$, which is the same size as the virus-lineage data analyzed in Section 4.5, are generated from the following scheme: Our target variable $X$ has a mixture distribution of a pointmass at 0 ($\nu = 1$ and
\( a_1 = 0 \), and an exponential distribution, i.e.

\[
X \sim \begin{cases} 
0, & \text{with probability } 0.9; \\
\text{Exp}(1), & \text{with probability } 0.1.
\end{cases}
\]

Note that the exponential density is supported on the positive half line, and has a jump discontinuity at the origin. The actual observation \( Y \) is determined as the sum of \( X \) and \( Z \), where \( Z \) is a noise variable generated from a population with mean 0 and standard deviation \( \sigma = 0.1 \). The distribution of \( Z \) is taken to be normal in Section 4.4.1. In Section 4.4.2, robustness of our method to a non-normal error distribution is studied. It also shows that the efficiency of the proposed methods in terms of the computing time.

### 4.4.1 Case 1: Correctly Specified Error Distribution

First, we generate the error variable \( Z \) from a normal distribution, which is the most commonly made assumption, with mean 0 and standard deviation \( \sigma = 0.1 \).

![Figure 4.1](image-url): (Case 1) LS-cdf based on the uniform weight over the range of observations. The left panel shows the kernel density estimate for the simulated pointmass estimators with the average (the solid vertical line) and the true value (the dashed vertical line). In the right panel, each gray curve is the individual density estimate, and the black thick solid curve is the average of the gray curves. For comparison, the true exponential density (the dashed curve) is displayed.

To implement the LS-cdf method, we first use a uniform weight function \( w(t) = 1_{[M_1, M_2]}(t) \), where \( M_1 = \min(Y) \) and \( M_2 = \max(Y) \). The left panel of Figure 4.1 is the scatter plot of the LS-cdf pointmass estimates. The average and the standard deviation of the 100 simulated estimates is 0.9146 and 0.0165, respectively. In the density estimation, shown in the right panel of Figure 4.1, the average of the 100 simulated density estimates are displayed as the
black solid curve. The average curve is lower than the true exponential density (the black dashed curve) near the origin, the location of the pointmass. One reason is the overestimation of the pointmass. Another possible reason is the sharp change of the true density near the origin, and the use of the linear interpolation. In fact, the average curve is not smooth enough, particularly between 0 and 2 where the true function changes quickly. This suggests the use of higher order interpolations.

In fact, we tested uniform weight functions with various values of \((M_1, M_2)\). However only the results from the case \((M_1, M_2) = (\min(Y), \max(Y))\) are displayed because of the following two reasons: First, conceptually it is natural that the integration in (4.5) covers at least the range of observations, which is the support of the empirical distribution function. In addition, simulation results are not quite different as long as the interval \([M_1, M_2]\) covers the range of \(Y\): On the other hand, when \((M_1, M_2)\) is narrower than the data range, the performances of the density estimation are much worse.

In addition to the uniform weight, we also apply a weight function \(w = F_n(1 - F_n)\), where \(F_n\) is the empirical distribution function. This \(w\) puts the most weight when \(F_n = 0.5\), and little weight on tails, and no weight outside the sample range. As shown in Figure 4.2, the most significant change is the performance of the density estimator in the tails: \(F_n\) is near 1 in the right tail, which implies almost zero weight. Hence the distance between the empirical distribution function and the modeled distribution function in the tail is not seriously counted in the estimation. This results in the poor performance of the density estimator in the right tail. On the other hand, the weight function gives large values near the origin due to the large probability mass at 0, which results in the small bias in the pointmass estimation.

Now, we move on to the LS-chf estimation, proposed in Section 4.2.2. The choice of the weight function in the LS-chf method is more challenging, because of the complexity of Fourier transformations. One fact which can be used here is that the smoothness of a function is highly related to the decaying rate of its Fourier transformation. Hence we expect the choice of \(w\) might affect the smoothness of the resulting estimator. Without any special information on the target distribution, a natural approach is to use a uniform weight \(w(t) = 1_{[-M, M]}(t)\) for some \(M\), which corresponds to no particular preference on the smoothness of the target
Figure 4.2: (Case 1) LS-cdf based on non-uniform weight function, \( w(y) = F_n(y)\{1 - F_n(y)\} \), which puts more weight on the center part of the distribution. Compared to the uniform weight results, the estimated density curve is smoother and has a thicker tail.

In the left panel of Figure 4.3, we see the estimation performance in term of the integrated bias, variance and mean squared error, as a function of \( M \). It is seen that the integrated mean squared error (IMSE) becomes stable for \( M \) large enough. The middle and the right panels of Figure 4.3 show the simulation results for \( M = 4 \), which gives the minimum IMSE. In the middle panel, the 100 simulated pointmass estimates are displayed along with the kernel density estimate; all estimates are distributed between 0.85 and 0.95 with the average 0.9081, which gives smaller bias than the LS-cdf methods. In the density envelope plot (the right panel), we can see the average estimator (the black solid curve) almost overlaps the true density (the black dashed curve).

Figure 4.3: (Case 1) LS-chf results. The left panel displays the integrated MSE of the generalized density estimator. The middle and the right panels show the histogram of the pointmass estimator and the density-envelope plot of the density estimator corresponding \( M = 4 \) which minimizes IMSE.

One thing to be noted here is that \( M \) affects more on the density estimation than the point
estimation. Three panels of Figure 4.4 show the density envelope plots for cases $M = 1, 8$ and 128, which gives some idea how $M$ affects the density estimation. First of all, $M$ controls the smoothness of the density estimator: as $M$ increases, $\hat{f}_c$ is smoother and has a heavier tail. In addition, too small $M$ gives large variance. The variance gets stabilized after some level of $M$. This role of $M$ is similar to that of the smoothness penalty parameter in (3.6).

One thing to be noted here is that $M$ has a stronger effect on the density estimation than on the point estimation. The three panels of Figure 4.4 show the density envelope plots for cases $M = 1, 8$ and 128, which gives an idea how $M$ affects the density estimation. First of all, $M$ controls the smoothness of the density estimator: as $M$ increases, $\hat{f}_c$ gets smoother and its tail gets thicker. In addition, extremely small values of $M$ give large estimation variances; the variance gets stabilized after some point. These results show that the role of $M$ looks similar to that of the smoothness penalty parameter $\lambda$ in (3.6).

4.4.2 Case 2: Misspecified Error Distribution

Here, we perform another simulation study to see the sensitivity of the proposed methods to the assumptions on measurement error distributions.

We assume that $Z$ comes from a normal distribution with mean 0 and standard deviation 0.1 in evaluating characteristic functions; however, the actual error $Z$ is generated from a uniform distribution with the same mean and standard deviation. These two distributions have the same first three moments, but the shapes of the density curves are totally different.
Other settings such as the sample size and the target distribution are exactly the same as the previous simulation. Also, uniform weight functions are used for both LS-sieve estimates.

Figure 4.5: (Case 2) LS-chf estimation. The left panel shows the integrated MSE along with integrated bias and variance. The middle and right panels show the result from $M = 4$, which is the optimal in terms of the IMSE.

Figure 4.5 shows the result of LS-chf. Compared to Figure 4.3 corresponding to normal $Z$, we can see that the IMSE (displayed in the left panel) is larger for large $M$; however, the difference is not very big. Moreover, when $M$ is relatively small, the IMSE is almost the same as that of Case 1. Especially in the optimal case ($M = 4$), shown in the middle (the pointmass estimation) and the right (the density estimation) panels, the result is quite similar to that of Case 1: in fact, the biases of both pointmass and density estimates are almost the same, but the estimation variances are a little larger.

Results of the LS-cdf are more interesting. Despite the violation of the normal assumption, the IMSE of the pointmass estimates is even smaller than the IMSE in Case 1. The IMSE of the density part is a little bigger compared to the IMSE of Case 1, but the increment is about 1.5% of the entire IMSE. More details can be found in Table 4.1. For comparison, results of the ML method in Chapter 3 are displayed as well. For ML and LS-chf, parameters are chosen based on the IMSE of $\hat{f}$, and the corresponding results are reported here. In addition, the uniform weight function is used in implementation of the LS-cdf.

In Table 4.1, we can see that LS methods give smaller IMSE than ML in both simulations, and the difference of IMSE(\hat{f}) between two simulations is also smaller. This indicates that the LS methods, especially the LS-cdf, are more robust than the ML-based method. Here, the most noticeable fact is that the LS methods require much shorter computation time, which
Table 4.1: Numerical comparison of the performances of the three methods: LS-cdf, LS-chf and ML. For both ML and LS-chf, parameters are chosen to minimize the IMSE of $\hat{f}$, and in LS-cdf, the uniform weight function is used.

<table>
<thead>
<tr>
<th></th>
<th>LS-cdf</th>
<th></th>
<th>LS-chf</th>
<th></th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 1</td>
</tr>
<tr>
<td>$\text{MSE}(\hat{p}) \times 10^3$</td>
<td>0.4854</td>
<td>0.4541</td>
<td>0.3684</td>
<td>0.4271</td>
<td>0.6134</td>
</tr>
<tr>
<td>$\text{IMSE}(\hat{f}_c)$</td>
<td>0.0738</td>
<td>0.0749</td>
<td>0.0692</td>
<td>0.0901</td>
<td>0.0774</td>
</tr>
<tr>
<td>$\text{IMSE}(\hat{f})$</td>
<td>0.0013</td>
<td>0.0012</td>
<td>0.0010</td>
<td>0.0012</td>
<td>0.0015</td>
</tr>
<tr>
<td>Time (sec/sim)</td>
<td>1.78</td>
<td>0.98</td>
<td>0.97</td>
<td>0.76</td>
<td>111.32</td>
</tr>
</tbody>
</table>

shows the efficiency of the LS methods.

4.5 Application to Virus-Lineage Data

In this section, we apply the proposed methods to the virus-lineage data (Burch et al., 2007) in estimating the distribution of mutation effects on fitness $S$. As done in Chapters 2 and 3, we first apply the proposed methods to estimate the distribution of the mutation effect on plaque sizes ($X$), and invert it to the distribution of $S$ using the relationship (2.11) and simple change of variables.

In order to reflect the nonnegativity of the deleterious mutation effects, we start with constructing an equally spaced grid of 10 bins ($r = 10$) with the length $h$ over $[0, 10h]$. The binwidth parameter $h$ is chosen so that

$$(\text{the maximum observation}) + 3\sigma \leq 10h,$$

where $\sigma$ is the standard deviation of $Z$. Then, the deleterious mutation effect is approximated as a discrete variable having values $x_j = (j - 0.5)h$, $j = 1, 2, \ldots, 10$, with corresponding probabilities $\theta_1, \ldots, \theta_{10}$, respectively.

Figure 4.6(a) shows the LS-cdf results based on the uniform weight over the data range: it gives the pointmass estimate $\hat{p}_1 = 0.8975$, which is a little smaller ($> 1\%$) than the pointmass
Figure 4.6: Panel (a) shows the LS-cdf density estimate of the mutation effects distribution with the pointmass estimate. Panel (b) shows the change of the LS-chf estimate of the pointmass probability $p_1$ as a function of the integration range $M$. Panel (c) shows some of the corresponding density estimates of $f_c$.

The density estimate $0.9095$ of the sieve-ML method in Chapter 3. The density estimate $\hat{f}_c$ is visually close to an exponential density. In fact, this $\hat{f}_c$ is close to the standard sieve ML estimate in Figure 3.3 but much smoother.

For the LS-chf method, uniform weight functions are also used. Figure 4.6(b) displays the change of the pointmass probability estimates according to the change of the range parameter $M$ in (4.6). Here, it is clearly seen that there two typical levels of estimates; $\hat{p}_1 \approx 0.915$ for small values of $M$, and $\hat{p}_1 \approx 0.836$ for large $M$. Recall the fact that in the simulation studies, the smallest IMSE of the generalized density estimates is obtained when $M$ is quite small, and gets stable when $M$ is large. From this point of view, it seems more reasonable to choose $\hat{p}_1 = 0.915$.

Figure 4.6(c) shows some of the LS-chf density estimates. Here, the density estimates from $M = 2^0, 2^1, 2^2$ are displayed because the estimates based on large $M$ are almost the same as the estimates from $M = 2^2$. When $M = 2^0$, the density estimate looks close to the LS-cdf estimate, but has a large bump near $x = 0.3$. Overall, the LS-chf tends to give rougher density estimates than the LS-cdf method.
4.6 Proofs and Technical Details

Here, we provide technical proofs of the theorems in Section 4.3. In addition, details on the implementation of the iterative estimation procedures are given.

4.6.1 Proof of Theorems

Proof of Theorem 4.3.1. Proofs in the cases of the empirical characteristic function and empirical distribution function are similar, so we shall treat only the latter.

For any \((p, \theta)\), define

\[
S(p, \theta) = \int \left\{ \hat{F}_n(y) - F^{SE}_Y(y|p, \theta) \right\}^2 w(y) dy,
\]

and

\[
I_1(p, \theta) = \int \left\{ F_Y(y) - F^{SE}_Y(y|p, \theta) \right\}^2 w(y) dy.
\]

Let \(\Pi\) be a class of all vectors \(p = (p_1, \ldots, p_{\nu+1})^T\) satisfying \(p_j \geq 0\) for all \(j\) and \(\sum p_j = 1\). And let \(\Theta\) be a class of \(r\)-dimensional vectors having the form \(\theta = (\theta_1, \ldots, \theta_r)\) with \(\theta_j \geq 0\) and \(\sum \theta_j = 1\). Define \((\hat{p}, \hat{\theta})\) be a global minimum of \(S(p, \theta)\) over \((p, \theta) \in \Pi \times \Theta\).

Note that

\[
S(p, \theta) \leq 2 \int \left\{ F^{SE}_Y(y|p, \theta) - F_Y(y) \right\}^2 w(y) dy + 2 \int \left\{ F_Y(y) - \hat{F}_n(y) \right\}^2 w(y) dy
\]

\[
= 2 I_1(p, \theta) + 2 I_2,
\]

where \(I_2 = \int \{F_Y(y) - \hat{F}_n(y)\}^2 w(y) dy\). From the integrability of \(w(\cdot)\) in (R4), and the fact

\[
\sup_y |\hat{F}_n(y) - F_Y(y)| \to 0 \quad \text{almost surely},
\]

77
by LDCT, we get

\[
0 \leq \lim_{n \to \infty} I_2 \leq \lim_{n \to \infty} 2 \int \left| \hat{F}_n(y) - F_Y(y) \right| w(y) dy = \int \lim_{n \to \infty} 2 \int \left| \hat{F}_n(y) - F_Y(y) \right| w(y) dy \\
\leq 2 \lim_{n \to \infty} \sup_y \left| \hat{F}_n(y) - F_Y(y) \right| w(y) dy = 0,
\]

(4.9)

almost surely.

In addition, it is straightforward to construct \((\hat{\mathbf{p}}, \hat{\mathbf{\theta}}) \in \Pi \times \Theta\) such that \(I_1(\hat{\mathbf{p}}, \hat{\mathbf{\theta}})\) converges to 0 as \(n \to \infty\) from (R1) and (R3). Therefore, from the definition of \((\hat{\mathbf{p}}, \hat{\mathbf{\theta}})\), it is obvious that

\[
0 \leq S(\hat{\mathbf{p}}, \hat{\mathbf{\theta}}) \leq S(\tilde{\mathbf{p}}, \tilde{\mathbf{\theta}}) \to 0.
\]

(4.10)

Hence, from (4.9) and (4.10), with probability 1,

\[
I_1(\hat{\mathbf{p}}, \hat{\mathbf{\theta}}) \leq 2 \left[ \left\{ F_Y(y) - \hat{F}_n(y) \right\}^2 + \left\{ \hat{F}_n(y) - F_{SE} Y(y | \hat{\mathbf{p}}, \hat{\mathbf{\theta}}) \right\}^2 \right] w(y) dy = 2 S(\hat{\mathbf{p}}, \hat{\mathbf{\theta}}) + 2 I_2 \to 0 \quad \text{as } n \to \infty.
\]

Since \(w\) is continuous and non-vanishing as assumed in (R4), it is bounded above zero on any finite interval. Hence convergence of \(I_1(\hat{\mathbf{p}}, \hat{\mathbf{\theta}})\) implies that the distribution estimator \(F_{SE} Y(\cdot | \hat{\mathbf{p}}, \hat{\mathbf{\theta}})\) converges to the true function \(F_Y(\cdot)\) on the range of \(Y\). Since both \(F_{SE} Y(\cdot | \hat{\mathbf{p}}, \hat{\mathbf{\theta}})\) and \(F_Y(\cdot)\) are monotonically increasing, bounded above by 1, and bounded below by 0, this results in the convergence on the whole real line. The theorem follows from this result and (R5). \(\square\)

Proof of Theorem 4.3.2. Similar to the proof of Theorem 4.3.1, for any \((\mathbf{p}, \mathbf{\theta})\), we define

\[
S_n(\mathbf{p}, \mathbf{\theta}) = \int \left\{ \hat{F}_n(y) - F_{SE} Y(y | \mathbf{p}, \mathbf{\theta}) \right\}^2 w_n(y) dy,
\]

and

\[
I_3(\mathbf{p}, \mathbf{\theta}) = \int \left\{ F_Y(y) - F_{SE} Y(y | \mathbf{p}, \mathbf{\theta}) \right\}^2 w_n(y) dy.
\]
and let \((\hat{p}, \hat{\theta})\) be a global minimum of \(S_n(p, \theta)\) on the parameter space \(\Pi \times \Theta\). Then,

\[
S_n(p, \theta) \leq 2I_3(p, \theta) + I_4
\]

and

\[
I_3(p, \theta) \leq 2S_n(p, \theta) + I_4,
\]

where \(I_4 = \int \{F_Y(y) - \hat{F}_n(y)\}^2w_n(y)dy\).

From the conditions (4.8), (R4) and boundedness of cumulative functions, by LDCT,

\[
\lim_{n \to \infty} I_4 = \int \lim_{n \to \infty} |F_n(y) - F_Y(y)|^2w(y)dy = 0.
\]

Let \(\hat{p} = (\hat{p}_1, \ldots, \hat{p}_{\nu+1})\) and \(\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_r)\) where

\[
\hat{p}_l = P(X = a_l) \quad \text{for} \quad l = 1, \ldots, \nu
\]

and

\[
\hat{\theta}_j = P(x_j - 0.5h \leq X_c \leq x_j + 0.5h) \quad \text{for} \quad j = 1, \ldots, r.
\]

Clearly, with probability 1, \(F^{SE}_Y|\hat{p}, \hat{\theta}\) converges to \(F_Y(y)\) almost everywhere from (R3). Therefore, again by LDCT, we can get the almost sure convergence of \(I_3(\hat{p}, \hat{\theta})\) to 0, which implies \(S_n(\hat{p}, \hat{\theta}) \to 0\) with probability 1. From the definition of \((\hat{p}, \hat{\theta})\), \(S_n(\hat{p}, \hat{\theta})\) also converges to 0 almost surely, hence so does \(I_3(\hat{p}, \hat{\theta})\).

To get the convergence of the distribution estimator, it is suffices to show the difference between \(I_1(\hat{p}, \hat{\theta})\) and \(I_3(\hat{p}, \hat{\theta})\) converges to 0.

\[
\sup_{p \in \Pi, \theta \in \Theta} \left| I_1(p, \theta) - I_3(p, \theta) \right| \leq \int \left| F_Y(y) - F^{SE}_Y(y|p, \theta) \right|^2 \cdot |w_n(y) - w(y)| dy
\]

\[
\leq 4 \int |w_n(y) - w(y)| dy \to 0, \quad \text{by LDCT}
\]

since \(|w_n(y) - w(y)|\) is bounded by a integrable function \(w^*(y) + w(y)\), and converges to 0 for all \(y\). Hence,

\[
\left| I_1(\hat{p}, \hat{\theta}) - I_3(\hat{p}, \hat{\theta}) \right| \leq \sup_{p \in \Pi, \theta \in \Theta} \left| I_1(p, \theta) - I_3(p, \theta) \right| \to 0,
\]

which completes the proof. \(\square\)
4.6.2 Details on Implementing the Proposed Methods

Since two methods in Section 4.2 are quite similar, we will mainly treat the LS-chf case, which is more complicate.

Suppose that $\theta$ is fixed. Then $S(p|\theta)$ has a quadratic form of $p$. In addition, the parameter space for $p$ is given by

$$\Pi = \left\{ p \in \mathbb{R}^{\nu+1} : p_l \geq 0, \text{ and } \sum_{l=1}^{\nu+1} p_l = 1 \right\},$$

which is a compact set. From these facts, we can confirm the existence and the uniqueness of the minimum of $S(p|\theta)$. The situation is the exactly same when $p$ is fixed; the differences are coefficients of the optimization problem, and the fact that $\theta$ is defined on

$$\Theta = \left\{ \theta \in \mathbb{R}^r : \theta_j \geq 0, \text{ and } \sum_{j=1}^r \theta_j = 1 \right\}.$$

Hence we suggest to use an iterative algorithm to achieve the global minimum of $S(\cdot, \cdot)$ as follows:

**Step 1.** Initialization: Set $p^{(0)}$ and $\theta^{(0)}$;

**Step 2.** Updating:

$$p^{(1)} = \arg\min_{p \in \Pi} \int |\hat{\phi}_n(t) - \varphi^SE_Y(t|p^{(0)}, \theta^{(0)})|^2 w(y)dy$$

$$= \arg\min_{p \in \Pi} p^T \int a_1(t)w(t)a_1^T(t)dt - 2 \int \left\{ b_1(t)w(t)a_1^T(t) + b_2(t)w(t)a_2^T(t) \right\} dt p$$

$$\theta^{(1)} = \arg\min_{\theta \in \Theta} \int |\hat{\phi}_n(t) - \varphi^SE_Y(t|p^{(1)}, \theta)|^2 w(y)dy$$

$$= \arg\min_{\theta \in \Theta} \theta^T \int a_3(t)w(t)a_3^T(t)dt - 2 \int \left\{ b_3(t)w(t)a_3^T(t) + b_4(t)w(t)a_4^T(t) \right\} dt \theta$$

for some functions $a_i$s and $b_i$s, and constants $c_1$ and $c_3$.

**Step 3.** Set $p^{(0)} = p^{(1)}$ and $\theta^{(0)} = \theta^{(1)}$, and repeat Step 2 until converges.
In particular, when $Z$ is symmetric about zero, i.e. the characteristic function of $Z$ is real-valued, the coefficient terms in Step 2 (a) can be explicitly given as

$$[a_1(t)]_l = \begin{cases} \varphi_Z(t) \cos(ta_l), & \text{for } 1 \leq l \leq \nu, \\ \varphi_Z(t) \sum_{j=1}^{r} \theta_j^{(0)} \cos(tx_j), & \text{for } l = \nu + 1, \end{cases}$$

$$[a_2(t)]_l = \begin{cases} \varphi_Z(t) \sin(ta_l), & \text{for } 1 \leq l \leq \nu, \\ \varphi_Z(t) \sum_{j=1}^{r} \theta_j^{(0)} \sin(tx_j), & \text{for } l = \nu + 1, \end{cases}$$

$$b_1(t) = \frac{1}{n} \sum_{k=1}^{n} \cos(tY_k) \quad \text{and} \quad b_2(t) = \frac{1}{n} \sum_{k=1}^{n} \sin(tY_k).$$

Similarly, for all $j = 1, \ldots, r$

$$[a_3(t)]_j = \varphi_Z(t)p_{\nu+1}^{(1)} \cos(tx_j), \quad [a_4(t)]_j = \varphi_Z(t)p_{\nu+1}^{(1)} \sin(tx_j),$$

$$b_3(t) = \frac{1}{n} \sum_{k=1}^{n} \cos(tY_k) - \varphi_Z(t) \sum_{l=1}^{\nu} p_l^{(1)} \cos(ta_l),$$

and

$$b_4(t) = \frac{1}{n} \sum_{k=1}^{n} \sin(tY_k) - \varphi_Z(t) \sum_{l=1}^{\nu} p_l^{(1)} \sin(ta_l).$$

In the case of LS-cdf, the algorithm is the exactly same, but has much simpler coefficient vectors; in the above equations,

$$[a_1(t)]_l = \begin{cases} F_Z(t - a_l), & \text{for } 1 \leq l \leq \nu, \\ \sum_{j=1}^{r} \theta_j^{(0)} F_Z(t - x_j), & \text{for } l = \nu + 1, \end{cases}$$

$$[a_2(t)]_l = 0 \quad \text{for all } l, \quad b_1(t) = \hat{F}_n(t), \quad \text{and} \quad b_2(t) = 0.$$

In addition, for all $j$,

$$[a_3(t)]_j = p_{\nu+1}^{(1)} F_Z(t - x_j), \quad [a_4(t)]_j = 0,$$

$$b_3(t) = \hat{F}_n(t) - \sum_{l=1}^{\nu} p_l^{(1)} F_Z(t - a_l), \quad \text{and} \quad b_4(t) = 0.$$
Future Work

Parameter Selection

In implementing the proposed estimators in practice, one of the most critical issue is the choice of the corresponding parameters: $T_n$ and $M_n$ in the direct deconvolution, $r$ (or equivalently $h$) and $\lambda$ in the sieve ML estimation, and $r$ and $M$ in the sieve LS methods.

In parameter selection, one natural approach is to minimize the (integrated) mean squared error of the estimator. However, this can not be used in real situations because the true value of $f_X$ is not known. Instead, Liu and Taylor (1989) suggested to use an upper bound of the mean squared error (UMSE). In our direct deconvolution estimation, we already obtain the upper bounds of the integrated bias and variance of the estimators in Lemmas 2.3.7 and 2.3.8, hence the UMSE-based method can be directly applied without any difficulty. Note that the actual performance of the UMSE-based method highly depends on the sharpness of the provided upper bound. An alternative is a more data-driven approach suggested by Hesse (1999); instead of the unknown true IMSE, an estimated IMSE, which is obtained by leave-one type cross validation, is used. Both methods are helpful especially in estimating $T_n$ and $M_n$ in the direct deconvolution, and $\lambda$ and $M$ for sieve estimation.

The selection of $r$, the number of bins used in approximating the continuous density in the sieve methods, can be studied in terms of the relationship between mixture models and deconvolution problems. Measurement error models are special cases of mixture models; in particular, after discretizing the density of the continuous component, the distribution of $Y$ can be easily understood as a mixture of homogeneous density functions with different location parameters. From this viewpoint, the choice of $r$ is equivalent to the choice of the number of known mixture components, hence AIC, BIC or any other statistics for model selection
can be used. More detailed discussion on model selection criteria and their properties can be found in McLachlan and Basford (1988).

More Asymptotic Properties
We showed the consistency of the proposed estimators under some suitable conditions. Consistency itself is an important property; however, convergence rates of the estimators can give more information. In this dissertation, we obtain the convergence rate only for the direct deconvolution case. Computation of the convergence rate is more challenging in the case of the sieve estimators due to the absence of the closed form of the estimators. We might need much stronger conditions on the error distribution to get the exact convergence rate.

Another interesting research area is asymptotic properties of the proposed estimator when \( \sigma \), the standard deviation of the error distribution, converges to 0 as the sample size \( n \) goes to infinity. This asymptotic scenario particularly makes sense in the repeated measures case. Delaigle (2008) gives some results for Fourier deconvolution estimators when the target variable has a continuous density under this assumption. In our case, due to the estimation of the pointmass, whose asymptotic properties are quite different from density estimation, the properties might be more complicated even in the case of direct deconvolution estimation.

Other Issues

1. Extension of the proposed method to several cases:
   Consideration of multi-dimensional target variables, or observations (or error variables) with some special correlation structure can widen the application range of the proposed method. In addition, applications of the proposed methods to the heterogeneous error distributions are also meaningful.

2. Goodness of fit test in measurement error models:
   There are several well-known methods to check whether the assumed distribution ap-
appropriately explains the data set; for example, the Kolmogorov-Smirnov test and the Anderson-Darling test. They are based on nonparametric methodology, and hence are robust. However, in measurement error models, it is not clear how to apply these methods. In addition, when the tests are applied to observations $Y_i$ s directly, they have low power, especially when the measurement error explains large proportion of the data. In the real application of this dissertation, we visually check the validity of the exponential assumption by using the density envelope plots as in Figure 3.4. Numerical testing could provide precise statistical inference relative to the graphical method.

3. Insufficient information on the error distributions and boundaries

In practice, the error distribution is rarely known, so it is often estimated. When the target variable has a continuous density, the performance of a Fourier deconvolution estimator is explored by Meister (2004) under a misspecified/unknown error distribution. When observations are measured repeatedly, like longitudinal data, the error distribution is estimable but is still not exact. In other cases, the problem is more serious. Hence, it is important to study how the proposed estimators will perform under incorrect assumptions on the error distribution, which provides the level of robustness of the estimators.

Similarly, we also assume that the support of the target distribution is known. However, in practice, sometimes we know that the boundaries exist, but not their exact locations. Meister (2006) explores this issue when the target distribution is purely continuous. If we can generalize the method to the mixture structure, the application range of the proposed methods can be extended.

4. Improvement of the direct deconvolution estimator:

One of the worst parts of the DC estimation studied in Chapter 2 is that the resulting density estimator is not a proper density. Like other Fourier deconvolution estimation methods, the resulting function has some negative values, and the integration may not be 1. Glad et al. (2003) suggest some possible solutions to construct a bona fide density function from the improper initial estimator, and study some properties of
the reconstructed estimator. It might be possible to improve the direct deconvolution estimator using these ideas.
Bibliography


