Direct Deconvolution Density Estimation of a Mixture Distribution Motivated by Mutation Effects Distribution

Mihee Lee[*], Christina Burch[†], Haipeng Shen[*], and J. S. Marron[*]

September 16, 2008

Abstract

The mutation effect distribution is essential to understand evolutionary dynamics. However, the existing studies on this problem have had limited resolution. So far, the most widely used method is to fit some parametric distribution, such as exponential, whose validity has not been checked. In this paper, we propose a nonparametric density estimator for the mutation effect distribution, based on a deconvolution approach. Consistency of the estimator is also established. Unlike the existing deconvolution estimators, we cover the case that the target variable has a mixture structure with a pointmass and a continuous component. To study the property of the proposed estimator, several simulation studies are performed. In addition, an application of modeling virus mutation effects is provided.

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

[*]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599. Email: {miwing, haipeng, marron}@email.unc.edu.

[†]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, NC 27599. Email: CBurch@bio.unc.edu

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. In this paper, we propose a nonparametric deconvolution estimator for the distribution.

Density estimation in measurement error models has been widely studied (see Carroll et al. (2006), and references within), but the existing methods only consider the case that the target variable has a continuous density function. In our motivating evolutionary study (Section 4), 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 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 paper, we focus on the case that the distribution of the target variable X is a mixture of a pointmass and a continuous distribution. Let X_d be the degenerate component of X, and let X_c be the continuous component. Then X can be represented as

$$X = \begin{cases} X_d & \text{with probability } \ p, \\ X_c & \text{with probability } \ 1-p, \end{cases} \tag{1}$$

where p is the unknown mixing probability. Suppose that we know the location of X_d, i.e. $\mathrm{P}(X_d = a) = 1$ for some known constant a. In all cases (simulations and real data analysis), $a = 0$ in this paper. Then the generalized density Cuevas and Walter (1992) of X, say f_X,

can be expressed as

$$f_X(x) = p\delta_a(x) + (1-p)f_c(x), \tag{2}$$

where δ_a denotes the Dirac delta at a, and f_c is the density of X_c.

Our interest is to estimate the above generalized density of X. According to (2), estimating the density of X is equivalent to estimating p and f_c. One problem is that X is unobservable in many cases. Instead of X itself, we can only observe the error compounded variable $Y = X + Z$, where Z is a measurement error with known density function f_Z, and it is assumed to be independent of X.

The remainder of the paper is organized as follows. In Section 2, we extend the idea of deconvolution estimation to scenarios where the target variable has a mixture distribution of a pointmass and a continuous component. Estimators for both the pointmass and the continuous density are derived. Their asymptotic properties are also provided in Section 2, with the technical proofs given in Section 5. Section 3 presents several simulation results to illustrate the performance of the estimators. In Section 4, the estimators are applied to the virus-lineage data of Burch et al. (2007).

2 The Estimators and Their Asymptotic Properties

In this section, we propose the direct deconvolution estimators of p and f_c of (2). The estimators are derived below in Sections 2.1 and 2.2 respectively, along with theorems about their asymptotic properties. Detailed proofs are provided in Section 5.

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), and Y denote the corresponding variable contaminated by the measurement error Z, i.e. $Y = X + Z$. 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. To the best of our knowledge, use of boundary information has not been studied in the context of measurement error models.

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

$$p = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \exp(-ita)\ \varphi_X(t)dt, \tag{3}$$

where φ_X is the characteristic function of X.

From (3), the pointmass p can be estimated by replacing φ_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 φ_Z is the known characteristic function of Z, and φ_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

$$\begin{aligned}\tilde{p} &= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \exp(-ita)\, \hat{\varphi}_X(t)\, dt, \\ &= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \exp(-ita) \cdot \frac{\hat{\varphi}_Y(t)}{\varphi_Z(t)}\, dt, \\ &= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \frac{1}{n} \sum_{j=1}^{n} \frac{\exp(it(Y_j - a))}{\varphi_Z(t)}\, dt. \end{aligned} \tag{4}$$

One thing to be noted is that p is a probability, and hence a real number. However the integrand of (4) 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

$$\hat{p} = \frac{1}{2nT_n} \sum_{j=1}^{n} Re \int_{-T_n}^{T_n} \frac{\exp(it(Y_j - a))}{\varphi_Z(t)}\, dt, \tag{5}$$

where *Re* denotes the real part of the complex integral.

The estimator $\hat{p}$ can be shown to be consistent as stated in Theorem 1. Below we first derive the mean and the variance of the estimator in Lemmas 1 and 2. All the proofs are given in Section 5.

LEMMA 1 *Let $\hat{p}$ be the estimator of p as defined in (5), 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*

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

where φ_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 derives the variance of $\hat{p}$. We assume that the distribution of the measurement error Z is symmetric about 0, a common assumption in measurement error models.

LEMMA 2 *Suppose that the distribution of Z is symmetric about 0. Then the variance of $\hat{p}$ is given by*

$$\begin{aligned}&\mathrm{Var}(\hat{p})\\ &= \frac{1}{2nT_n^2}\int_0^{T_n}\int_0^{T_n}\left[\frac{Re\{\varphi_V(s+t)+\varphi_V(s-t)\}-2Re\{\varphi_V(s)\}Re\{\varphi_V(t)\}}{\varphi_Z(s)\varphi_Z(t)}\right]dsdt,\end{aligned}$$

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

$$\begin{aligned}\mathrm{Var}(\hat{f}_n(x)) &= \frac{1}{n\pi^2}\int_0^{T_n}\int_0^{T_n}\left[\frac{1}{2}Re\{\varphi_V(s+t)+\varphi_V(s-t)\}-Re\{\varphi_V(s)\}Re\{\varphi_V(t)\}\right]\\ &\qquad\times\frac{\varphi_K(sh_n)\varphi_K(th_n)}{\varphi_Z(s)\varphi_Z(t)}\,dsdt.\end{aligned}$$

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 $\mathrm{Var}(\hat{p})$ converges to 0 much faster than $\mathrm{Var}(\hat{f}_n(x))$. In fact,

$$\frac{\mathrm{Var}(\hat{p})}{\mathrm{Var}(\hat{f}_n(x))} = O(T_n^{-2}), \quad \text{as} \quad n \to \infty.$$

Based on the above two lemmas, we conclude that $\hat{p}$ is consistent under some suitable conditions in Theorem 1.

THEOREM 1 *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*

$$T_n \to \infty, \quad \frac{1}{n^{1/2}\, T_n} \int_0^{T_n} \frac{1}{\varphi_Z(t)} dt \to 0 \tag{6}$$

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 1 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, $T_n = \alpha \log^{1/2} n$ for any $0 < \alpha < 1$ satisfies (6). In this case, the variance of the estimator is of the order $\log^{-1/2} n$, i.e. $\mathrm{Var}(\hat{p}) = O(\log^{-1/2} n)$ as $n \to \infty$.

2.2 Density Estimation of the Continuous Component f_c

To estimate f_c, we also use the Inverse-Fourier transformation. In particular, when φ_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 be continuous with density f_c, so its characteristic function φ_c is integrable. In addition, the mixture structure of X suggests that $\varphi_c(t)$ can be expressed as

$$\varphi_c(t) = \frac{\varphi_X(t) - p \cdot \exp(ita)}{1 - p}, \tag{7}$$

where $\varphi_X(t)$ can be estimated in the same manner as discussed above in Section 2.1. Then,

f_c can be estimated as

$$\begin{aligned}
\tilde{f}_c(x) &= \lim_{M\to\infty} \frac{1}{2\pi} \int_{-M}^{M} \hat{\varphi}_c(t) \exp(-itx) dt \\
&= \lim_{M\to\infty} \frac{1}{2\pi} \int_{-M}^{M} \left[\frac{\hat{\varphi}_X(t) - p \ \exp(ita)}{1-p} \right] \exp(-itx) dt \\
&= \lim_{M\to\infty} \frac{1}{2\pi(1-p)} \int_{-M}^{M} \left[\frac{1}{n} \sum_{j=1}^{n} \frac{\exp\big(it(Y_j - x)\big)}{\varphi_Z(t)} - p \ \exp\big(it(a-x)\big) \right] dt.
\end{aligned}$$

As in the pointmass estimation, $\tilde{f}_c$ is not guaranteed to be a real-valued function. Moreover, the computation of $\tilde{f}_c$ 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

$$\hat{f}_c(x) = \frac{1}{2\pi n(1-\hat{p})} \sum_{j=1}^{n} Re \int_{-M_n}^{M_n} \left[\frac{\exp\big(it(Y_j - x)\big)}{\varphi_Z(t)} - \hat{p} \ \exp\big(it(a-x)\big) \right] dt. \tag{8}$$

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

Theorems 2 and 3 below provide some asymptotic properties of $\hat{f}_c(x)$. For any $x \neq a$, we show in Theorem 2 that the proposed density estimator is a consistent estimator of $f_c(x)$ under some suitable conditions. In addition, under stronger conditions, Theorem 3 establishes the consistency of $\hat{f}_c(x)$ at $x = a$. The proofs of the theorem are provided in Section 5.

THEOREM 2 *Suppose that the conditions in Theorem 1 hold. In addition, suppose that*

$$M_n \to \infty, \quad n^{-1/2} \int_0^{M_n} \frac{1}{\varphi_Z(t)} dt \to 0 \tag{9}$$

as n goes to infinity. Then $\hat{f}_c(x)$ converges to $f_c(x)$ in probability for any $x \neq a$.

THEOREM 3 *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* (9)*, suppose that*

$$M_n = o(T_n), \quad n^{-1/2} \int_0^{T_n} \frac{1}{\varphi_Z(t)} dt = O(1), \tag{10}$$

as $n \to \infty$. *Then* $\hat{f}_c(x)$ *is a consistent estimator of* $\hat{f}_c(x)$ *at* $x = a$.

REMARK 4 When comparing Theorem 3 with Theorem 2, 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 Simulation Studies

In this section, we perform three simulation studies to investigate the performance and properties of the estimators proposed in Section 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 $\mathcal{N}(3, 1)$, $\mathcal{N}(0, 1)$ 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.

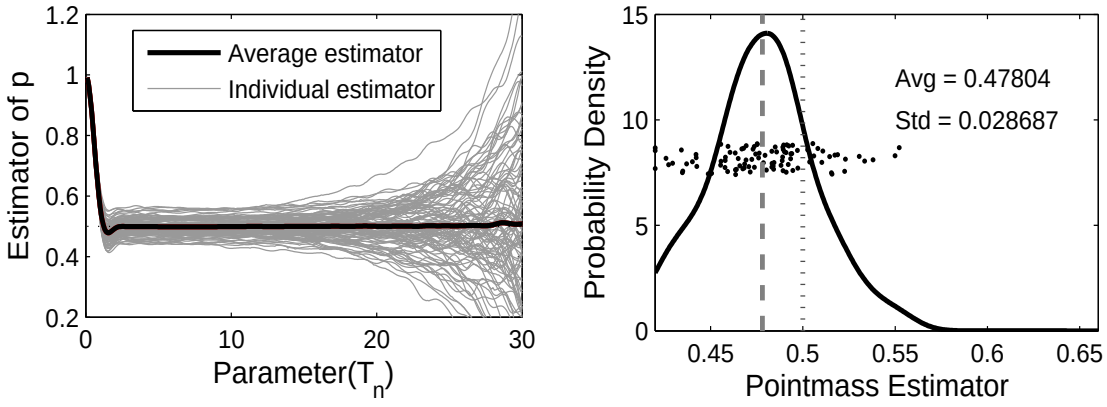
3.1 Case 1: Mixture of $\mathcal{N}(3, 1)$ and the Pointmass

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.

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

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.



(a) (b)

Figure 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.

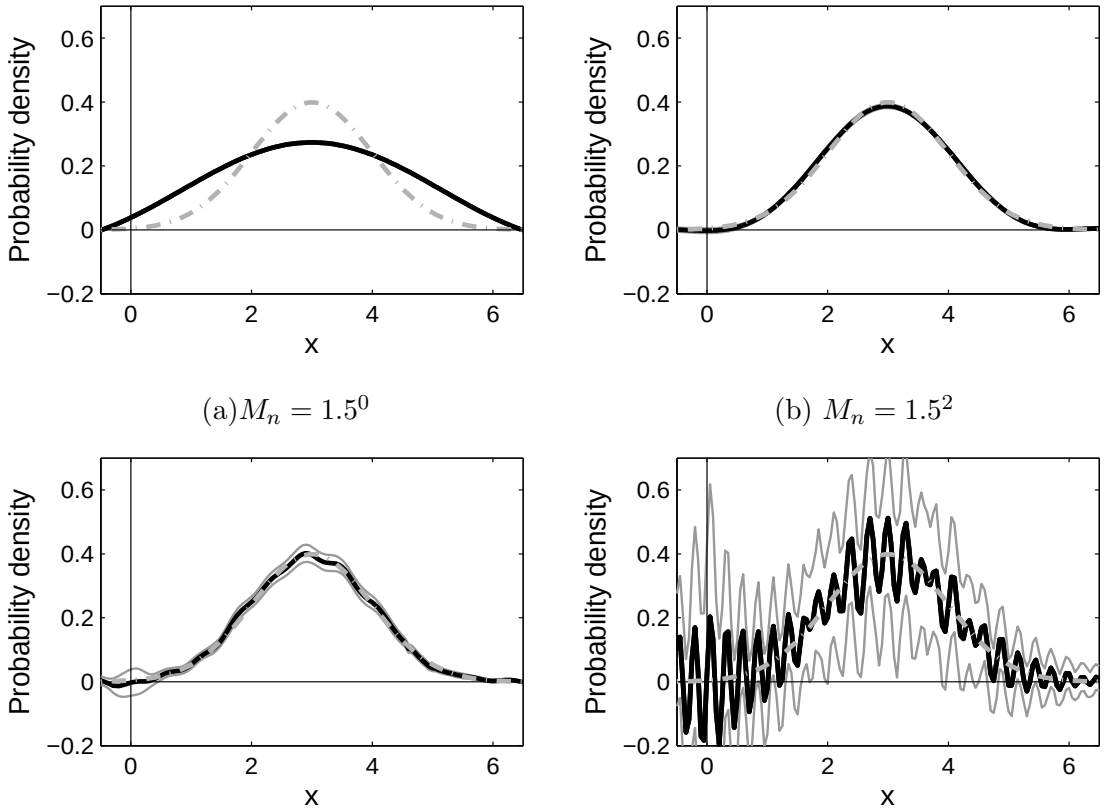
Figure 1 summarizes the performance of the pointmass estimator for 100 simulated data sets. Figure 1(a) shows the change in the pointmass estimator as a function of the integration parameter T_n of (5), where the vertical axis shows the value of $\hat{p}$. The gray curves show the pointmass estimators from the 100 samples and the black solid curve is the average of the 100 estimators. According to Figure 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 1(b), we choose a specific value of $\hat{p}$, from each gray curve Figure 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 estimator, if it lies between 0 and 1. Otherwise, e.g. there is no local minimum, T_n which gives the smallest difference of $\hat{p}$ is selected, and the corresponding $\hat{p}$ is chosen as our estimator. Figure 1(b) shows these 100 estimators with their density (the solid curve), which is estimated by the kernel smoothing method. 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 Figure 1, 3 and 5 to make the comparison clear.

Figure 2 plots the density estimator in (8) for various values of M_n. The true value of $p = 0.5$ is used in estimating the density. 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. 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



(a)$M_n = 1.5^0$ (b) $M_n = 1.5^2$

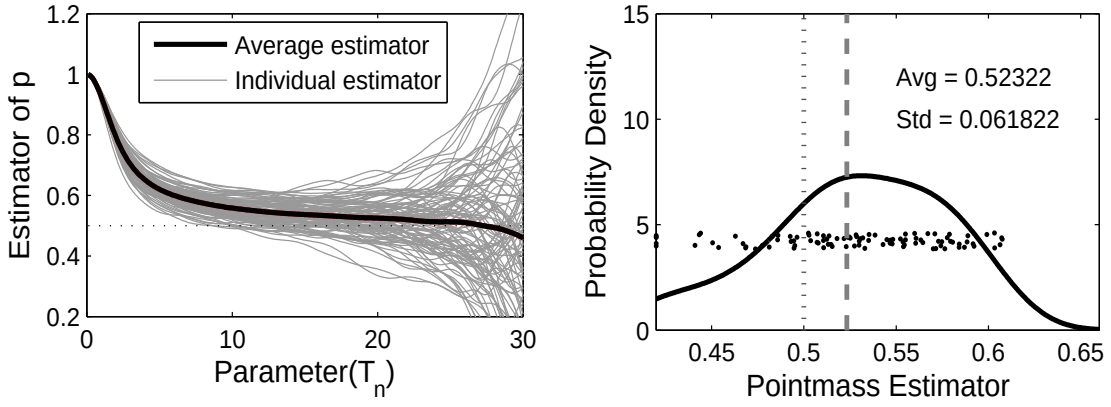
(c)$M_n = 1.5^6$ (d) $M_n = 1.5^8$

Figure 2: (Case 1) 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 $\pm$ 2 standard error, based on the 100 random samples.

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 $x \neq a$.

3.2 Case 2: Mixture of $\mathcal{N}(0,1)$ 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 $L = 100$.

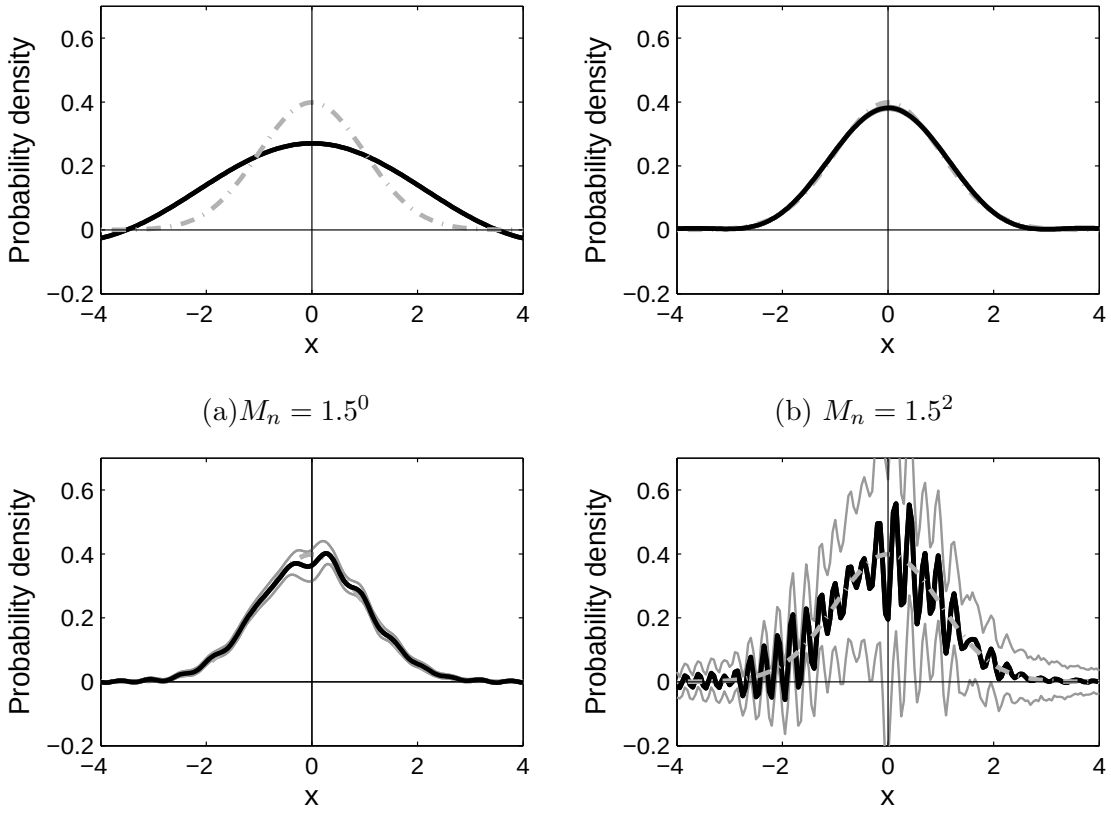


(a) (b)

Figure 3: (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 3(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. When the first local minimum is selected as the estimator, as shown in Figure 3(b), we can see the pointmass is overestimated (the average of the 100 estimators is around 0.52).



(a)$M_n = 1.5^0$ (b) $M_n = 1.5^2$

(c)$M_n = 1.5^6$ (d) $M_n = 1.5^8$

Figure 4: (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 $\pm$ 2 standard error.

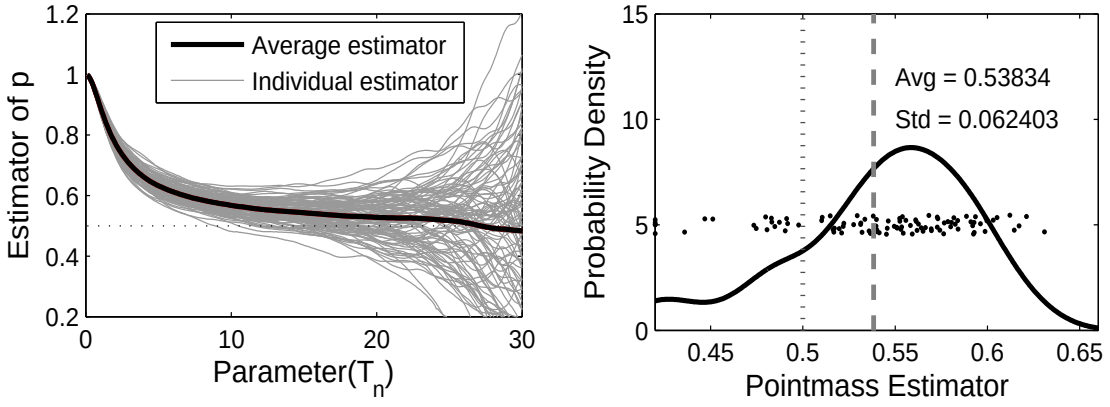
Figure 4 shows the result of the density estimation, which is also similar to the previous simulation. One big difference from the previous simulation is the trend of the standard error. In the current simulation, both the estimator $\hat{f}_c$ and the standard error are almost symmetric about 0. In 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.

3.3 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.

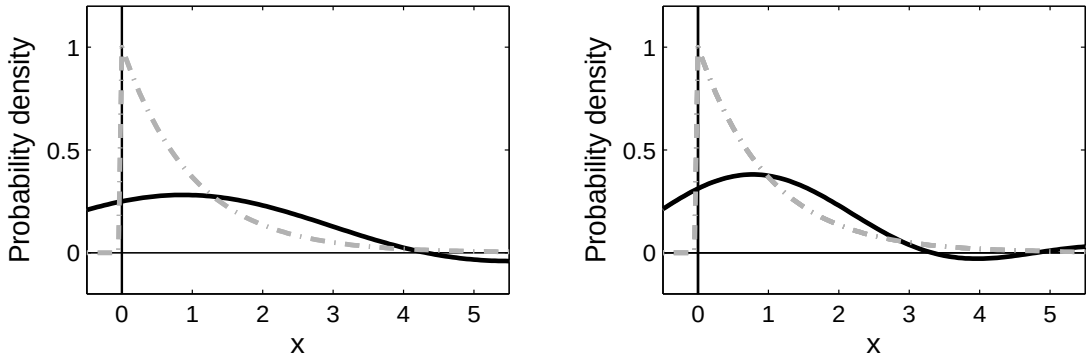
The difficulty in estimating the exponential density is that it has a non-smooth left boundary, so the estimation would not be accurate near the left boundary (at 0). Moreover, the location of the pointmass is near the peak of the exponential component. Like the second case, the estimation of both p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



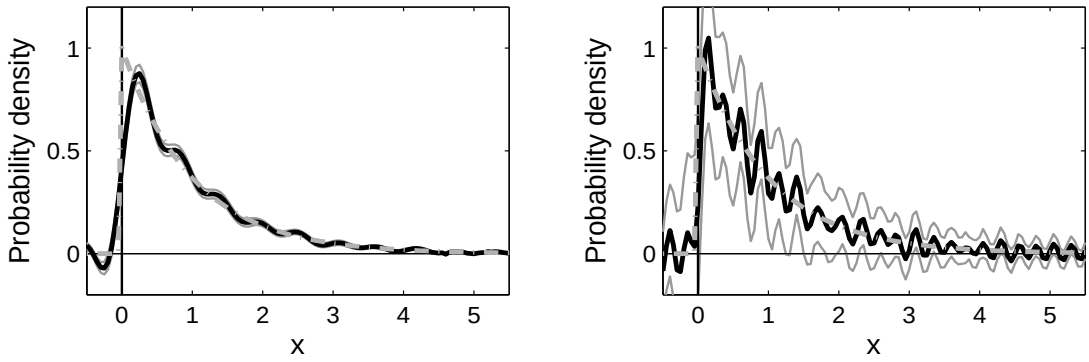
(a) (b)

Figure 5: (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.

As shown in Figure 5, the pointmass is a little overestimated. The estimation variance and bias are similar with those of the second case, but slightly larger. The estimation of f_c



(a)$M_n = 1.5^0$ (b) $M_n = 1.5^2$



(c)$M_n = 1.5^6$ (d) $M_n = 1.5^8$

Figure 6: (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.

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 6(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.

4 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. In 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. Then the mutation effect on fitness is defined as the reduction in plaque sizes. 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:

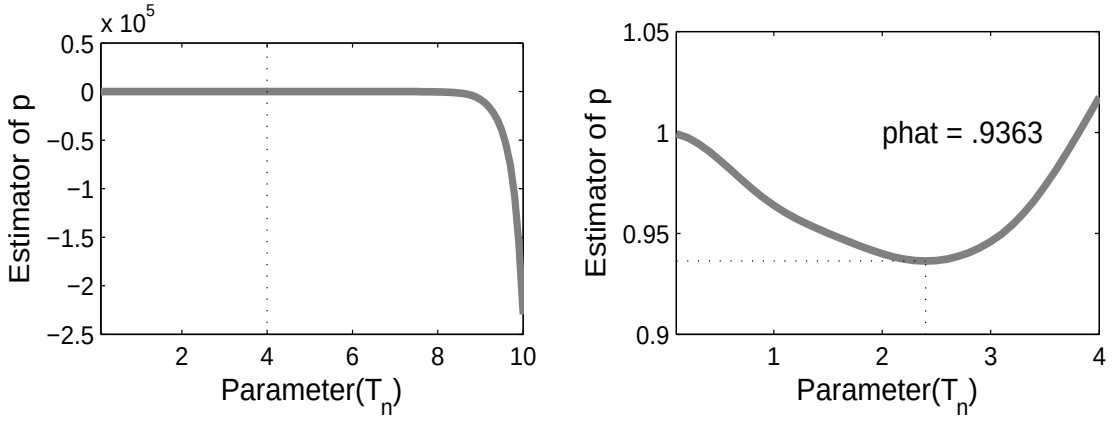
(i) No mutation occurs, or only silent mutations occur.

(ii) A deleterious mutation occurs.

Since the silent mutation does not affect fitness, theoretically the plaque size does not change, so the mutation effect is 0 for case (i). On the other hand, deleterious mutations reduce the plaque sizes, so the deleterious mutation effect 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.

We consider the pointmass estimation result first. Figure 7 plots the $\hat{p}$ versus T_n. In Figure 7(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 (5) 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 7(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 3, 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 $T_n = 2.4$.

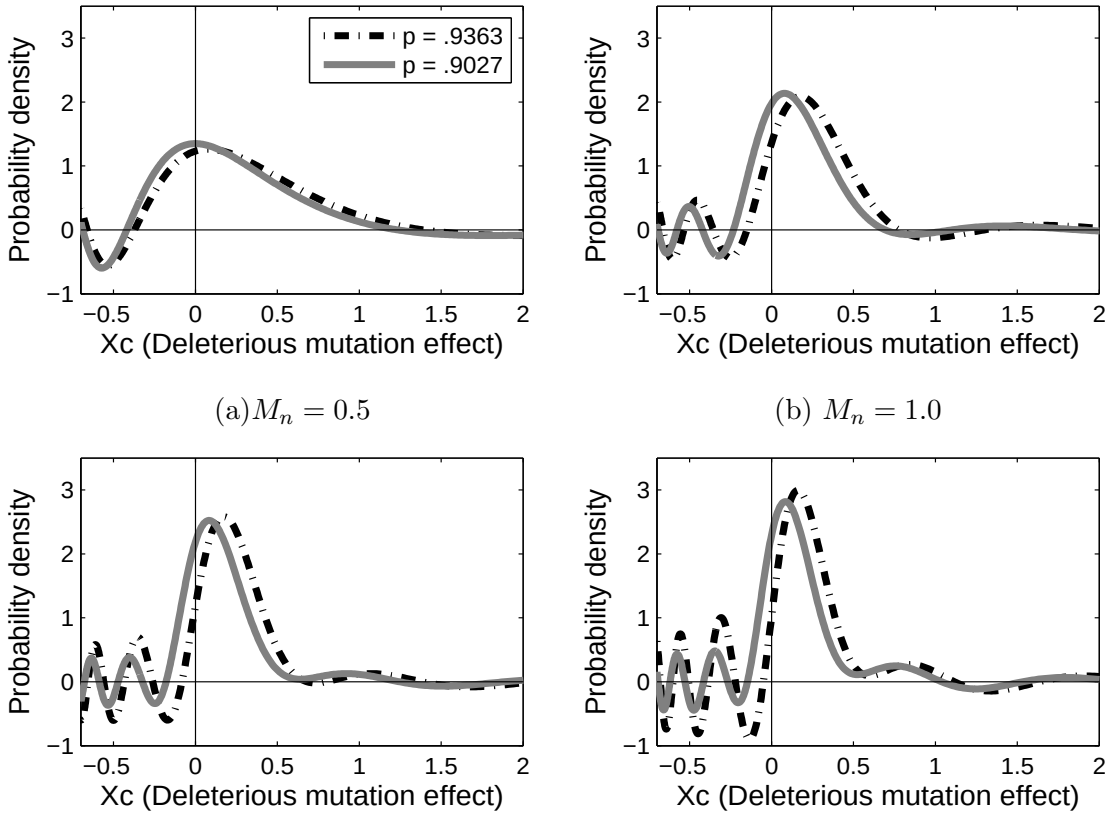


(a) (b)

Figure 7: 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 8 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



(a)$M_n = 0.5$ (b) $M_n = 1.0$

(c)$M_n = 1.3$ (d) $M_n = 1.5$

Figure 8: These plots show the density estimator $\hat{f}_c$ for different values of M_n. In each plot, the black dash-dot curve is the estimator when $\hat{p} = 0.9363$, and the gray curve is when $\hat{p} = 0.9027$.

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 8, 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 8(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 8(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.

5 Theoretical Proofs

In this section, we provide technical proofs for the lemmas and theorems in Section 2. 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 1 It can be seen by the simple change of variable technique. According to the expression of $\hat{p}$,

$$\begin{aligned} E\big(\hat{p}\big) &= E\left[\frac{1}{2T_n} Re\left\{\int_{-T_n}^{T_n} \frac{\exp\big(it(Y-a)\big)}{\varphi_Z(t)} dt\right\}\right] \\ &= \int_{-\infty}^{\infty}\int_{-T_n}^{T_n} \frac{1}{2T_n} Re\left\{\frac{\exp\big(it(y-a)\big)}{\varphi_Z(t)}\right\} f_Y(y)\, dt\, dy. \end{aligned}$$

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, $|e^{it(y-a)}|$ 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:

$$\begin{aligned} E\big(\hat{p}\big) &= \frac{1}{2T_n} Re \int_{-T_n}^{T_n} \exp(-ita)\varphi_X(t) dt \\ &= \frac{1}{2T_n} \int_{-T_n}^{T_n} \exp(-ita)\Big\{p \cdot \exp(ita) + (1-p)\varphi_c(t)\Big\} dt \\ &= p + \frac{1-p}{2T_n} \int_{-T_n}^{T_n} \exp(ita)\ \varphi_c(t) dt. \end{aligned}$$

This completes the proof. ■

Proof of Lemma 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

$$\begin{aligned}\mathrm{Var}(\hat{p}) &= \frac{1}{4nT_n^2}\mathrm{Var}\left(\int_{-T_n}^{T_n}\frac{\cos t(Y-a)}{\varphi_Z(t)}dt\right)\\ &= \frac{1}{nT_n^2}E\left(\int_0^{T_n}\frac{\cos tV - E(\cos tV)}{\varphi_Z(t)}dt\right)^2,\end{aligned}$$

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

$$\begin{aligned}&\mathrm{Var}(\hat{p})\\ &= \frac{1}{2nT_n^2}\int_0^{T_n}\int_0^{T_n}\frac{E\{\cos(s+t)V+\cos(s-t)V\} - 2E(\cos sV)E(\cos tV)}{\varphi_Z(s)\varphi_Z(t)}\,ds\,dt\\ &= \frac{1}{2nT_n^2}\int_0^{T_n}\int_0^{T_n}\frac{Re\{\varphi_V(s+t)+\varphi_V(s-t)\} - 2Re\{\varphi_V(s)\}Re\{\varphi_V(t)\}}{\varphi_Z(s)\varphi_Z(t)}\,ds\,dt.\end{aligned}$$

This completes the proof. ■

Proof of Theorem 1 To show the consistency, we will show that both the bias and the variance of $\hat{p}$ converges to 0. From Lemma 1,

$$\begin{aligned}\mathrm{bias}(\hat{p}) &= \frac{1-p}{2T_n}\int_{-T_n}^{T_n}\exp(-ita)\,\varphi_c(t)dt\\ &= \frac{(1-p)\pi}{T_n}\cdot\frac{1}{2\pi}\int_{-T_n}^{T_n}\exp(-ita)\,\varphi_c(t)dt.\end{aligned}$$

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 $|\varphi_V(t)| \le 1$ for any t,

$$\left|Re\{\varphi_V(s+t)+\varphi_V(s-t)\} - 2Re\;\varphi_V(s)\,Re\;\varphi_V(t)\right| \le 4.$$

Then the variance of $\hat{p}$ is bounded by

$$\begin{aligned}\mathrm{Var}\big(\hat{p}\big) &\leq \frac{2}{nT_n^2}\int_0^{T_n}\int_0^{T_n}\frac{1}{\varphi_Z(s)\varphi_Z(t)}\,ds\,dt \\ &= 2\left(\frac{1}{n^{1/2}\,T_n}\int_0^{T_n}\frac{1}{\varphi_Z(t)}dt\right)^2.\end{aligned}$$

Hence it converges to 0 according to (6). ■

Proof of Theorem 2 First, we divide $\hat{f}_c(x)$ into three parts:

$$\begin{aligned}\hat{f}_c(x) &= \frac{1}{2\pi(1-\hat{p})}\int_{M_n}^{M_n} Re\left\{\frac{1}{n}\sum_{j=1}^{n}\frac{\exp\big(it(Y_j-x)\big)}{\varphi_Z(t)} - \hat{p}\cdot\exp\big(it(a-x)\big)\right\}dt \\ &= \frac{1-p}{1-\hat{p}}\left[\frac{Re}{2\pi(1-p)}\int_{-M_n}^{M_n}\left\{\frac{1}{n}\sum_{j=1}^{n}\frac{\exp\big(it(Y_j-x)\big)}{\varphi_Z(t)} - p\cdot\exp\big(it(a-x)\big)\right\}dt\right. \\ &\qquad\left. + \frac{Re}{2\pi(1-p)}\int_{M_n}^{M_n}(p-\hat{p})\exp\big(it(a-x)\big)dt\right] \\ &= T_1(T_2+T_3),\end{aligned}$$

where

$$\begin{aligned}T_1 &= \frac{1-\hat{p}}{1-p}, \\ T_2 &= \frac{1}{2\pi(1-p)}Re\int_{-M_n}^{M_n}\left\{\frac{1}{n}\sum_{j=1}^{n}\frac{\exp\big(it(Y_j-x)\big)}{\varphi_Z(t)} - p\cdot\exp\big(it(a-x)\big)\right\}dt, \\ \text{and}\quad T_3 &= \frac{1}{2\pi(1-p)}Re\int_{M_n}^{M_n}(p-\hat{p})\exp\big(it(a-x)\big)dt.\end{aligned}$$

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 1, 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 $x \neq a$,

$$\begin{aligned} T_3 &= \big(p - \hat{p}\big) \cdot Re \int_{-M_n}^{M_n} e^{it(a-x)} dt \\ &= \big(p - \hat{p}\big) \int_{-M_n}^{M_n} \cos t(a-x) dt \\ &= 2\big(p - \hat{p}\big) \frac{\sin M_n(a-x)}{a-x}. \end{aligned} \tag{11}$$

Here, $|\sin M_n(a-x)|$ is uniformly bounded by 1, $|T_3| \leq 2|p - \hat{p}|/|a - x|$. Moreover, 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 $\mathrm{Var}(T_2) \to 0$ as n goes to infinity. From the definition of T_2,

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

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

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

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

$$\begin{aligned} E(T_2) &= \frac{1}{2\pi(1-p)} Re \int_{-M_n}^{M_n} \left\{ \frac{\varphi_Y(t) \exp(-itx)}{\varphi_Z(t)} - p \cdot \exp\big(it(a-x)\big) \right\} dt \\ &= \frac{1}{2\pi(1-p)} Re \int_{M_n}^{M_n} \Big\{ \varphi_X(t) - p \cdot \exp(ita) \Big\} \exp(-itx)\ dt. \\ &= \frac{1}{2\pi} Re \int_{M_n}^{M_n} \varphi_c(t)\ \exp(-itx)\ dt. \end{aligned}$$

The last equality comes from the mixture structure of X. According to (7), we get the fact that $(\varphi_X(t) - p \cdot \exp(ita))/(1-p)$ is the same as $\varphi_c(t)$. Since M_n goes to infinity as n increases,

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

as n goes to infinity.

The next part show the calculation of the variance of T_2, which is very similar to the computation of $\mathrm{Var}(\hat{p})$ in the proof of Lemma 2.

$$
\begin{aligned}
&\mathrm{Var}(T_2) \\
&= \mathrm{Var}\left[\frac{1}{2n\pi(1-p)}\sum_{j=1}^{n}\int_{-M_n}^{M_n}\left\{\frac{\cos t(Y_j-x)}{\varphi_Z(t)} - p\cdot\exp\big(it(a-x)\big)\right\}dt\right] \\
&= \mathrm{Var}\left[\frac{1}{2n\pi(1-p)}\sum_{j=1}^{n}\int_{-M_n}^{M_n}\frac{\cos t(Y_j-x)}{\varphi_Z(t)}dt\right] \\
&= \frac{1}{4n\pi^2(1-p)^2}\mathrm{Var}\left[\int_{-M_n}^{M_n}\frac{\cos t(Y-x)}{\varphi_Z(t)}dt\right] \\
&= \frac{1}{n\pi^2(1-p)^2}\,E\left[\int_0^{M_n}\frac{\cos tV - E\cos tV}{\varphi_Z(t)}dt\right]^2 \quad \text{where} \quad V=Y-x \\
&= \frac{1}{n\pi^2(1-p)^2}\int_0^{M_n}\int_0^{M_n}\left[\frac{Re\big\{\varphi_V(s+t)+\varphi_V(s-t)\big\}}{2} - Re\big\{\varphi_V(s)\big\}Re\big\{\varphi_V(t)\big\}\right] \\
&\qquad\qquad \times\frac{1}{\varphi_Z(s)\varphi_Z(t)}\,ds\,dt \\
&\le \frac{2}{\pi^2(1-p)^2}\left(n^{-1/2}\int_0^{M_n}\frac{1}{\varphi_Z(t)}dt\right)^2.
\end{aligned}
$$

From (9), 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 3 The proof is the same as the proof of Theorem 2, except the convergence of T_3 in (11).

Since $M_n \to \infty$ and $M_n = o(T_n)$, T_n also goes to infinity as $n \to \infty$. In addition, by (10),

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

This means that all conditions in Theorem 1 are satisfied, so $\hat{p}$ converges to p in probability.

Then T_1 and T_2 in the proof of Theorem 2 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.

The difficulty of providing the convergence of T_3 comes from the fact that the integration in (11) is not bounded when $x = a$. Since (11) is the same as $2M_n(\hat{p} - p)$, 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 (10) is satisfied,

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

In addition,

$$\begin{aligned} &\mathrm{Var}\big(M_n(\hat{p} - p)\big) \\ &= \frac{M_n^2}{2nT_n^2} \int_0^{T_n} \int_0^{T_n} \frac{Re\big\{\varphi_V(s+t) + \varphi_V(s-t)\big\} - 2Re\big\{\varphi_V(s)\big\} Re\big\{\varphi_V(t)\big\}}{\varphi_Z(s)\varphi_Z(t)}\, ds\, dt \\ &\le 2 \left(\frac{M_n}{n^{1/2} T_n} \int_0^{T_n} \frac{1}{\varphi_Z(t)} dt \right)^2 \to 0. \end{aligned}$$

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. ■

Acknowledgements

J. S. Marron, Haipeng Shen and Mihee Lee are partially supported by NSF grant DMS-0606577. Christina Burch is partially supported by NIH grant R01-GM067940.

References

Billingsley, P. (1995), *Probability and Measure*, Wiley Interscience, 3rd ed.

Burch, C., Guyader, S., Samarov, D., and Shen, H. (2007), "Experimental estimate of the abundance and effects of nearly neutral mutations in the RNA virus ϕ6," *Genetics*, 176, 467–476.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. (2006), *Measurement Error in Nonlinear Models*, Chapman & Hall, 2nd ed.

Chaudhuri, P. and Marron, J. S. (2000), "Scale space view of curve estimation," *The Annals of Statistics*, 28, 408–428.

Cuevas, A. and Walter, G. (1992), "On estimation of generalized densities," *Communications in Statistics - Theory and Methods*, 21, 1807–1821.

Elena, S., Ekunwe, L., Hajela, N., Oden, S., and Lenski, R. (1998), "Distribution of fitness effects caused by random insertion mutations in Escherichia coli," *Genetica*, 102/103, 349–358.

Liu, M. and Taylor, R. (1989), "A consistent nonparametric density estimator for the deconvolution problem," *The Canadian Journal of Statistics*, 17, 427–438.

Sanjuan, R., Moya, A., and Elena, S. (2004), "The distribution of fitness effects by single-nucleotide substitutions in an RNA virus," *PNAS*, 101, 8396–8401.