Nonparametric and parametric estimators of prevalence from group testing data with aggregated covariates

Aurore Delaigle and Wen-Xin Zhou
Department of Mathematics and Statistics, University of Melbourne, VIC 3010, Australia

Abstract: Group testing is a technique employed in large screening studies involving infectious disease, where individuals in the study are grouped before being observed. Parametric and nonparametric estimators of conditional prevalence have been developed in the group testing literature, in the case where the binary variable indicating the disease status is available only for the group, but the explanatory variable is observed for each individual. However, for reasons such as the high cost of assays, the confidentiality of the patients, or the impossibility of measuring a concentration under a detection limit, the explanatory variable is observable only in an aggregated form and the existing techniques are no longer valid. We develop consistent parametric and nonparametric estimators of the conditional prevalence in this complex problem. We establish theoretical properties of our estimators and illustrate their practical performance on simulated and real data. We extend our techniques to the case where the group status is measured imperfectly, and to the setting where the covariate is aggregated and the individual status is available.

Keywords: averaged biomarker, binary outcome, conditional probability, confidentiality, cost of assays, nonparametric regression with aggregated data, nonparametric regression with pooled biomarker, pooled data.

1 Introduction

In large screening studies related to rare infectious disease, individuals are often pooled before being tested for the disease. There, instead of observing the result $Y = 0$ or 1 of the test for each individual, we observe only the result Y^* of the test for groups of individuals. Originally, this group testing approach was introduced to gain time and money; see Dorfman (1943) for an example with US soldiers tested for syphilis. Nowadays this group testing procedure is employed in much more diverse contexts, such as detection of pollution by a toxic substance, or the estimation of proportion of transgenic corn; see, for example, Montesinos-López et al. (2012, 2013).

In group testing problems, it is often of interest to estimate the probability of contamination given a covariate X; that is, $p(x) = P(Y = 1|X = x)$ (we use contamination as a generic term, which can represent contamination by a pollutant, the presence of a disease, evidence of genetic manipulation, etc). Parametric methods of estimation and inference have been developed in the literature to estimate this conditional probability under various settings; see for example Vansteelandt et al. (2000), Xie (2001), Bilder and Tebbs (2009), Chen et al. (2009) and Huang and Tebbs (2009). For semiparametric and nonparametric methods, we refer to Delaigle and Meister (2011), Delaigle and Hall (2012), Li and Xie (2012), Wang et al. (2013), Wang et al. (2014) and Delaigle et al. (2014). See also Chen and Swallow (1990), Farrington (1992), Hardwick et al. (1998), Gastwirth and Johnson (1994) and Hung and Swallow (2000) for related work on estimation and inference about prevalence.

The existing techniques for estimating p rely on the fact that X is available at an individual level, whereas in some cases it is only available in an aggregated form at the group level. Most often, continuous biomarkers are pooled because the cost of assays is too high to be able to take individual measurements; see for example Weinberg and Umbach (1999), Caudill (2010) and Zhang and Albert (2011), who discuss studies where X is the exposure to a toxic substance, and Faraggi et al. (2003), Liu and Schisterman (2003) and Mitchell et al. (2014), who discuss pooling of biomarkers more generally. Grouping can also be applied for other reasons such as preserving confidentiality of participants in a study (Gastwirth and Hammick, 1989), to sparingly use non-renewable blood specimens (Weinberg and Umbach, 1999), or because X represents the concentration of a substance whose individual value is below the detection limit, whereas the group concentration can be measured. For instance, such grouping was introduced by the National Centers for Disease Control and Prevention in the NHANES 2005–2006 study, to measure the concentration of organochlorine

pesticides and metabolites. See also Caudill (2010) for other related examples. When X is pooled, existing methods cannot be pressed into service.

Our goal is to develop new methods (nonparametric and parametric) for estimating p from group testing data with aggregated covariates. Nonparametric estimation from aggregated data was considered by others before, but not in our context of group testing outcomes variables. Meister (2007) considered the problem of estimating the density of X from a sample of aggregated data. In Linton and Whang (2002), the goal was to estimate a regression curve $E(Y|X)$, where X and Y are both continuous random variables, which can only be observed in the form of averages contaminated by additive noise (in our case Y is a binary variable observed in the form of a maximum Y^* within a group). In Zhang and Albert (2011), the focus was on estimating $E(Y|X)$ parametrically when Y is binary and X is continuous and aggregated, but in their case the response Y was not grouped. Our problem, which combines aspects of group testing with aspects of aggregated data, is particularly difficult. A first nontrivial task is to express the function p in terms of quantities that can be estimated from our indirect data. A second challenge is to find a way to define a local polynomial estimator for this problem. In the standard setting, this estimator is defined through a least squares equation which cannot be computed when only grouped and aggregated data are available. We circumvent this difficulty through an ingenious use of Fourier transforms and the development of empirical estimators thereof.

In Section 2, we introduce our regression model. In Sections 3 and 4, we introduce our nonparametric and parametric estimators, respectively; their theoretical properties are presented in Section 5. In Section 6, we suggest a data-driven procedure for bandwidth selection and illustrate the numerical performance of our estimators on simulated and real data. In Section 7.1, we extend our results to the case where the group status is observed with errors, and in Section 7.2 we discuss the extension of

our nonparametric estimators to the setting of Zhang and Albert (2011), where the individual Y data are available, and only the X observations are aggregated. Proofs of the theoretical results are provided in Section A and in a supplemental file.

2 Model, data and distributions

2.1 Model and data

We are interested in estimating the regression curve $p(x) = P(Y = 1|X = x)$, where Y is a binary variable indicating the presence ($Y = 1$) or absence ($Y = 0$) of a contaminant or a disease, given that the covariate X takes the value x. The ideal, unobserved, data consist of independent and identically distributed (i.i.d.) pairs (X_{ij}, Y_{ij}), $i = 1, \ldots, \nu$, $j = 1, \ldots, n$. The index $_{ij}$ represents the ith individual from the jth group, and the group size $\nu \geq 2$. We let $N = n\nu$ denote the total number of individuals, and let f_X denote the density of the X_{ij}'s.

In our group testing context, the (X_{ij}, Y_{ij})'s are not available. Instead we observe $(S_1, Y_1^*), \ldots, (S_n, Y_n^*)$, where for $j = 1, \ldots, n$, $Y_j^* = \max_{i=1,\ldots,\nu} Y_{ij}$ denotes the result of the test carried out on the entire jth group, and $S_j = \sum_{i=1}^{\nu} X_{ij}$ is the aggregated value of X in the jth group.

Remark 2.1. In Section 7.2, we will show how to extend our ideas to the setting considered by Zhang and Albert (2011), where the X_{ij}'s are aggregated, but the Y_{ij}'s are not grouped, i.e. where we observe (S_j, Y_{ij}), for $i = 1, \ldots, \nu$ and $j = 1, \ldots, n$.

2.2 Conditional distribution of the data

We wish to construct parametric and nonparametric estimators of the curve p from the data $(S_1, Y_1^*), \ldots, (S_n, Y_n^*)$. The methodologies we shall suggest exploit the con-

ditional distribution of $Y_j^*|S_j$. To derive this distribution, for all $x \in \mathbb{R}$ let

$$q(x) = 1 - p(x) = P(Y = 0|X = x) \quad \text{and} \quad m(x) = q(x)f_X(x). \tag{2.1}$$

Recall that $Y = 0$ or 1 is a Bernoulli random variable, so that Y_j^* also follows a Bernoulli distribution. Now

$$\begin{aligned}
&P(Y_j^* = 0, S_j \le x) \\
&= \int_{-\infty}^{\infty} P\left(Y_{1j} = \cdots = Y_{\nu j} = 0\,, \sum_{i=2}^{\nu} X_{ij} \le x - x_1 \middle| X_1 = x_1\right) f_X(x_1)\,dx_1 \\
&= \int_{-\infty}^{x} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} m\left(u - \sum_{k=1}^{\nu-1} x_k\right) \prod_{k=1}^{\nu-1} m(x_k)\,dx_k\,du \\
&= \int_{-\infty}^{x} m^{*\nu}(u)\,du\,,
\end{aligned}$$

where we use $f^{*\nu} = f * \cdots * f$ to denote the ν-fold convolution of a function f. Letting $f_S = f_X^{*\nu}$ denote the density of the S_j's, this implies that

$$P(Y_j^* = 0|S_j = x) = m^{*\nu}(x)/f_S(x) \tag{2.2}$$

and $P(Y_j^* = 1|S_j = x) = 1 - f_S^{-1}(x)m^{*\nu}(x)$.

3 Nonparametric estimators

3.1 Nadaraya-Watson type estimator

We start by constructing a basic ratio-type estimator similar in spirit to the Nadaraya-Watson estimator employed in standard nonparametric regression problems. Specifically, following (2.1), we write $p(x) = 1 - f_X^{-1}(x)m(x)$ and assume that

$$m_0 = \int_{-\infty}^{\infty} \{1 - p(x)\} f_X(x)\,dx = \int_{-\infty}^{\infty} m(x)\,dx > 0. \tag{3.1}$$

For any function $f \in L_1(\mathbb{R})$, let $\phi_f(t) = \int_{-\infty}^{\infty} e^{\mathrm{i}tx} f(x)\,dx$ denote its Fourier transform, and for any random variable U, denote by $\phi_U(t) = E(e^{\mathrm{i}tU})$ its characteristic function.

Throughout we assume that

$$\phi_X(t) \neq 0 \quad \text{and} \quad \phi_m(t) \neq 0, \quad \text{for all } t \in \mathbb{R}. \tag{3.2}$$

We estimate f_X and m separately. For f_X, we can follow Meister's (2007) nonparametric approach, as follows. Recall that, by the Fourier inversion theorem, if $\phi_X \in L_1(\mathbb{R})$, then $f_X(x) = (2\pi)^{-1} \int_{-\infty}^{\infty} e^{-\mathrm{i}tx}\, \phi_X(t)\, dt$. Therefore, to construct an estimator of f_X it suffices to construct an estimator of ϕ_X. As in Meister (2007), using the aggregated data, first we estimate $\phi_S(t) = \{\phi_X(t)\}^\nu$ by

$$\hat{\phi}_S(t) := n^{-1} \sum_{j=1}^{n} e^{\mathrm{i}tS_j}. \tag{3.3}$$

Among the multiple complex-valued roots of the complex-valued function $\hat{\phi}_S$, to find the one which corresponds to a valid estimator of ϕ_X we use a slightly modified version of Meister's (2007) estimator; see Section C.1 for details. Then, as in Meister (2007), we estimate f_X by

$$\hat{f}_X(x) = (2\pi)^{-1} \int_{-\infty}^{\infty} e^{-\mathrm{i}tx}\, \phi_K(th) \hat{\phi}_X(t)\, dt, \tag{3.4}$$

where ϕ_K denotes the Fourier transform of a symmetric kernel function K, and $h = h_n > 0$ is a bandwidth. As noted in the introduction, Linton and Whang (2002) considered a related problem of nonparametric curve estimation from aggregated data which also involves estimating ϕ_X from averaged data. Their approach is similar to the one discussed above, except that it implicitly assumes that ϕ_S is real, which, in general, is only satisfied when X is a symmetric random variable.

To estimate the function m, again by the Fourier inversion theorem, if $\phi_m \in L_1(\mathbb{R})$, then $m(x) = (2\pi)^{-1} \int_{-\infty}^{\infty} e^{-\mathrm{i}tx}\, \phi_m(t)\, dt$. Next, (2.2) implies that $m^{*\nu} \leq f_S$, and thus the Fourier transform of $m^{*\nu}$ exists. To estimate this Fourier transform, recall that basic properties of the Fourier transform imply $\phi_{m^{*\nu}}(t) = \{\phi_m(t)\}^\nu$. For $j = 1, \ldots, n$,

let $Z_j^* = 1 - Y_j^*$ so that $m^{*\nu}(x) = f_S(x)P(Z_j^* = 1|S_j = x)$ in view of (2.2). Taking the Fourier transform of both sides of this equation, we deduce that

$$\phi_{m^{*\nu}}(t) = E\{P(Z_j^* = 1|S_j)e^{\mathrm{i}tS_j}\} = E(Z_j^* e^{\mathrm{i}tS_j}). \tag{3.5}$$

Equation (3.5) suggests a natural estimator $\hat{\phi}_{m^{*\nu}}(t) := n^{-1}\sum_{j=1}^n Z_j^* e^{\mathrm{i}tS_j}$ of $\phi_{m^{*\nu}}(t)$, from which we deduce an estimator $\hat{\phi}_m$ of ϕ_m; see Section C.1 for full details. We then estimate $m(x)$ by

$$\hat{m}(x) = (2\pi)^{-1}\int_{-\infty}^{\infty} e^{-\mathrm{i}tx}\,\phi_K(th)\hat{\phi}_m(t)\,dt, \tag{3.6}$$

and define our first nonparametric estimator of $p(x)$ by

$$\hat{p}(x) = 1 - \hat{m}(x)/\hat{f}_X(x). \tag{3.7}$$

3.2 Local polynomial type estimator

Next we construct a version of the more general and widely used local polynomial estimator, which can be computed from our data. In addition to estimating p, it can estimate derivatives $p^{(d)}$, $d \geq 1$. In the standard i.i.d. case, which corresponds to the case where we observe the data (X_{ij}, Y_{ij}), $i = 1, \ldots, \nu$ and $j = 1, \ldots, n$, the ℓth order local polynomial estimator of p is obtained as follows. First, approximate p in a neighbourhood of x by an ℓth order polynomial, that is $p(u) \approx \beta_0 + \beta_1(u-x) + \ldots + \beta_\ell(u-x)^\ell$. Then, at each x, estimate the local coefficients $\beta_k = \beta_k(x)$ by $\hat{\beta}_k = \hat{\beta}_k(x)$ for $k = 0, \ldots, \ell$, obtained via minimising the following weighted least-squares sum with respect to the β_k's:

$$\sum_{i,j}\left\{Y_{ij} - \sum_{k=0}^{\ell}\beta_k(X_{ij}-x)^\ell\right\}^2 K_h(X_{ij}-x), \tag{3.8}$$

where $K_h(y) = h^{-1}K(y/h)$, K is a kernel function and $h = h_n > 0$ is a bandwidth. For $d \leq \ell$, the ℓth order local polynomial estimator of $p^{(d)}(x)$ is defined

by $\hat{p}^{(d)}(x) = d!\,\hat{\beta}_d(x)$, which can be written as $\hat{p}^{(d)}(x) = d!\,h^{-d}\,\mathbf{e}_d^{\mathrm{T}}\,\mathbf{S}_N^{-1}\,\mathbf{T}_N$, where $\mathbf{e}_d = (0,\ldots,0,1,0\ldots.,0)^{\mathrm{T}}$ with 1 at the $(d+1)$th entry, $\mathbf{S}_N = (S_{N,k,k'})_{0\le k,k'\le \ell}$ and $\mathbf{T}_N = (T_{N,0},\ldots,T_{N,\ell})^{\mathrm{T}}$, with $S_{N,k,k'}(x) = (Nh^{k+k'})^{-1}\,\sum_{j=1}^{n}\,\sum_{i=1}^{\nu}\,K_h(X_{ij}-x)\,(X_{ij}-x)^{k+k'}$ and $T_{N,k}(x) = (Nh^{k})^{-1}\,\sum_{j=1}^{n}\,\sum_{i=1}^{\nu}\,Y_{ij}\,K_h(X_{ij}-x)\,(X_{ij}-x)^{k}$.

It does not seem possible to develop an analogue of (3.8) here, since we observe neither the X_{ij}'s nor the Y_{ij}'s. Instead we could construct versions of $S_{N,k,k'}$ and $T_{N,k}$ that can be computed from our data and have the same asymptotic limit as $S_{N,k,k'}$ and $T_{N,k}$. In errors-in-variables problems that do not involve group testing data, faced with a similar problem, Delaigle et al. (2009) constructed, in the Fourier domain, conditionally unbiased estimators of each $K_h(X_{ij}-x)\,(X_{ij}-x)^{k+k'}$. We cannot do this in our highly nonlinear context, but we too shall exploit Fourier properties.

Instead of directly estimating $p^{(d)}$, it is simpler to start with an estimator of $q^{(d)}$. Let $Z_{ij} = 1-Y_{ij}$, let $U_{N,k}$ denote the version of $T_{N,k}$ with each Y_{ij} replaced by Z_{ij}, and let $\mathbf{U}_N = (U_{N,0},\ldots,U_{N,\ell})^{\mathrm{T}}$. If the (X_{ij},Z_{ij})'s were observed, we could compute the standard local polynomial estimator of $q^{(d)}(x)$ defined by $\hat{q}^{(d)}(x) = d!\,h^{-d}\,\mathbf{e}_d^{\mathrm{T}}\,\mathbf{S}_N^{-1}\,\mathbf{U}_N$. We construct versions $\hat{S}_{N,k,k'}$ and $\hat{U}_{N,k}$ of $S_{N,k,k'}$ and $U_{N,k}$ that can be computed from our data, as follows. In Section C.2 we prove that

$$\phi_{S_{N,k,k'}}(t) = \mathrm{i}^{-k-k'}\,\hat{\phi}_{X,\mathrm{emp}}(t)\phi_K^{(k+k')}(-ht),\ \ \phi_{U_{N,k}}(t) = \mathrm{i}^{-k}\,\hat{\phi}_{ZX,\mathrm{emp}}(t)\phi_K^{(k)}(-ht), \quad (3.9)$$

where $\hat{\phi}_{X,\mathrm{emp}}(t) = N^{-1}\sum_{j=1}^{n}\,\sum_{i=1}^{\nu}\,e^{\mathrm{i}tX_{ij}}$ is an unbiased estimator of $\phi_X(t)$, and $\hat{\phi}_{ZX,\mathrm{emp}}(t) = N^{-1}\sum_{j=1}^{n}\,\sum_{i=1}^{\nu}\,Z_{ij}\,e^{\mathrm{i}tX_{ij}}$ is an unbiased estimator of

$$\phi_{ZX}(t) = E(Z_{ij}e^{\mathrm{i}tX_{ij}}) = E\{q(X_{ij})e^{\mathrm{i}tX_{ij}}\} = \phi_{qf_X}(t) = \phi_m(t).$$

Of course we cannot compute $\hat{\phi}_{X,\mathrm{emp}}$ and $\hat{\phi}_{ZX,\mathrm{emp}}$, but instead we can use the estimators $\hat{\phi}_X$ and $\hat{\phi}_m$ that we constructed in Section 3.1. Using Fourier inversion, this motivates us to define $\hat{\mathbf{S}}_N = (\hat{S}_{N,k,k'})_{0\le k,k'\le \ell}$ and $\hat{\mathbf{U}}_N = (\hat{U}_{N,0},\ldots,\hat{U}_{N,\ell})^{\mathrm{T}}$, where

$$\hat{S}_{N,k,k'}(x) = (2\pi\,\mathrm{i}^{k+k'})^{-1}\int e^{-\mathrm{i}tx}\,\hat{\phi}_X(t)\phi_K^{(k+k')}(-ht)\,dt,$$

$$\hat{U}_{N,k}(x) = (2\pi \mathrm{i}^k)^{-1} \int e^{-\mathrm{i}tx}\, \hat{\phi}_m(t) \phi_K^{(k)}(-ht)\, dt.$$

Recalling that $p = 1 - q$, we define our local polynomial type estimator of $p^{(d)}$ by

$$\hat{p}^{\,(d)} = I\{d = 0\} - \hat{q}^{\,(d)}, \quad \text{where} \quad \hat{q}^{\,(d)} = d!\, h^{-d}\, \mathbf{e}_d^{\mathrm{T}}\, \hat{\mathbf{S}}_N^{-1}\, \hat{\mathbf{U}}_N. \tag{3.10}$$

When $\ell = 0$ and $d = 0$, this estimator reduces to the estimator derived in Section 3.1.

Remark 3.1. (Which order ℓ should we use in practice?) In standard nonparametric regression problems, the local linear estimator (local polynomial estimator with $\ell = 1$) is almost always preferred to the Nadaraya-Watson estimator (local polynomial estimator with $\ell = 0$). In our case too, we found that the local linear estimator gave better results, and this is the estimator we recommend using in practice.

4 Parametric estimator of p

As commonly assumed in the group testing literature, we sometimes have at our disposal a parametric model for p. There, p takes a parametric form $p_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta} \in \Theta$ is a d-dimensional parameter, and $\Theta \subseteq \mathbb{R}^d$ is a compact set. Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$. It follows from our calculations in Section 2.2 that the conditional likelihood of the $Y_j^*|S_j$'s is given by

$$L(Y_1^*, \ldots, Y_n^* | S_1, \ldots, S_n) = \prod_{j=1}^{n} f(Y_j^*, S_j \,|\, \boldsymbol{\theta}),$$

where, for $y = 0$ or 1, $f(y, s \,|\, \boldsymbol{\theta}) = \{1 - m_{\boldsymbol{\theta}}^{*\nu}(s)/f_S(s)\}^y \{m_{\boldsymbol{\theta}}^{*\nu}(s)/f_S(s)\}^{1-y} = y\{1 - m_{\boldsymbol{\theta}}^{*\nu}(s)/f_S(s)\} + (1-y)\{m_{\boldsymbol{\theta}}^{*\nu}(s)/f_S(s)\}$, with $m_{\boldsymbol{\theta}} = q_{\boldsymbol{\theta}}\, f_X = (1 - p_{\boldsymbol{\theta}}) f_X$.

Ideally we would estimate $\boldsymbol{\theta}_0$ by the vector $\hat{\boldsymbol{\theta}}$ that maximises the log-likelihood, or equivalently, that maximises $\ell_n(\boldsymbol{\theta}) := (2n)^{-1} \sum_{j=1}^{n} \ln \big\{ f(Y_j^*, S_j \,|\, \boldsymbol{\theta})\, f_S(S_j) \big\}$. However, in the most general case, no parametric model is available for f_X, so that f_X and f_S are unknown. Denote the standard kernel density estimator of $f_S(s)$ by

$\hat{f}_S(s) = (nb)^{-1} \sum_{j=1}^n \tilde{K}\{(s - S_j)/b\}$, where $\tilde{K}$ is a kernel function and $b = b_n > 0$ is a bandwidth, and let $\hat{f}_X$ as defined at (3.4). The two equivalent forms of $f(y, s \,|\, \boldsymbol{\theta})$ suggest two ways of estimating $\boldsymbol{\theta}_0$: estimate $\boldsymbol{\theta}_0$ by the vector $\tilde{\boldsymbol{\theta}} \in \Theta$ that maximises

$$\tilde{\ell}_n(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{j=1}^n \ln \left(\max \left[c_0, Y_j^* \hat{f}_S(S_j) + (1 - 2Y_j^*) \big\{ (q_{\boldsymbol{\theta}} \hat{f}_X)^{*\nu}(S_j) \big\} \right] \right), \tag{4.1}$$

under the constraint that $0 \le \hat{p}_\theta \le 1$, and where $c_0 > 0$ is a small constant, or estimate $\boldsymbol{\theta}_0$ by $\hat{\boldsymbol{\theta}} \in \Theta$ that maximises

$$\hat{\ell}_n(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{j=1}^n \left[Y_j^* \ln \big\{ \hat{f}_S(S_j) - (q_{\boldsymbol{\theta}} \hat{f}_X)^{*\nu}(S_j) \big\} + (1 - Y_j^*) \ln \big\{ (q_{\boldsymbol{\theta}} \hat{f}_X)^{*\nu}(S_j) \big\} \right] \tag{4.2}$$

subject to

$$\hat{f}_S(S_j) - (q_{\boldsymbol{\theta}} \hat{f}_X)^{*\nu}(S_j) \ge c_0 \quad \text{and} \quad (q_{\boldsymbol{\theta}} \hat{f}_X)^{*\nu}(S_j) \ge c_0 \tag{4.3}$$

for all $1 \le j \le n$. The two approaches are essentially equivalent, and the theory established in Theorem 5.3 for $\hat{\boldsymbol{\theta}}$ can be easily adapted to $\tilde{\boldsymbol{\theta}}$. We refer to Section D.1 for details of implementation.

Remark 4.1. The case where we have a parametric model for f_X, say $f_{\boldsymbol{\gamma}}$ with $\boldsymbol{\gamma} \in \mathbb{R}^{d_2}$, is simpler. There, the corresponding parametric form of f_S is $f_{\boldsymbol{\gamma}}^{*\nu}$. As long as $\boldsymbol{\gamma}$ is identifiable from the S_j's, the latter can be used to estimate $\boldsymbol{\gamma}$ by a standard approach such as maximum likelihood. Let $f_{\hat{\boldsymbol{\gamma}}}$ and $f_{\hat{\boldsymbol{\gamma}}}^{*\nu}$ denote the corresponding estimators of f_X and f_S, respectively. To find the maximum likelihood estimator of $\boldsymbol{\theta}$, we proceed as in the case where f_X and f_S are estimated nonparametrically, except that we replace the nonparametric estimators of f_X and f_S by $f_{\hat{\boldsymbol{\gamma}}}$ and $f_{\hat{\boldsymbol{\gamma}}}^{*\nu}$.

5 Theoretical properties

5.1 Smoothness classes

Estimating the regression curve p from aggregated and group testing data is a difficult inverse problem connected to the so-called deconvolution problem studied in

Fan (1991). In both cases, the covariates are indirectly observed. In our case, we can only observe $S_j = \sum_{i=1}^{\nu} X_{ij}$ whereas in the deconvolution case, we have access to data $W_{ij} = X_{ij} + U_{ij}$ observed with additive errors U_{ij}. The rates of convergence of nonparametric estimators in deconvolution problems are notoriously slow, and depend strongly on the rate of decay of the characteristic function ϕ_U of the U_{ij}'s. A usual distinction is made between cases where U is ordinary smooth or supersmooth. Similarly, the behaviour of our estimator depends on the behaviour of ϕ_X and ϕ_m and we make the same distinction between ordinary smooth and supersmooth classes.

A random variable X (resp., a function m) is ordinary smooth of order $\beta > 1$ (resp., $\kappa > 1$) if there exists a constant $C_1 > 1$ (resp., $C_2 > 1$) such that for every $t \in \mathbb{R}$,

$$C_1^{-1}(1+|t|)^{-\beta} \leq |\phi_X(t)| \leq C_1(1+|t|)^{-\beta} \tag{5.1}$$

$$\Big(\text{resp., } C_2^{-1}(1+|t|)^{-\kappa} \leq |\phi_m(t)| \leq C_2(1+|t|)^{-\kappa}\Big). \tag{5.2}$$

A random variable X (resp., a function m) is supersmooth of order $\rho > 0$ (resp., $\varrho > 0$) if there exist constants ρ_0, $\gamma > 0$ and $C_3 > 1$ (resp., ϱ_0, $\varsigma > 0$ and $C_4 > 1$) such that for every $t \in \mathbb{R}$,

$$C_3^{-1}(1+|t|)^{\rho_0}\exp(-|t|^{\rho}/\gamma) \leq |\phi_X(t)| \leq C_3(1+|t|)^{\rho_0}\exp(-|t|^{\rho}/\gamma) \tag{5.3}$$

$$\Big(\text{resp., } C_4^{-1}(1+|t|)^{\varrho_0}\exp(-|t|^{\varrho}/\varsigma) \leq |\phi_m(t)| \leq C_4(1+|t|)^{\varrho_0}\exp(-|t|^{\varrho}/\varsigma)\Big). \tag{5.4}$$

5.2 Asymptotic properties of the nonparametric estimator

In this section we derive theoretical properties of our estimator $\hat{p}$ defined in Section 3.1. The proofs are long and technical, and we defer most of them to a supplementary file. Similar results could be established for the estimator defined Section 3.2, but require even longer and more technical arguments. Therefore we leave them for

future work. In Appendix B.1, we recall the basic notations $\asymp$, $\sim$, $\vee$ and $\wedge$ used below.

Various sorts of smoothness combinations of m and f_X are possible. Here we consider the cases where f_X and m are both supersmooth or ordinary smooth. In the ordinary smooth case, we make the following regularity assumptions:

(Co1) (5.1) and (5.2) hold, $\sup_{t\in\mathbb{R}} |\phi'_X(t)\cdot t^{\beta+1}| < \infty$ and $\sup_{t\in\mathbb{R}} |\phi'_m(t)\cdot t^{\kappa+1}| < \infty$.

(Co2) K is symmetric with $\int K(x)\,dx = 1$, and for some $c_K < \infty$, $\phi_K(t) = 0$ for all $|t| > c_K$. Moreover, $\sup_{t\in\mathbb{R}} |\phi_K^{(\ell)}(t)| < \infty$ for $\ell = 0, 1$.

Condition (Co2) is fairly standard, especially in related deconvolution problems. It is satisfied by kernels traditionally employed there, such as the infinite order sinc kernel K whose Fourier transform is defined by $\phi_K(t) = I_{[-1,1]}(t)$, and the second order kernels K whose Fourier transform is given by $\phi_K(t) = (1-t^2)^q \cdot I_{[-1,1]}(t)$ for some positive integer q. Recall that a second order kernel is a kernel that satisfies $\mu_{K,0} = 1$, $\mu_{K,1} = 0$ and $0 < |\mu_{K,2}| < \infty$, where, for $\ell = 0, 1, \ldots$, we use the notation $\mu_{K,\ell} = \int_{-\infty}^{\infty} u^\ell K(u)\,du$.

The theorem below establishes asymptotic properties of our estimator $\hat{p}$ in the ordinary smooth case; see Section A.1 for its proof. We use the following notation: for f a function or a random variable and for h a bandwidth, let

$$B_{K,f}(x;h) = \frac{1}{2\pi}\int e^{-\mathrm{i}tx}\{\phi_K(th) - 1\}\phi_f(t)\,dt\ ,\ V_{K,f}(h) = \frac{1}{2\pi}\int \frac{|\phi_K(t)|^2}{|\phi_f(t/h)|^{2(\nu-1)}}\,dt.$$

Theorem 5.1. Assume that $f_X(x) > 0$, that conditions (Co1) and (Co2) hold and that $h = h_n \to 0$ and $nh^{2\nu\max(\beta,\kappa)+1/2} \to \infty$ as $n \to \infty$. Then $\hat{p}$ at (3.7) satisfies

$$\hat{p}(x) - p(x) = T_p(x) + o_P\big\{n^{-1/2}h^{-(\nu-1)(\beta\vee\kappa)-1/2}\big\}, \tag{5.5}$$

where $T_p(x)$ is a random variable such that

$$E\{T_p(x)\} = B_p(x), \quad \mathrm{Var}\{T_p(x)\} = V_p(x) := V_{p,1}(x)\{1 + o(1)\} + V_{p,2}(x),$$

with $B_p(x) = f_X^{-1}(x)\big\{q(x)B_{K,X}(x;h) - B_{K,m}(x;h)\big\}$ and

$$V_{p,1}(x) = \{nh\,\nu^2 f_X^2(x)\}^{-1}\big\{m^{*\nu}(x)V_{K,m}(h) + q^2(x)f_X^{*\nu}(x)V_{K,X}(h)\big\}, \tag{5.6}$$

$$V_{p,2}(x) = -\{nh\,\pi\nu^2 f_X^2(x)\}^{-1}q(x)m^{*\nu}(x)\int_{-\infty}^{\infty}\frac{|\phi_K(t)|^2}{\{\phi_X(t/h)\phi_m(-t/h)\}^{\nu-1}}\,dt. \tag{5.7}$$

In addition, if p and f_X are twice differentiable and their second derivative is α-Hölder continuous for some $0 < \alpha \le 1$, and K is such that $\mu_{K,1} = 0$ and $\int |u|^{2+\alpha}|K(u)|\,du < \infty$, then $B_p(x) = \big\{\frac{1}{2}p''(x) + p'(x)f_X'(x)f_X^{-1}(x)\big\}\mu_{K,2}\,h^2 + o(h^2)$.

Abusing terminology, we shall refer to B_p and V_p as the bias and variance of our estimator. Let $\mathcal{I}$ be a bounded interval such that $\inf_{x\in\mathcal{I}} f_X(x) > 0$ and $\sup_{x\in\mathcal{I}} m(x) < \infty$, and with slight abuse of terminology, define the mean integrated squared error (MISE) of our estimator on $\mathcal{I}$ by $\mathrm{MISE} = \int_{\mathcal{I}}\{B_p^2(x) + V_p(x)\}\,dx$. We deduce from the theorem that $\mathrm{MISE} = O(h^C) + O\{n^{-1}h^{-2(\kappa\vee\beta)(\nu-1)-1}\}$, where $C = 4$ if we use a second order kernel, and $C = 2(\beta\wedge\kappa) - 1$ if we use an infinite order kernel such as the sinc kernel or a kernel satisfying (F.1) in Section F. Taking a bandwidth of order $h \asymp n^{-1/\{2(\kappa\vee\beta)(\nu-1)+C+1\}}$, we get $\mathrm{MISE} = O[n^{-C/\{2(\kappa\vee\beta)(\nu-1)+C+1\}}]$.

The next theorem describes asymptotic properties of $\hat{p}$ in the supersmooth case. For brevity, here we assume that ρ and ϱ in (5.3) and (5.4) are both greater than or equal to 1. See Section F in the supplementary file, where we state and prove a version of the theorem (see Theorem F.1) without this restriction on ρ and ϱ. The proof of Theorem 5.2 is almost exactly identical to that of Theorem F.1.

Theorem 5.2. Assume that conditions (5.3), (5.4) and (Co2) hold, and that the bandwidth h satisfies $h = h_n = \max\big\{(d\,\gamma\log n)^{-1/\rho}, (d\,\varsigma\log n)^{-1/\varrho}\big\}$ for some $0 < d < (2\nu)^{-1}$. Then

$$\hat{p}(x) - p(x) = Q_p(x) + O_P\big\{(\log n)^{c_2}\,n^{-c_1}\big\}, \tag{5.8}$$

where $c_1 = \min\{2\tau^{\rho\vee\varrho}\, d,\, 1-(2\nu-1)d\}$, $Q_p(x)$ is a random variable satisfying $E\{Q_p(x)\} = f_X^{-1}(x)\big\{q(x)B_{K,X}(x;h) - B_{K,m}(x;h)\big\}$ and $\mathrm{Var}\{Q_p(x)\} = O\big\{n^{-1+2(\nu-1)d}\,(\log n)^{c_2}\big\}$, for some constant $c_2 > 0$ that does not depend of n.

It can be deduced from the theorem that, with the bandwidth stated there, the MISE of our estimator is dominated by the bias contribution, and is of order $O\Big(\max\big[\exp\{-\tau^{\varrho}(d\,\gamma\log n)^{\varrho/\rho}/\varsigma\}, \exp\{-\tau^{\rho}(d\,\gamma\log n)^{\rho/\varrho}/\gamma\}\big]\Big)$. In Section F, we show (Theorem F.1) that an improved rate can be achieved by using two bandwidths. However, it is not clear how these could be chosen effectively in practice.

Remark 5.1. (Asymptotic normality of $\hat{p}$). In Section E.4, under appropriate conditions, we prove that in the ordinary smooth case, we have

$$\{V_p(x)\}^{-1/2}\{\hat{p}(x) - p(x) - B_p(x)\} \xrightarrow{D} N(0,1) \quad \text{as } n\to\infty\,, \tag{5.9}$$

where $B_p(x)$ and $V_p(x)$ are as in Theorem 5.1. A similar result can be established in the supersmooth case, using arguments and technical conditions that are similar in spirit to those used by Fan (1991) and Delaigle et al. (2009).

5.3 Asymptotic properties of the parametric estimator

To establish asymptotic properties of the estimator $\hat{\boldsymbol{\theta}}$ defined in Section 4, let $p_0(x) = P(Y=1|X=x\,;\boldsymbol{\theta}_0)$, $m_0 = q_0 f_X = (1-p_0)f_X$, and for all $s\in\mathbb{R}$ and $\boldsymbol{\theta}\in\Theta$, let

$$\Lambda(s,\boldsymbol{\theta}) = \frac{1-2Y^*}{Y^*\, f_S(s) + (1-2Y^*)m_{\boldsymbol{\theta}}^{*\nu}(s)}\,\{m'_{\boldsymbol{\theta}} * m_{\boldsymbol{\theta}}^{*(\nu-1)}\}(s), \tag{5.10}$$

where $m'_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}}\, m_{\boldsymbol{\theta}}$, with $\nabla_{\boldsymbol{\theta}}$ the gradient. Put $Q_0(\boldsymbol{\theta}) = E\{\ln f(Y^*, S\,|\,\boldsymbol{\theta})\}$. The next theorem summarises the asymptotic properties of $\hat{\boldsymbol{\theta}}$. See Section A.2 for a proof.

Theorem 5.3. Assume that conditions (P1)–(P5) in Appendix B.2 hold, that $\boldsymbol{\theta}_0$ uniquely maximises $Q_0(\boldsymbol{\theta})$ subject to $\boldsymbol{\theta}\in\Theta$, and that $\boldsymbol{\theta}_0$ is an interior point of Θ. Then, as long as the constant c_0 in (4.3) is small enough, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{D} N\big(0, \boldsymbol{\Sigma}_0^{-1}\big)$, where $\boldsymbol{\Sigma}_0 = E\{\Lambda(S,\boldsymbol{\theta}_0)\Lambda(S,\boldsymbol{\theta}_0)^{\mathrm{T}}\}$.

6 Numerical properties

6.1 Bandwidth selection for the nonparametric estimator

Choosing a good data-driven bandwidth in our context is particularly difficult. In standard regression problems, the most popular technique is arguably the plug-in approach, where the bandwidth is chosen by minimising an estimator of $\int_a^b (B^2 + V) f_X$, with B and V denoting the dominating parts of the "bias" and "variance" of the regression estimator, and $[a, b]$ typically denoting quantiles of f_X. In our case, the "bias" and "variance" of our estimators take complex forms involving unknown quantities that are difficult to estimate in practice, and thus computing a plug-in bandwidth seems too challenging to be practical. A practical bandwidth that is often easier to compute is the cross-validation bandwidth. However, to compute this bandwidth we need to calculate the regression estimator at the X_{ij}'s, which is not feasible in our case since we observe only indirect data. With this in mind, our goal is not to construct a consistent estimator of the optimal bandwidth, but rather to suggest a bandwidth that can be expected to give reasonable practical results, as we explain next. Here we provide the details for the local linear estimator, which this is the estimator we recommend using in practice, but the same ideas can be used for the local polynomial estimator of another order.

Let $\hat{p}$ denote our local linear estimator of p computed from the grouped data (S_j, Z_j^*), $j = 1, \ldots, n$. We wish to construct an approximation to the bandwidth h_0 that minimises the distance $D = \int_a^b (\hat{p} - p)^2 f_X$, where a and b denote the 0.05 and 0.95 quantiles of the distribution of X. As argued above, estimating D directly would be too challenging. Instead, we compute another bandwidth h_1, which is appropriate for estimating $p^*(s) = E(Z_j^* | S_j = s)$ by the standard local linear estimator computed from the data (S_j, Z_j^*), $j = 1, \ldots, n$. Of course, h_1 is not necessarily a good approximation of h_0. However, if we could construct a reasonable approximation $\hat{c}$ of

$c = h_0/h_1$, then we could approximate h_0 by $\hat{h}_0 = \hat{c}\, h_1$.

We propose to approximate c by $\hat{c} = h_0^*/h_1^*$, where h_0^* is a version of h_0 constructed from artificial data $(\tilde{X}_{ij}, \tilde{Y}_{ij})$, $i = 1, \ldots, \nu$, $j = 1, \ldots, n$, generated under a parametric model for f_X and p, and where h_1^* is a version of h_1 constructed from data $(\tilde{S}_j, \tilde{Y}_j^*)$, $j = 1, \ldots, n$ obtained by randomly pooling the $(\tilde{X}_{ij}, \tilde{Y}_{ij})$'s in groups of size ν, in the same way as the original data were pooled. Of course, this parametric model is usually wrong as in principle we do not know the correct parametric model (otherwise we would not estimate p nonparametrically). In particular, as for the SIMEX types of bandwidths suggested by Delaigle and Meister (2007) and Delaigle and Hall (2008), $\hat{c}$ is not usually a consistent estimator of c. However, and paraphrasing Delaigle and Hall (2008), since the $(\tilde{S}_j, \tilde{Y}_j^*)$'s measure the (S_j, Y_j^*)'s in the same way as the $(\tilde{X}_{ij}, \tilde{Y}_{ij})$'s measure the (X_{ij}, Y_{ij})'s, we can expect that the relationship between h_0 and h_1 is well approximated by that between h_0^* and h_1^*.

To implement these ideas in practice, we proceed as follows:

1. Compute h_1, the standard plug-in bandwidth for the standard local linear estimator of $p^*(s) = E(Z_j^*|S_j = s)$, computed using the data (S_j, Z_j^*), $j = 1, \ldots, n$.
2. Let $\hat{\mu}_X = \hat{\mu}_S/\nu$ and $\hat{\sigma}_X = \hat{\sigma}_S/\sqrt{\nu}$, where $\hat{\mu}_S$ and $\hat{\sigma}_S$ denote the empirical mean and standard deviation of the S_j's, and let $\tilde{p}_0$ denote a quadratic spline estimator of p computed with $K = 2$ knots located at the quantiles $(k+1)/(K+2)$, $k = 1, 2$, of $\tilde{f}_X$, where $\tilde{f}_X$ is the density of $N(\hat{\mu}_X, \hat{\sigma}_X^2)$. For $i = 1, \ldots, \nu$, $j = 1, \ldots, n$, generate $(\tilde{X}_{ij}, \tilde{Y}_{ij})$ by taking $\tilde{X}_{ij} \sim N(\hat{\mu}_X, \hat{\sigma}_X^2)$ and $\tilde{Y}_{ij}|\tilde{X}_{ij} \sim$Bernoulli$\{\tilde{p}_0(\tilde{X}_{ij})\}$.
3. For $j = 1, \ldots, n$, let $\tilde{Y}_j^* = \max_{i=1,\ldots,\nu} \tilde{Y}_{ij}$, $\tilde{S}_j = \sum_{i=1}^{\nu} \tilde{X}_{ij}$, and put $\tilde{Z}_j^* = 1 - \tilde{Y}_j^*$.
4. Let h_0^* denote the bandwidth that minimises $\int_a^b (\hat{p}_0 - \tilde{p}_0)^2 \tilde{f}_X$, where $\hat{p}_0$ denotes the version of $\hat{p}$ computed from the data $(\tilde{S}_j, \tilde{Z}_j^*)$, $j = 1, \ldots, n$, and where a and b are the 0.05 and 0.95 quantiles of the $N(\hat{\mu}_X, \hat{\sigma}_X^2)$ distribution, respectively.
5. Let h_1^* denote the standard plug-in bandwidth for the standard local linear estimator of $\tilde{p}^*(s) = E(\tilde{Z}_j^*|\tilde{S}_j = s)$, computed using the data $(\tilde{S}_j, \tilde{Z}_j^*)$, $j = 1, \ldots, n$.
6. Let $\hat{c} = h_0^*/h_1^*$ and take $\hat{h}_0 = \hat{c}\, h_1$.

As for the SIMEX bandwidth of Delaigle and Hall (2008), to avoid too strong a dependence of h_0^* and h_1^* on the sample $(\tilde{X}_{ij}, \tilde{Y}_{ij})$, we generate $B = 500$ such samples,

thereby obtaining B pairs of bandwidths $(h^*_{1,b}, h^*_{2,b})$, for $b = 1, \ldots, B$, and we compute h^*_1 and h^*_2 by taking $h^*_1 = B^{-1} \sum_{b=1}^{B} h^*_{1,b}$ and $h^*_0 = B^{-1} \sum_{b=1}^{B} h^*_{0,b}$. We conclude this section by a few remarks regarding implementation. First, our choice of locating the knots at quantiles $(k+1)/(K+2)$ is borrowed from Ruppert et al. (2003), page 126. Second, to implement step 4, we do a grid search on the interval $[h^*_1/\sqrt{\nu}, 3h^*_1]$. Third, when fitting the quadratic spline in step 2, we used the method of Section 4 with f_X estimated nonparametrically.

6.2 Simulations

We illustrate the performance of our parametric and nonparametric estimators of p via the following simulated examples:

(i) $p(x) = e^{-5+1.4\,x}/(1 + e^{-5+1.4\,x})$ and $X \sim N(2, 9/16)$.

(ii) $p(x) = 0.5\, e^{-6+2.5\,x}/(1 + e^{-6+2.5\,x})$ and $X \sim N(2, 9/16)$.

(iii) $p(x) = \min\{1, \max(0, 0.03 - 0.05\,x + 0.04\,x^2)\}$, $X = 4 - Z/4$ and $Z \sim \chi^2(8)$.

(iv) $p(x) = 0.25\{m_1(x)\}^{0.25} \cdot 1\{x \le 3\} + m_2(x) \cdot 1\{x > 3\}$, where $m_1(x) = 0.2\,\phi_{2.5,0.4}(x) + \phi_{3,0.225}(x) + \phi_{3.5,0.375}(x) + \phi_{4,0.5}(x)$, with $\phi_{\mu,\sigma}$ the density of a $N(\mu, \sigma^2)$, $m_2(x) = (x-3)^2/2 + 0.3115$, and $X \sim f_X = 0.35\,\phi_{1.5,0.45/\sqrt{2}} + 0.65\,\phi_{1.75,1/\sqrt{2}}$.

We applied the parametric estimator of Section 4 and the local linear estimator (local polynomial estimator from Section 3.2 with $\ell = 1$) to data simulated from the above four models, where we grouped the data in groups of size $\nu = 4$ and $\nu = 8$, for N ranging from $2,000$ to $10,000$. As suggested by a referee, in the parametric case we used parametric models $p_{\boldsymbol{\theta}}$ and $f_{\boldsymbol{\gamma}}$ for both p and f_X, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^T \in \mathbb{R}^d$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{d_2})^T \in \mathbb{R}^{d_2}$, for some $d, d_2 \ge 1$, and where we fitted $\boldsymbol{\gamma}$ by maximum likelihood (see Remark 4.1).

For the local linear estimator, we chose the bandwidth as in Section 6.1, and used a kernel often employed in the deconvolution literature, defined by $\phi_K(t) =$

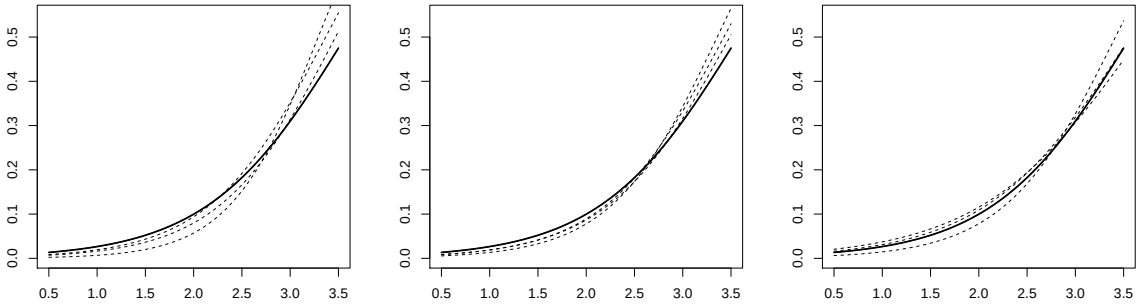

Figure 1: Parametric estimator of p for three samples coming from model (i) with $\nu = 8$ and $N = 2,000$ (left), $N = 5,000$ (middle) and $N = 10,000$ (right), and corresponding to the 1st, 2nd and 3rd quartiles of the ISEs. The continuous line depicts the true p.

$(1-t^2)^3 \cdot I_{[-1,1]}(t)$. To illustrate the usefulness of a nonparametric estimator, we considered correctly and incorrectly specified parametric models for p and f_X. In case (i) we used the correct model $p_{\boldsymbol{\theta}}(x) = e^{\theta_1+\theta_2 x}/(1+e^{\theta_1+\theta_2 x})$ and $f_{\boldsymbol{\gamma}} = \phi_{\gamma_1,\gamma_2}$. In case (ii) we used either the correct model $p_{\boldsymbol{\theta}}(x) = \theta_3\, e^{\theta_1+\theta_2 x}/(1+e^{\theta_1+\theta_2 x})$ and $f_{\boldsymbol{\gamma}} = \phi_{\gamma_1,\gamma_2}$ (case (ii.a)) or the model from (i), whose $p_{\boldsymbol{\theta}}$ is thus incorrect (case (ii.b)). In case (iii), we used the correct model $p_{\boldsymbol{\theta}}(x) = \min\{1, \max(0, \theta_1 + \theta_2\, x + \theta_3\, x^2)\}$, and for $f_{\boldsymbol{\gamma}}$ we either used the correct parametric model (case (iii.a)) for which $X = \gamma_1 + \gamma_2\, \chi^2(\gamma_3)$, or an incorrect parametric model (case (iii.b)), where we assumed that X had a Laplace(γ_1) distribution. In case (iv), we used the models from case (i) for p and f_X. Since p is a probability curve, in each case we truncated all estimators to the range $[0, 1]$.

We know from Delaigle and Meister (2011) that grouping the Y_{ij}'s does not affect the convergence rates of nonparametric estimators, and in our context of grouped Y_{ij}'s and aggregated X_{ij}'s, it is the aggregation of the X_{ij}'s that causes the deterioration of convergence rates of nonparametric estimators. To illustrate this in practice, in each case we also computed the local linear estimator of Delaigle and Meister (2011) (denoted below by DM) using the grouped but non-aggregated data (X_{ij}, Z_j^*), and Delaigle and Meister's (2011) plug-in bandwidth. We also computed the standard local linear estimator (denoted below by LL) using the ideal individual data (X_{ij}, Y_{ij})

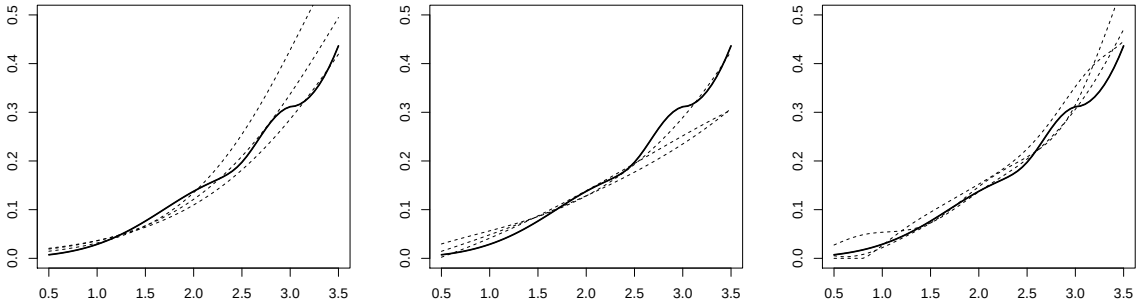

Figure 2: Parametric (left) and nonparametric (middle) estimator of p based on the (S_j, Z_j^*)'s, and Delaigle and Meister's (2011) estimator of p based on the (X_{ij}, Z_j^*)'s (right) for three samples coming from model (iv) with $N = 5,000$ and $\nu = 4$, and corresponding to the 1st, 2nd and 3rd quartiles of the ISEs. The continuous line depicts the true p.

and a standard plug-in bandwidth.

We simulated $1,000$ samples from each model, for each ν and each N. We present a summary of the results in Table 1. For each case shown in the table and each estimator p^{est} of p, we provide the median and the interquartile range of the $1,000$ values of the integrated squared error $\text{ISE} = \int_{0.5}^{3.5}(p^{\text{est}} - p)^2$. Overall the results reflect our theoretical properties: as ν decreases or N increases, the results improve, the correctly specified parametric estimator often works better than the nonparametric estimator, but the latter usually works better than the parametric estimator with an incorrectly specified $p_{\boldsymbol{\theta}}$ or $f_{\boldsymbol{\gamma}}$, such as in cases (ii.b), (iii.b) and (iv). Comparing our new local linear estimator with the LL and DM estimators, we can see that in most cases, as expected by the theory, grouping the data has less impact than aggregating the data. In particular, as N increases, the DM estimator, which is computed from the (X_{ij}, Z_j^*)'s, improves faster than the new estimator, which is computed from the (S_j, Z_j^*)'s.

The results are illustrated in Figures 1 to 3, as well as in Figure 5 in Section 7.1. In each graph, for a given estimator we show three estimated curves computed from the three samples that gave the first, second and third quartiles of the $1,000$ ISEs.

Table 1: Simulation results for models (i) to (iv). The numbers show $10^3\times$ Median integrated squared error (interquartile range) calculated from $1,000$ simulated samples, using our parametric (P) or our nonparametric (NP) estimators computed from the (S_j, Z_j^*)'s, Delaigle and Meister's (2011) estimator (DM) computed from the (X_{ij}, Z_j^*)'s or the standard local linear estimator (LL) computed from the (X_{ij}, Y_{ij})'s.

		$\nu = 4$				$\nu = 8$			
Model	N	P	NP	DM	LL	P	NP	DM	LL
(i)	$2 \cdot 10^3$	1.00[2.09]	4.39[2.85]	2.58[3.00]	0.81[0.95]	2.57[6.18]	5.18[4.51]	5.58[7.30]	0.81[0.95]
	$5 \cdot 10^3$	0.39[0.86]	3.90[2.09]	1.23[1.17]	0.38[0.43]	1.12[2.53]	4.55[2.85]	2.43[2.87]	0.38[0.43]
	10^4	0.22[0.49]	3.87[1.79]	0.67[0.59]	0.22[0.24]	0.51[1.37]	4.57[2.14]	1.44[1.55]	0.22[0.24]
(ii.a)	$2 \cdot 10^3$	4.34[4.29]	3.15[3.35]	3.20[3.84]	1.01[1.16]	6.95[9.77]	3.96[5.68]	7.64[10.7]	1.01[1.16]
	$5 \cdot 10^3$	2.84[3.21]	2.54[2.11]	1.45[1.61]	0.46[0.47]	4.63[5.10]	3.65[3.87]	3.33[4.20]	0.46[0.47]
	10^4	1.93[2.70]	2.32[1.61]	0.82[0.78]	0.25[0.27]	3.85[3.34]	3.33[2.92]	1.94[2.15]	0.25[0.27]
(ii.b)	$2 \cdot 10^3$	4.46[3.20]	3.15[3.35]	3.20[3.84]	1.01[1.16]	7.06[9.22]	3.96[5.68]	7.64[10.7]	1.01[1.16]
	$5 \cdot 10^3$	3.77[1.55]	2.54[2.11]	1.45[1.61]	0.46[0.47]	4.90[4.18]	3.65[3.87]	3.33[4.20]	0.46[0.47]
	10^4	3.58[1.07]	2.32[1.61]	0.82[0.78]	0.25[0.27]	3.98[2.18]	3.33[2.92]	1.94[2.15]	0.25[0.27]
(iii.a)	$2 \cdot 10^3$	1.97[3.88]	4.71[4.10]	1.93[2.24]	0.57[0.63]	7.43[15.7]	6.10[6.30]	3.99[4.93]	0.57[0.63]
	$5 \cdot 10^3$	0.64[1.27]	3.87[3.09]	0.83[0.85]	0.24[0.28]	3.01[5.74]	5.77[4.34]	1.85[2.07]	0.24[0.28]
	10^4	0.31[0.62]	3.05[2.52]	0.52[0.47]	0.14[0.14]	1.49[3.11]	5.52[3.33]	1.03[1.13]	0.14[0.14]
(iii.b)	$2 \cdot 10^3$	4.57[3.85]	4.71[4.10]	1.93[2.24]	0.57[0.63]	7.73[7.67]	6.10[6.30]	3.99[4.93]	0.57[0.63]
	$5 \cdot 10^3$	3.95[2.00]	3.87[3.09]	0.83[0.85]	0.24[0.28]	5.90[4.28]	5.77[4.34]	1.85[2.07]	0.24[0.28]
	10^4	3.90[1.66]	3.05[2.52]	0.52[0.47]	0.14[0.14]	5.15[2.88]	5.52[3.33]	1.03[1.13]	0.14[0.14]
(iv)	$2 \cdot 10^3$	5.75[13.0]	4.65[5.88]	4.80[7.38]	2.88[4.73]	9.96[27.0]	5.93[8.39]	7.95[12.9]	2.88[4.73]
	$5 \cdot 10^3$	4.11[7.25]	3.91[4.75]	2.28[3.49]	1.48[2.16]	5.65[12.3]	5.13[5.85]	3.72[5.58]	1.48[2.16]
	10^4	4.23[5.75]	3.44[3.59]	1.25[1.69]	0.83[1.14]	5.20[9.80]	5.00[4.03]	2.11[3.03]	0.83[1.14]

Figure 1 shows these curves in case (i) for the parametric estimator, when $\nu = 8$ and $N = 2,000$ to $10,000$; it illustrates the good properties of our parametric estimator. In Figure 2, we depict the estimated curves for our (misspecified) parametric estimator and our nonparametric estimator computed from the data (S_j, Z_j^*), in case (iv). To illustrate the negative impact that aggregating the X_{ij}'s has on estimators, we also show the curves for the DM estimator, computed from the non-aggregated (X_{ij}, Z_j^*)'s. Next, we illustrate the impact of fitting a wrong parametric model. In Figure 3, we compare our parametric estimator for cases (ii.a) and (ii.b) with our nonparametric estimator. In this example, the true curve is a logistic curve divided by two, and takes values between 0 and 0.5, whereas the incorrect parametric model uses a logistic curve which takes values between 0 and 1. As a result, the incorrectly specified parametric estimator has difficulties to recover the right-most part of the curve, but is not too

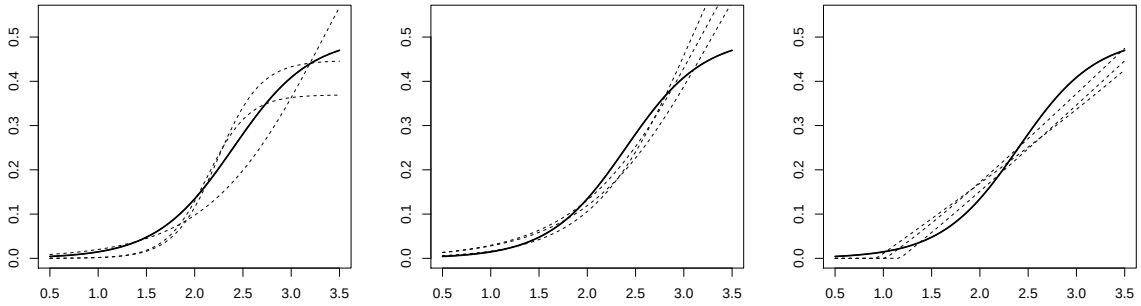

Figure 3: Correctly (left) and incorrectly (middle) specified parametric estimator, and nonparametric estimator of p (right) based on the (S_j, Z_j^*)'s, for three samples coming from model (ii) with $N = 2,000$ and $\nu = 4$, and corresponding to the 1st, 2nd and 3rd quartiles of the ISEs. The continuous line depicts the true p.

bad elsewhere.

6.3 Illustration with real data

To illustrate the performance of our estimators on real data, we applied them to data from the National Health and Nutrition Examination Surveys (NHANES), collected between 1999 and 2000. These data are non-grouped, which is convenient to illustrate the effect that grouping has on estimators. We note that our analysis of these data is purely illustrative, and as others before, we ignore issues such as sampling weights. In this example, we take Y to be the indicator of the presence of hepatitis B core antibody in the patient's serum or plasma, and X to be the logarithm of the level of gamma glutamyl transaminase (GGT), a biomarker for liver disease (elevated levels of GGT are typical for patients with hepatitis).

Before grouping the data, we computed the standard local linear estimator of p with a standard plug-in bandwidth. We denote this estimator by $\hat{p}_0$. Then we randomly grouped the data in groups of sizes $\nu = 4$ and $\nu = 8$. We repeated this 200 times, creating in this way 200 grouped samples of (S_j, Y_j^*)'s of each size ν. We applied to those data our local linear estimator from Section 3.2 and our parametric

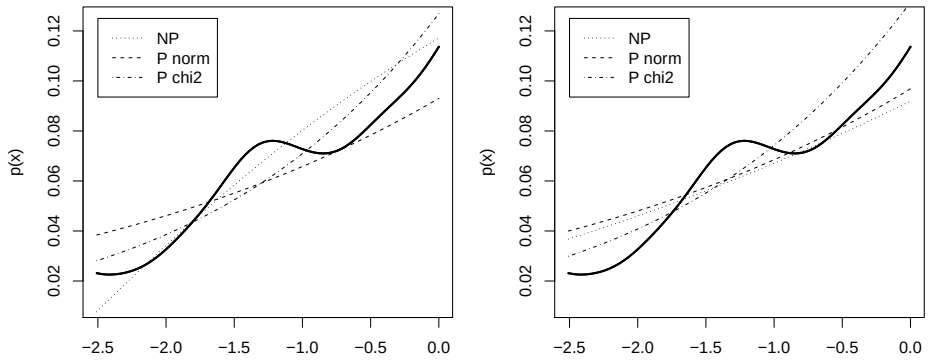

Figure 4: Nonparametric (NP) estimator of p and parametric estimator of p assuming that $X \sim N(\gamma_1, \gamma_2)$ (P norm) or $X \sim \gamma_1 + \gamma_2\, \chi^2(\gamma_3)$ (P chi2), in the NHANES examples when $\nu = 4$ (left) or $\nu = 8$ (right). The thick line depicts the estimator $\hat{p}_0$ computed from the non-grouped data.

estimator from Section 4. For the parametric model for p, we used a logistic curve as in example (i) in Section 6.2, and for f_X we considered two models: $X \sim \gamma_1 + \gamma_2\, \chi^2(\gamma_3)$ and $X \sim N(\gamma_1, \gamma_2)$. We estimated all parameters as in Section 6.2.

As in Section 6.2, for each ν, we computed 200 values of the integrated squared error, which, for each estimator p^{est}, we define here as ISE $= \int_a^b (p^{\text{est}} - \hat{p}_0)^2$, where a and b are the empirical 0.025 and 0.975 quantiles of X, respectively. In Figure 4 we show, for $\nu = 4$ and $\nu = 8$, the curve $\hat{p}_0$ together with the parametric estimator for $X \sim \gamma_1 + \gamma_2\, \chi^2(\gamma_3)$ and for $X \sim N(\gamma_1, \gamma_2)$ and the nonparametric estimator, corresponding in each case to the median ISE value. Of course here we do not know the true curve p, but we can see that, given the difficulty of estimating curves from grouped and aggregated data, the estimated curves are relatively close to the nonparametric estimator $\hat{p}_0$ computed from the non-grouped data, indicating that in this example our estimators worked reasonably well. While the logistic model is a reasonable approximation in this case, the local linear estimator has a smaller median ISE than the two versions of the parametric estimator.

7 Extensions

7.1 Imperfect tests

In real applications, the group status Y_j^* can be observed imperfectly. For example, this is often the case if Y_j^* is obtained through a blood test. Two types of errors may arise: a negative group is declared positive, and a positive group is declared negative. In this case, the procedures developed in the previous sections are not consistent and need to be modified to take the errors into account. If we let $\tilde{Y}_j^*$ denote the observed status of the jth group, then the accuracy of the test is measured by the specificity Sp and the sensitivity Se, which are defined by $\text{Sp} = P(\tilde{Y}_j^* = 0 | Y_j^* = 0)$, $\text{Se} = P(\tilde{Y}_j^* = 1 | Y_j^* = 1)$. As often in the literature (see, e.g. Vansteelandt et al., 2000), we assume that the observed status $\tilde{Y}_j^*$ depends only on the true status Y_j^*, and depends neither on the group size nor on X.

It follows from our calculations in Section 3.2 that, to construct a local polynomial estimator of p, we need consistent estimators of ϕ_X and ϕ_m. Since the S_j^*'s are not affected by the imperfectly observed Y_j^*'s, we can define $\hat{\phi}_X$ as in Section 3.1. To define a consistent estimator of ϕ_m, let $\tilde{Z}_j^* = 1 - \tilde{Y}_j^*$ and $\tilde{m}(x) = P(\tilde{Y} = 0 | X = x) f_X(x)$. It can be proved (see Appendix C.3) that

$$\begin{aligned}\phi_{\tilde{m}^{*\nu}}(t) :=& \int_{-\infty}^{\infty} e^{\mathrm{i}tx} P(\tilde{Z}_j^* = 1 | S_j = x) f_S(x)\, dx \\ =&(\text{Se} + \text{Sp} - 1)\phi_{m^{*\nu}}(t) + (1 - \text{Se})\phi_S(t). \end{aligned} \tag{7.1}$$

We can estimate $\phi_{\tilde{m}^{*\nu}}(t)$ consistently by $\hat{\phi}_{\tilde{m}^{*\nu}}(t) = n^{-1} \sum_{j=1}^n \tilde{Z}_j^* \exp(\mathrm{i}tS_j)$. Taking $\hat{\phi}_S$ as in (3.3), we deduce that $\phi_{m^{*\nu}}(t)$ can be estimated consistently by

$$\hat{\phi}_{m^{*\nu}}^{\text{corr}}(t) = (\text{Se} + \text{Sp} - 1)^{-1} \hat{\phi}_{\tilde{m}^{*\nu}}(t) - (1 - \text{Se})(\text{Se} + \text{Sp} - 1)^{-1} \hat{\phi}_S(t).$$

Proceeding as in Section 3.1, we deduce from there a consistent estimator $\hat{\phi}_m^{\text{corr}}$ of ϕ_m.

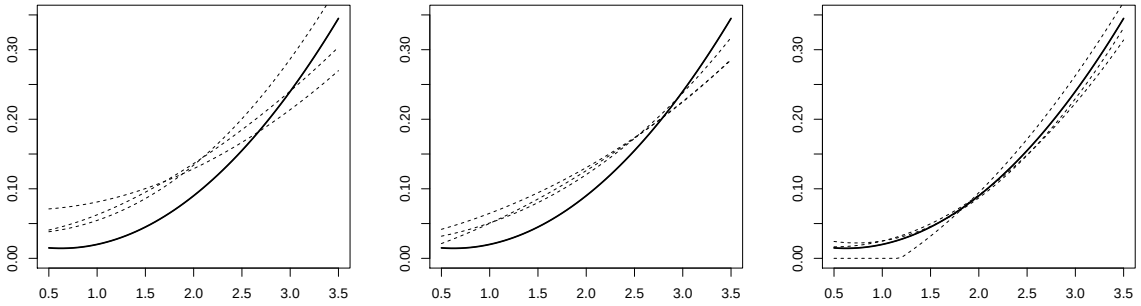

Figure 5: Estimators computed from imperfectly observed data. Uncorrected nonparametric estimator (left), corrected nonparametric estimator (middle) and corrected correctly specified parametric estimator (right) of p for three samples coming from model (iii.a) with $N = 10,000$ and $\nu = 4$, and corresponding to the 1st, 2nd and 3rd quartiles of the ISEs. The continuous line depicts the true p.

Using calculations similar to those leading to (3.10), we can define a local polynomial estimator of p, by

$$\hat{p}^{\,\mathrm{corr}} = 1 - \hat{q}^{\,\mathrm{corr}} \quad \text{with} \quad \hat{q}^{\,\mathrm{corr}} = \mathbf{e}_1^{\mathrm{T}}\, \hat{\mathbf{S}}_N^{-1}\, \hat{\mathbf{U}}_N^{\mathrm{corr}}, \tag{7.2}$$

where $\hat{\mathbf{S}}_N$ as in Section 3.2 and $\hat{\mathbf{U}}_N^{\mathrm{corr}} = (\hat{U}_{N,0}^{\mathrm{corr}}, \ldots, \hat{U}_{N,\ell}^{\mathrm{corr}})^{\mathrm{T}}$, with

$$\hat{U}_{N,k}^{\mathrm{corr}}(x) = (2\pi \mathrm{i}^k)^{-1} \int e^{-\mathrm{i}tx}\, \hat{\phi}_m^{\mathrm{corr}}(t) \phi_K^{(k)}(-ht)\, dt.$$

In the parametric case, we show in Appendix C.3 that the conditional likelihood of the $\{\tilde{Y}_j^* | S_j\}$'s is given by

$$\prod_{j=1}^{n} \left\{ (\mathrm{Se} + \mathrm{Sp} - 1) \frac{m_{\boldsymbol{\theta}}^{*\nu}(S_j)}{f_S(S_j)} + (1 - \mathrm{Se}) \right\}^{1-\tilde{Y}_j^*} \left\{ (1 - \mathrm{Sp} - \mathrm{Se}) \frac{m_{\boldsymbol{\theta}}^{*\nu}(S_j)}{f_S(S_j)} + \mathrm{Se} \right\}^{\tilde{Y}_j^*}. \tag{7.3}$$

Therefore, in that case, we can consistently estimate $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}^{\mathrm{corr}}$, obtained by maximising (7.3).

To illustrate the finite sample performance of these estimators, we applied them to $1,000$ samples of sizes $N = 2,000$, $5,000$ and $10,000$ generated from models (i), (ii.a) and (iii.a) from Section 6.2. We created pools of size $\nu = 4$, and introduced errors in

Table 2: Simulation results for models (i), (ii.a) and (iii.a) in the case of imperfectly observed Y_j^*'s. The numbers show $10^3\times$ Median integrated squared error (interquartile range) calculated from $1,000$ simulated samples, using the corrected parametric (CP), uncorrected parametric (UCP), corrected nonparametric (CNP) or uncorrected nonparametric (UCNP) estimators computed from grouped and aggregated data.

Model	N	CP	UCP	CNP	UCNP
(i)	$2\cdot 10^3$	1.21[2.66]	1.86[2.60]	4.57[2.97]	5.55[2.53]
	$5\cdot 10^3$	0.50[1.03]	1.23[1.19]	4.00[1.99]	4.94[1.68]
	10^4	0.26[0.56]	0.94[0.66]	3.95[1.64]	4.76[1.40]
(ii.a)	$2\cdot 10^3$	4.55[4.63]	5.01[5.16]	3.23[3.69]	4.42[3.69]
	$5\cdot 10^3$	3.06[2.94]	3.35[3.67]	2.60[2.28]	3.74[2.17]
	10^4	2.03[2.71]	1.96[2.51]	2.33[1.65]	3.39[1.42]
(iii.a)	$2\cdot 10^3$	2.56[4.69]	3.02[4.65]	4.47[4.29]	6.94[4.69]
	$5\cdot 10^3$	0.82[1.70]	1.72[1.73]	3.43[3.04]	5.71[3.29]
	10^4	0.41[0.78]	1.27[0.96]	2.94[2.33]	5.32[2.47]

the Y_j^*'s as above, in such a way that Se $= 0.996$ and Sp $= 0.923$, as in Xie (2001). We compared four estimators: the corrected local linear estimator ($\hat{p}^{\,\mathrm{corr}}$ in (7.2) with $\ell = 1$), the corrected parametric estimator $m_{\hat{\boldsymbol{\theta}}^{\mathrm{corr}}}$, the uncorrected local linear estimator, that is $\hat{p}$ in (3.10) with $\ell = 1$, but with the Y_j^*'s replaced by their imperfect version $\tilde{Y}_j^*$, and the uncorrected parametric estimator $m_{\hat{\boldsymbol{\theta}}}$, computed as in Section 4, but with the Y_j^*'s replaced by the $\tilde{Y}_j^*$'s. As in Section 6.2, we measured the performance of each estimator through the ISE. The results, presented in Table 2 show that the ISE of the corrected estimators is smaller than that of the uncorrected estimators. Unsurprisingly, the results improve as sample size increases and the parametric estimator, which is computed under the correct parametric model in all three cases, often outperforms the nonparametric one. See also Figure 5, where, in case (iii.a), we show estimated curves obtained using the uncorrected nonparametric estimator, the corrected nonparametric estimator and the corrected parametric estimator.

7.2 Nonparametric estimator when the individual Y_{ij}'s are observed

There are many applications where the individual binary Y_{ij}'s are observed, and only the X_{ij}'s are aggregated. There, we observe data (S_j, Y_{ij}), for $i = 1, \ldots, \nu$ and $j = 1, \ldots, n$. This problem was studied in the parametric context by Zhang and Albert (2011), who also discuss interesting epidemiological applications. We can exploit our ideas to derive a nonparametric estimator in this case too, as follows.

Using calculations similar to those in Section 2.2, it can be proved that

$$P(Y_{ij} = 1, S_j \leq x) = \int_{-\infty}^{\infty} P\{S_j^{(i)} \leq x - u\} p(u) f_X(u)\, du\,,$$

where $S_j^{(i)} = S_j - X_{ij}$. This implies that $f_S(x)P(Y_{ij} = 1|S_j = x) = (pf_X) * f_X^{*(\nu-1)}(x)\,.$ Taking the Fourier transform on both sides of this equation we deduce that

$$\phi_{SY}(t) \equiv \int e^{\mathrm{i}tx} f_S(x) P(Y_{ij} = 1|S_j = x)\, dx = \phi_{pf_X}(t)\, \phi_X^{\nu-1}(t)\,.$$

Now, calculations similar to those from Section 3.1 lead to $\phi_{SY}(t) = E(e^{\mathrm{i}tS_j} Y_{ij})$, which can be estimated by $\hat{\phi}_{SY}(t) = N^{-1} \sum_{i,j} e^{\mathrm{i}tS_j} Y_{ij}$. Moreover we can estimate $\phi_X^{\nu-1}$ by $(\hat{\phi}_X)^{\nu-1}$, with $\hat{\phi}_X$ as in Section 3.1. From there, we can estimate $(pf_X)(x)$ by

$$\widehat{(pf_X)}(x) = (2\pi)^{-1} \int e^{-\mathrm{i}tx} \hat{\phi}_{SY}(t) \hat{\phi}_X^{1-\nu}(t)\, \phi_K(ht) \cdot 1\{|\hat{\phi}_X(t)| > t_n\}\, dt\,.$$

Finally, we can estimate $p(x)$ by $\hat{p}(x) = \widehat{(pf_X)}(x)/\hat{f}_X(x)\,.$

Theoretical properties of this estimator can be derived along the lines of the theory from Section 5.2. As already noted in Section 6.2, aggregating the X_{ij}'s has much more impact on nonparametric estimators than grouping the Y_{ij}'s does (recall from Delaigle and Meister, 2011, that pooling the Y_{ij}'s does not impact the convergence rates of nonparametric estimators). It is therefore unsurprising that the estimator derived in the previous paragraph has rates similar to those in Section 5.2.

8 Supplemental Materials

Supplemental materials are submitted together with this manuscript and are intended to be for online publication only. Some technical notations and conditions can be found in Section B of the file. Details of the methodology are gathered in Section C. Long and technical arguments used to prove Theorem 5.1 can be found in Section D, and the proof of a more complex version of Theorem 5.2 is given in Section E.

Acknowledgement

This research was supported by a grant and a fellowship from the Australian Research Council. The NHANES data are available from the NHANES website of the Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, MD, USA: `http://wwwn.cdc.gov/nchs/nhanes/search/nhanes99_00.aspx`.

References

[1] Bilder, C.R. and Tebbs, J.M. (2009). Bias, efficiency, and agreement for group-testing regression models. *J. Statist. Comput. Simul.* **79**, 67–80.

[2] Caudill, S.P. (2010). Characterizing populations of individuals using pooled samples. *J. Expo. Sci. Environ. Epidemiol.* **20**, 29–37.

[3] Chen, C.L. and Swallow, W.H. (1990). Using group testing to estimate a proportion, and to test the binomial model. *Biometrics* **46**, 1035–1046.

[4] Chen, P., Tebbs, J.M. and Bilder, C.R. (2009). Group testing regression models with fixed and random effects. *Biometrics* **65**, 1270–1278.

[5] Delaigle, A., Fan, J. and Carroll, R.J. (2009). A design-adaptive local polynomial estimator for the errors-in-variables problem. *J. Amer. Statist. Assoc.* **104**, 348–359.

[6] Delaigle, A. and Hall, P. (2008). Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *J. Amer. Statist. Assoc.* **103**, 280–287.

[7] Delaigle, A. and Hall, P. (2012). Nonparametric regression with homogeneous group testing data. *Ann. Statist.* **40**, 131–158.

[8] Delaigle, A., Hall, P. and Wishart, J. (2014). New approaches to non- and semi-parametric regression for univariate and multivariate group testing data. *Biometrika* **101**, 567–585.

[9] Delaigle, A. and Meister, A. (2007). Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *J. Amer. Statist. Assoc.* **102**, 1416–1426.

[10] Delaigle, A. and Meister, A. (2011). Nonparametric regression analysis for group testing data. *J. Amer. Statist. Assoc.* **106**, 640–650.

[11] Dorfman, R. (1943). The detection of defective members of large populations. *Ann. Math. Statist.* **14**, 436–440.

[12] Fan, J. (1991). Asymptotic normality for deconvolution kernel density estimators. *Sankhyā Ser. A* **53**, 97–110.

[13] Faraggi, D., Reiser, B. and Schisterman, E.F. (2003). ROC curve analysis for biomarkers based on pooled assessments. *Statist. Med.* **22**, 2515-2527.

[14] Farrington, C. (1992). Estimating prevalence by group testing using generalized linear models. *Statist. Med.* **11**, 1591–1597.

[15] Gastwirth, J.L. and Hammick, P.A. (1989). Estimation of prevalence of a rare disease, preserving the anonymity of the subjects by group testing: Application to estimating the prevalence of AIDS antibodies in blood donors. *J. Statist. Plann. Inf.* **22**, 15–27.

[16] Gastwirth J.L. and Johnson, W.O. (1994). Screening with cost-effective quality control: potential applications to HIV and drug testing. *J. Amer. Statist. Assoc.* **89**, 972–981.

[17] Hardwick, J., Page, C. and Stout, Q. (1998). Sequentially deciding between two experiments for estimating a common success probability. *J. Amer. Statist. Assoc.* **93**, 1502–1511.

[18] Huang, X. and Tebbs, J.M. (2009). On latent-variable model misspecification in structural measurement error models for binary response. *Biometrics* **65**, 710–718.

[19] Hung, M.C. and Swallow W.H. (2000). Use of binomial group testing in tests of hypotheses for classification or quantitative covariables. *Biometrics* **56**, 204–212.

[20] Li, M. and Xie, M. (2012). Nonparametric and semiparametric regression analysis of group testing samples. *Int. J. Stats. Med. Res.* **1**, 60–72.

[21] Linton, O. and Whang, Y.L. (2002). Nonparametric estimation with aggregated data. *Econom. Theory* **18**, 420–468.

[22] Liu, A. and Schisterman, E.F. (2003). Comparison of diagnostic accuracy of biomarkers with pooled assessments. *Biometrical Journal* **45**, 631–644.

[23] Meister, A. (2007). Optimal convergence rates for density estimation from grouped data. *Statist. Probab. Lett.* **77**, 1091–1097.

[24] Mitchell, E.M., Lyles, R.H., Manatunga, A.K., Danaher, M., Perkins, N.J. and Schisterman, E.F. (2014). Regression for skewed biomarker outcomes subject to pooling. *Biometrics* **70**, 202–211.

[25] Montesinos-López, O.A., Montesinos-López, A., Crossa, J. and Eskridge, K. (2012). Sample size under inverse negative binomial group testing for accuracy in parameter estimation. *PloS ONE* **7**, e32250.

[26] Montesinos-López, O.A., Montesinos-López, A., Crossa, J. and Eskridge, K. (2013). Sample size for detecting transgenic plants using inverse binomial group testing with dilution effect. *Seed Science Research*, to appear.

[27] Newey, W.K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, Edited by: McFadden, D. and Engler, R. Vol. 5, Amesterdam: North-Holland.

[28] Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33**, 1065–1076.

[29] Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge University Press.

[30] Vansteelandt, S., Goetghebeur, E. and Verstraeten, T. (2000). Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics* **56**, 1126–1133.

[31] Wang, D., Zhou, H. and Kulasekera, K.B. (2013). A semi-local likelihood regression estimator of the proportion based on group testing data. *J. Nonparametr. Statist.* **25**, 209–221.

[32] Wang, D., McMahan, C.S., Gallagher, C.M. and Kulasekera, K.B. (2014). Semiparametric group testing regression models. *Biometrika* **101**, 587–598.

[33] Weinberg, C.R. and Umbach, D.M. (1999). Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics* **55**, 718–726.

[34] Xie, M. (2001). Regression analysis of group testing samples. *Statist. Med.* **20**, 1957–1969.

[35] Zhang, Z. and Albert, P.S. (2011). Binary regression analysis with pooled exposure measurements: A regression calibration approach. *Biometrics* **67**, 636–645.

A Proof of Theorems 5.1 and 5.3

A.1 Proof of Theorem 5.1

To prove Theorem 5.1, we first need to investigate the asymptotic properties of the density estimator $\hat{f}_X(x)$ at (3.4). This is done in Proposition A.1 below. Although the proof of Theorem 5.1 relies on notations and arguments used in the proof of Proposition A.1, the latter proof is quite technical and is relegated to Section E.1 of the supplemental file. The proof of Theorem 5.1 refers to equations defined in Section E.1, which can be recognised through their numbering (E.2), (E.1), etc.

Proposition A.1. Assume that condition (5.1) and (Co2) hold, that $h \to 0$ as $n \to \infty$ and that for some $\delta > 0$, $nh^{2\nu\beta+\delta} \to \infty$ as $n \to \infty$. Then $\hat{f}_X(x) - f_X(x) = T_X(x) + o_P\{n^{-1/2}h^{-(\nu-1)\beta-1/2}\}$, where $T_X(x)$ is a random variable such that $E\{T_X(x)\} = B_X(x)$ and $\mathrm{Var}\{T_X(x)\} = V_X(x)\{1+o(1)\}$, with $B_X(x) = B_{K,X}(x;h)$ and

$$V_X(x) = (\nu^2\, nh)^{-1}\, f_X^{*\nu}(x) V_{K,X}(h). \tag{A.1}$$

In addition, if f_X is twice differentiable with an α-Hölder continuous second derivative for some $0 < \alpha \leq 1$, and K is such that $\mu_{K,1} = 0$ and $\int |u|^{2+\alpha}|K(u)|\, du < \infty$, then $B_X(x) = \frac{1}{2} f_X''(x)\mu_{K,2}\, h^2 + o(h^2)$.

Proof of Theorem 5.1. Throughout, we let const. denote a generic finite positive constant and without loss of generality, we take $c_K = 1$ in condition (Co2). Recalling the notation in (3.4) and (3.6), we write

$$\delta_X(t) = \hat{\phi}_X(t) - \phi_X(t), \quad \delta_m(t) = \hat{\phi}_m(t) - \phi_m(t). \tag{A.2}$$

Moreover, define $\mathcal{B}_n = \{t \in \mathbb{R} : |t| \leq 1/h\}$ and

$$\mathcal{E}_{1n} = \big\{t \in \mathbb{R} : |\Delta_1(t)| \leq |\phi_S(t)|/2\big\} \quad \text{with} \quad \Delta_1(t) = \hat{\phi}_S(t) - \phi_S(t), \tag{A.3}$$

$$\mathcal{E}_{2n} = \big\{u \in \mathbb{R} : |\Delta_2(u)| \leq |\phi_{m^{*\nu}}(u)|/2\big\} \quad \text{with} \quad \Delta_2(t) = \hat{\phi}_{m^{*\nu}}(t) - \phi_{m^{*\nu}}(t). \tag{A.4}$$

By condition (Co1), we have $\inf_{t\in\mathcal{B}_n} \min\{|\phi_X(t)|^\nu, |\phi_m(t)|^\nu\} \geq c\, h^{(\beta\vee\kappa)\nu} \geq 2/n$, for some constant $c > 0$ and all sufficiently large n.

Recall that $\hat{p}(x) = 1 - \hat{q}(x)$ with $\hat{q}(x) = \hat{m}(x)/\hat{f}_X(x)$, and write

$$\hat{q}(x) - q(x) = \hat{f}_X^{-1}(x)\{\hat{m}(x) - q(x)\hat{f}_X(x)\} = Q_1(x) + Q_2(x), \tag{A.5}$$

where $Q_1(x) = f_X^{-1}(x)\{\hat{m}(x) - q(x)\hat{f}_X(x)\}$ and $Q_2(x) = \{\hat{m}(x) - q(x)\hat{f}_X(x)\}\big\{\hat{f}_X^{-1}(x) - f_X^{-1}(x)\big\}$. In what follows, we study $Q_1(x)$ and $Q_2(x)$ separately.

For the numerator of $Q_1(x)$ (see equations (3.4) and (3.6)), taking $\ell = 2$ in (E.5) with X replaced by m we have, for $t \in \mathcal{E}_{2n} \cap \mathcal{B}_n$, $\hat{\phi}_m(t) = \phi_m(t) + \chi_{m,1}(t) + \chi_{m,2}(t) -$

$\chi_{m,3}(t) + \chi_{m,4}(t) + O\{\chi_{m,5}(t)\}$, where $\chi_{m,1}(t) = \nu^{-1}\{\phi_m(t)\}^{1-\nu}\Delta_2(t)$, $\chi_{m,2}(t) = (2\nu^2)^{-1}(1-\nu)\{\phi_m(t)\}^{1-2\nu}\Delta_2^2(t)$, $\chi_{m,3}(t) = \{\chi_{m,1}(t) + \chi_{m,2}(t)\}I_{\mathcal{E}_{2n}^c}(t)$, $\chi_{m,4}(t) = \delta_m(t)I_{\mathcal{E}_{2n}^c}(t)$ and $\chi_{m,5}(t) = |\phi_m(t)|^{1-3\nu}|\Delta_2(t)|^3$.

Similarly, following the proof that leads to (E.20), it can be shown that

$$\begin{aligned}
&\hat{m}(x) - (2\pi)^{-1}\int e^{-\mathrm{i}tx}\phi_K(th)\phi_m(t)\,dt \\
&\quad = (2\pi)^{-1}\int e^{-\mathrm{i}tx}\phi_K(th)\{\chi_{m,1}(t) + \chi_{m,2}(t)\}\,dt + O_P\big\{n^{-3/2}h^{-(3\nu-1)\kappa-1}\big\}, \\
&\hat{f}_X(x) - (2\pi)^{-1}\int e^{-\mathrm{i}tx}\phi_K(th)\phi_X(t)\,dt \\
&\quad = (2\pi)^{-1}\int e^{-\mathrm{i}tx}\phi_K(th)\chi_0(t)\,dt \\
&\qquad + O_P\big\{n^{-1}h^{-(2\nu-1)\beta-1/2}\big\} + O_P\big\{n^{-3/2}h^{-(3\nu-1)\beta-1}\big\},
\end{aligned}$$

where $\chi_0(t) = \nu^{-1}\{\phi_X(t)\}^{1-\nu}\Delta_1(t)$.

By Plancherel's isometry, we deduce from the above calculations that

$$\begin{aligned}
f_X(x)Q_1(x) &= -f_X(x)B_p(x) + (2\pi)^{-1}\int e^{-\mathrm{i}tx}\,\phi_K(th)\big\{\chi_{m,1}(t) - q(x)\,\chi_0(t)\big\}\,dt \\
&+ (2\pi)^{-1}\int e^{-\mathrm{i}tx}\,\phi_K(th)\chi_{m,2}(t)\,dt + O_P\big\{n^{-1}h^{-(2\nu-1)\beta-1/2} + n^{-3/2}h^{-(3\nu-1)(\beta\vee\kappa)-1}\big\},
\end{aligned}$$

where $B_p(x)$ is defined given in the statement of Theorem 5.1. In the above expression, using the argument leading to (E.11) we get $\int e^{-\mathrm{i}tx}\,\phi_K(th)\chi_{m,2}(t)\,dt = O_P\{n^{-1}h^{-(2\nu-1)\kappa-1/2}\}$. Moreover, put

$$Q_{1,1}(x) = \{2\pi f_X(x)\}^{-1}\int e^{-\mathrm{i}tx}\,\phi_K(th)\big\{\chi_{m,1}(t) - q(x)\,\chi_0(t)\big\}\,dt. \tag{A.6}$$

As long as $f_X(x)$ is bounded away from zero and $nh^{2\nu(\beta\vee\kappa)+1/2} \to \infty$, we have

$$Q_1(x) = -B_p(x) + Q_{1,1}(x) + O_P\big\{n^{-1/2}h^{-(\nu-1)(\beta\vee\kappa)-1/2}\big\}. \tag{A.7}$$

Next, note that $EQ_{1,1}(x) = 0$ for $Q_{1,1}(x)$ as in (A.6). Then we derive the asymptotic expression of the variance of $Q_{1,1}(x)$ using an argument similar to that used to

deal with $R_1(x)$ in (E.6). First, rewrite it as $\{nf_X(x)\}^{-1}\sum_j(\xi_{nj}-E\xi_{nj})$, where

$$\xi_{nj} = Z_j^* K_{m,h}(x-S_j) - q(x)K_{X,h}(x-S_j) \tag{A.8}$$

with $K_{X,h}(y)=h^{-1}K_X(y/h)$ and $K_{m,h}(y)=h^{-1}K_m(y/h)$ for K_X as in (E.7) and

$$K_m(y) = (2\pi\nu)^{-1}\int e^{-\mathrm{i}ty}\,\phi_K(t)\{\phi_m(t/h)\}^{1-\nu}\,dt. \tag{A.9}$$

Then, Lemma E.4 in Section E establishes the asymptotic expression of the variance of ξ_{nj}, such that, $\mathrm{Var}\big(n^{-1}\sum_j\xi_{nj}\big) = f_X^2(x)\big[V_{p,1}(x)\{1+o(1)\}+V_{p,2}(x)\big]$ for $V_{p,1}(x)$ and $V_{p,2}(x)$ as in (5.6) and (5.7), respectively. This, together with (A.7), gives us

$$Q_1(x) = -T_p(x) + o_P\big\{n^{-1/2}h^{-(\nu-1)(\beta\vee\kappa)-1/2}\big\}, \tag{A.10}$$

where $T_p(x)$ is given in the statement of Theorem 5.1.

Finally, we address the order of $Q_2(x)=\{\hat m(x)-q(x)\hat f_X(x)\}\{\hat f_X^{-1}(x)-f_X^{-1}(x)\}$. By conditions (Co1) and (Co2), and if $nh^{(2\nu-1)\beta+1/2}\to\infty$ as $n\to\infty$, Proposition A.1 with $\delta=\frac{1}{2}$ yields that $\hat f_X^{-1}(x)-f_X^{-1}(x)=o_P(1)$. Further, it follows from the definition of Q_1 and the fact $f_X(x)>0$ that $\hat m(x)-q(x)\hat f_X(x)=O_P(Q_1)$. Together with (A.5) and (A.10), this proves (5.5) and thus completes the proof of the theorem. □

A.2 Proof of Theorem 5.3

For brevity, we prove the result only for $\nu=2$ and dimension $d=2$, but our arguments can be directly extended to deal with $\nu>2$ and $d>2$. Let $q'_{\boldsymbol\theta}=\nabla_{\boldsymbol\theta}\,q_{\boldsymbol\theta}$, so that $\nabla_{\boldsymbol\theta}(q_{\boldsymbol\theta}f_X)^{*2}=2\,m'_{\boldsymbol\theta}*m_{\boldsymbol\theta}$, and let $\ell_0(\boldsymbol\theta)=\frac{1}{2}\,E[Y^*\ln\{f_S(S)-m^{*2}_{\boldsymbol\theta}(S)\}+(1-Y^*)\ln m^{*2}_{\boldsymbol\theta}(S)]$. Under our conditions, $\boldsymbol\theta_0$ uniquely maximises $\ell_0(\boldsymbol\theta)$ subject to $\boldsymbol\theta\in\Theta$. Finally recall the definition of $\hat\ell_n$ at (4.2).

Step 1: Consistency. Using an argument similar to that used in the proofs of Propositions A.1 and F.1, it can be proved that, under condition (5.1), $E\|\hat f_X-$

$f_X\|_2^2 \le \text{const.}\{n^{-1/2}h^{-1} + n^{-1}h^{-2\beta-1} + h^{2\beta-1}\}$, and under condition (5.3) with $h = (d\,\gamma \log n)^{-1/\rho}$ for some $0 < \delta \le \frac{1}{4}$, $E\|\hat{f}_X - f_X\|_2^2 \le \text{const.}\,(\log n)^{\text{const.}} n^{-2\delta}$. Consequently, for h as in (P5), we have $\|\hat{f}_X - f_X\|_2 = o_P(1)$.

Moreover, note that $(q_{\boldsymbol{\theta}}\hat{f}_X)^{*2}(s) = \int q_{\boldsymbol{\theta}}(x)\hat{f}_X(x)q_{\boldsymbol{\theta}}(s-x)\hat{f}_X(s-x)\,dx$. Using Hölder's inequality, we get, for any $s \in \mathbb{R}$ and $\boldsymbol{\theta} \in \Theta$,

$$\begin{aligned}\big|(q_{\boldsymbol{\theta}}\hat{f}_X)^{*2}(s) - (q_{\boldsymbol{\theta}}f_X)^{*2}(s)\big| &= \big|\{q_{\boldsymbol{\theta}}(\hat{f}_X - f_X)\}^{*2}(s) + 2\{q_{\boldsymbol{\theta}}(\hat{f}_X - f_X)\} * (q_{\boldsymbol{\theta}}f_X)(s)\big| \\ &\le \text{const.}\big(\|\hat{f}_X - f_X\|_2 \cdot \|f_X\|_2 + \|\hat{f}_X - f_X\|_2^2\big),\end{aligned}$$

where in the last step we used the fact that $q_{\boldsymbol{\theta}} \le 1$. Therefore,

$$(q_{\boldsymbol{\theta}}\hat{f}_X)^{*2}(s) = (q_{\boldsymbol{\theta}}f_X)^{*2}(s) + o_P(1) \tag{A.11}$$

holds uniformly in $s \in \mathbb{R}$ and $\boldsymbol{\theta} \in \Theta$.

For the kernel estimator $\hat{f}_S$, it is well-known that $\|\hat{f}_S - f_S\|_\infty = \sup_{s\in\mathbb{R}} |\hat{f}_S(s) - f_S(s)| = o_P(1)$, under conditions (P2), (P4) and (P5) (see, e.g. Parzen, 1962). This together with (A.11) imply that, uniformly for $1 \le j \le n$ and $\boldsymbol{\theta} \in \Theta$, $\hat{f}_S(S_j) = f_S(S_j) + o_P(1)$ and $(q_{\boldsymbol{\theta}}\hat{f}_X)^{*2}(S_j) = (q_{\boldsymbol{\theta}}f_X)^{*2}(S_j) + o_P(1)$. Subsequently, by the law of large numbers and the compactness of Θ, $\hat{\ell}_n(\boldsymbol{\theta}) = (2n)^{-1}\sum_{j=1}^n \big[Y_j^* \ln\big\{f_S(S_j) - m_{\boldsymbol{\theta}}^{*2}(S_j)\big\} + (1-Y_j^*)\ln m_{\boldsymbol{\theta}}^{*2}(S_j)\big] + o_P(1) = \ell_0(\boldsymbol{\theta}) + o_P(1)$ holds uniformly in $\boldsymbol{\theta} \in \Theta$.

On the other hand, by condition (P2), the constraints in (4.3) are satisfied with probability tending to one, as long as the constant c_0 is chosen small enough, say $c_0 < \min\{c_1, c_1(c_2^{-1}-1)\}$ for c_1 and c_2 as in (P2). Therefore, $\sup_{\boldsymbol{\theta}\in\Theta}|\hat{\ell}_n(\boldsymbol{\theta}) - \ell_0(\boldsymbol{\theta})| \xrightarrow{P} 0$, and applying the fundamental consistency result for extremum estimators (e.g. Theorem 2.1 in Newey and McFadden, 1994), we obtain $\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$.

Step 2: Asymptotic normality. To establish the asymptotic normality for $\hat{\boldsymbol{\theta}}$, we proceed by verifying the assumptions of Theorem 3.1 of Newey and McFadden (1994).

Put $\hat{\mathcal{C}}_n = \big\{\boldsymbol{\theta} \in \Theta : \boldsymbol{\theta} \text{ satisfies (4.3)}\big\} \subseteq \Theta$. Recall that $\sup_{\boldsymbol{\theta}\in\Theta}\|\hat{m}_{\boldsymbol{\theta}} - m_{\boldsymbol{\theta}}\|_\infty = o_P(1)$ and $\|\hat{f}_S - f_S\|_\infty = o_P(1)$, where $\hat{m}_{\boldsymbol{\theta}} = q_{\boldsymbol{\theta}}\hat{f}_X$. Together with condition (P2), this

implies that there exists a neighbourhood $\mathcal{N} \subset \Theta$ of $\boldsymbol{\theta}_0$ such that $P(\mathcal{N} \subset \hat{\mathcal{C}}_n) \to 1$ and $P(\hat{\boldsymbol{\theta}} \in \mathcal{N}) \to 1$ provided $c_0 < c_1 \min(1, c_2^{-1} - 1)$.

By condition (P1), $\hat{\ell}_n(\boldsymbol{\theta})$ is twice differentiable. Using the same argument as the one that leads to (A.11), it can be proved that

$$\nabla_{\boldsymbol{\theta}} \hat{\ell}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{j=1}^{n} \frac{(1 - 2Y_j^*)\{(q'_{\boldsymbol{\theta}} \hat{f}_X) * (q_{\boldsymbol{\theta}} \hat{f}_X)\}(S_j)}{Y_j^* \, \hat{f}_S(S_j) + (1 - 2Y_j^*)(q_{\boldsymbol{\theta}} \hat{f}_X)^{*2}(S_j)} = \frac{1}{n} \sum_{j=1}^{n} \Lambda(S_j, \boldsymbol{\theta}) + o_P(1)$$

and $\nabla^2_{\boldsymbol{\theta}} \hat{\ell}_n(\boldsymbol{\theta}) = n^{-1} \sum_{j=1}^{n} \mathbf{G}(S_j, \boldsymbol{\theta}) + o_P(1) = \mathbf{H}(\boldsymbol{\theta}) + o_P(1)$ hold uniformly in $\boldsymbol{\theta} \in \Theta$, where $\Lambda(\cdot, \boldsymbol{\theta})$ is given in (5.10), $\mathbf{G}(s, \boldsymbol{\theta}) = \big\{\nabla_{\boldsymbol{\theta}} \Lambda(s, \boldsymbol{\theta})\big\}^{\mathrm{T}}$, $\mathbf{H}(\boldsymbol{\theta}) = E\{\mathbf{G}(S, \boldsymbol{\theta})\}$ and $\nabla^2_{\boldsymbol{\theta}}(f)$ denotes the Hessian of f. Note that $E\{\Lambda(S, \boldsymbol{\theta}_0)\} = 0$, so that, by the central limit theorem, $\sqrt{n}\, \nabla_{\boldsymbol{\theta}} \hat{\ell}_n(\boldsymbol{\theta}_0) \xrightarrow{D} N\big(0, \boldsymbol{\Sigma}_0\big)$, where $\boldsymbol{\Sigma}_0 = E\{\Lambda(S, \boldsymbol{\theta}_0)\Lambda(S, \boldsymbol{\theta}_0)^{\mathrm{T}}\}$.

Finally, under conditions (P1) and (P2), for the above neighbourhood $\mathcal{N}$ of $\boldsymbol{\theta}_0$, $E \sup_{\boldsymbol{\theta} \in \mathcal{N}} \|\nabla_{\boldsymbol{\theta}} \Lambda(S, \boldsymbol{\theta})\| < \infty$, with $\|\cdot\|$ the Euclidean norm. By Lemma 3.6 of Newey and McFadden (1994), we have $\boldsymbol{\Sigma}_0 = \mathbf{H}_0 = \mathbf{H}(\boldsymbol{\theta}_0)$. The proof follows by Theorem 3.1 of Newey and McFadden (1994), using the fact that $\mathbf{H}_0^{-1} \boldsymbol{\Sigma}_0 \mathbf{H}_0 = \boldsymbol{\Sigma}_0^{-1}$. □