Biometrika (2015), **xx**, x, *pp*. 1–16

Printed in Great Britain

Nonparametric methods for group testing data, taking dilution into account

BY A. DELAIGLE, P. HALL

School of Mathematics and Statistics, University of Melbourne, Parkville, Victoria 3010, Australia.

A.Delaigle@ms.unimelb.edu.au halpstat@ms.unimelb.edu.au

SUMMARY

Group testing methods are used widely to assess the presence of a contaminant, based on measurements of the concentration of a biomarker, for example to test the presence of a disease in pooled blood samples. The test would be perfect if it produced a positive result whenever the contaminant was present, and a negative result otherwise. However, in practice the test is always at least somewhat imperfect, for example because it is sensitive to the proportion of contaminated items in the group, rather than to the sheer existence of one or more contaminated items. We develop a nonparametric method for accommodating this dilution effect. Our approach allows us to estimate, under minimal assumptions, the probability $m(x)$ that an item is contaminated, conditional on the value x of an explanatory variable, and to estimate the probability, q, that an individual chosen at random is disease free, and the specificity Sp, and the sensitivity Se, of the test. These are all ill-posed problems, where poor convergence rates are usually encountered. However, despite these pessimistic expectations, our estimators of q, Sp and Se are root-N consistent, where N denotes the total number of individuals in all the groups, and our estimator of $m(x)$ converges at the rate it would enjoy if q, Sp and Se were known.

Some key words: Bandwidth choice; biomarker; blood testing; count data; kernel methods; local polynomial methods; nonparametric regression; sensitivity; specificity.

1. INTRODUCTION

We consider studies where the interest is to model the relation $m(x) = \mathrm{pr}(Y = 1 \mid X = x)$ between a binary random variable $Y = 0$ or 1 and an explanatory variable X. For example, Y may represent the health status of a patient, the presence or absence of a toxic or polluting substance in a water or milk sample, or the transgenic status of a plant; and X may represent age, cholesterol level, or the location of a river or a field.

For reasons such as time and cost restrictions and technical complexity, it is not always possible to observe directly the status of each individual or item in the study. Instead, the individuals are pooled randomly into J groups of respective sizes $\nu_1, \ldots, \nu_J$, and we can observe only independent group testing data (X_{ij}, Y_j^*) $(j = 1, \ldots J;\ i = 1, \ldots, \nu_j)$, where Y_j^* denotes the observed status of the jth group and X_{ij} is an explanatory variable for the ith individual in the jth group. See for example Dorfman (1943), Gastwirth and Hammick (1989), Chen and Swallow (1990) and Farrington (1992). For each group, the status Y_j^* is usually obtained through a test which is often imperfect, and which produces errors when Y_j^* is not equal to the true status of the jth group.

Characteristics of the problem that are of practical interest include the curve m, and q, Sp and Se introduced in the abstract. Methods for estimating m have been suggested by Vansteelandt et al. (2000), Xie (2001), Chen et al. (2009), Delaigle and Meister (2011), Delaigle and Hall (2012) and Li and Xie (2012). For simplicity, those authors assumed that the errors of the tests from which one obtains the Y_j^*s did not depend on the ν_js nor the X_{ij}s, but as pointed by Wein and Zenios (1996) and Zenios and Wein (1998), they often depend on the group sizes, and ignoring this leads to biased estimators. Taking these issues into account, McMahan et al. (2013) suggested parametric estimators. We construct nonparametric estimators, focusing on the case where the X_{ij}s are univariate. In the multivariate case, our techniques could be extended, for example, by adapting to our setting semiparametric approaches such as the single index and partially linear models developed by Delaigle et al. (2014).

2. ESTIMATING m WHEN SPECIFICITY AND SENSITIVITY OF THE TEST ARE KNOWN

2·1. *Model and data*

Our model and main data are as in McMahan et al. (2013). We observe a sample of independent vectors $(X_{1j}, \ldots, X_{\nu_j j}, Y_j^*)$ $(j = 1, \ldots, J)$, where the X_{ij}s are independent and Y_j^* is a binary $0/1$ variable representing the result of a test carried on the jth group. The test is imperfect, and Y_j^* may differ from the true, unobserved, disease status, $\widetilde{Y}_j^*$, of the jth group, where $\widetilde{Y}_j^* = \max_{i=1,\ldots,\nu_j} \widetilde{Y}_{ij}$, with $\widetilde{Y}_{ij}$ denoting the unobserved true status, 0 or 1, of the ith individual from the jth group, and with the $(X_{ij}, \widetilde{Y}_{ij})$s independent and identically distributed. Our goal is to estimate

$$m(x) = \mathrm{pr}\big(\widetilde{Y}_{ij} = 1 \,\big|\, X_{ij} = x\big) . \tag{1}$$

As in Wein and Zenios (1996), Zenios and Wein (1998) and McMahan et al. (2013), Y_j^* is obtained through measurements of a continuous quantity B, for example, in the case of disease testing, a biomarker concentration, or, in more general settings, the continuous level or concentration of a chemical substance. Let B_{ij} denote the unobserved concentration or level of a chemical substance for the ith individual from the jth group. In the biomarker example, the concentration in the jth pool is usually taken to be the average concentration $\bar{B}_j = \nu_j^{-1} \sum_{i=1}^{\nu_j} B_{ij}$. We shall follow this convention here, although other models could be adopted, allowing for example the average biomarker concentration to take into account potential imperfect mixing or unequal volumes in the samples. In practice, $\bar{B}_j$ is measured through a complex process that typically incurs measurement errors. For example, instead of the biomarker concentration $\bar{B}_j$ we may observe the optical density reading W_j; see Wein and Zenios (1996) and Zenios and Wein (1998). Following McMahan et al.'s (2013) Section 4, we assume that the biomarker is measured with error, that is, we observe independent variables $W_1, \ldots, W_J$ satisfying

$$W_j = \bar{B}_j + U_j \quad (j = 1, \ldots, J), \tag{2}$$

where $\bar{B}_j$ and U_j are independent and $U_j \sim f_U$, and that the result of the jth test is given by

$$Y_j^* = 1\big(W_j > t_0^{(j)}\big) , \tag{3}$$

with $t_0^{(j)}$ a cutoff point. As in McMahan et al. (2013) we shall treat $t_0^{(j)}$ as fixed and known. In practice it can be chosen so as to minimise the variance of an estimator of m; see Section 5·3. Other error models are possible, for example in the case where there are errors induced by the limit of detection of the measurement device.

We adopt McMahan et al.'s (2013) assumption that the error density f_U is known, although the unknown error case could also be treated, using for example methods analogous to those employed in Delaigle et al. (2008) or Delaigle and Hall (2015). Throughout we assume that

$$\text{for all } t\,, \quad |\phi_U(t)| \neq 0, \text{ where } \phi_U(t) = E(e^{itU}), \tag{4}$$

which is satisfied by many common distributions. As implicitly assumed by McMahan et al. (2013), we assume throughout that U_j is independent of $\widetilde{Y}_j^*$, and that, for all x,

$$\mathrm{pr}\biggl(W_j \le x \;\bigg|\; \sum_{i=1}^{\nu_j} \widetilde{Y}_{ij}, X_{1j}, \ldots, X_{\nu_j j}\biggr) = \mathrm{pr}\biggl(W_j \le x \;\bigg|\; \sum_{i=1}^{\nu_j} \widetilde{Y}_{ij}\biggr)\,.$$

We let $N = \sum_{j=1}^{J} \nu_j$ denote the total number of individuals in the sample, we let

$$\mathrm{Sp}^{(j)} = \mathrm{pr}\bigl(W_j \le t_0^{(j)} \,\big|\, \widetilde{Y}_j^* = 0\bigr) \tag{5}$$

denote the specificity of a the test performed on a group of size ν_j, and we denote the sensitivity of the test performed on a group of size ν_j, in which exactly k individuals are positive, by

$$\mathrm{Se}^{(j,k)} = \mathrm{pr}\biggl(W_j > t_0^{(j)} \;\bigg|\; \sum_{i=1}^{\nu_j} \widetilde{Y}_{ij} = k\biggr) \quad (k = 1, \ldots, \nu_j\,). \tag{6}$$

Remark 1. The estimator of m we introduce in Section 2·2 can be employed in the setting where the biomarker is measured without error, i.e., when $U_j = 0$. As we shall see, it is easy to adapt to this case the estimators of specificity and sensitivity developed in Sections 3·2 and 3·3.

2·2. *Oracle local polynomial estimator of m for imperfectly observed grouped data*

If the $(X_{ij}, \widetilde{Y}_{ij})$s were observed, we could estimate $m(x)$ at (1) nonparametrically by a standard local polynomial estimator of order p. Let $\mathcal{K}$ be a function called kernel, $h_0 > 0$ a parameter called bandwidth, and $\mathcal{K}_{h_0} = h_0^{-1}\mathcal{K}(\cdot/h_0)$. In the Supplementary Material, we recall the construction of this estimator, which can be written as $\widehat{m}(x) = e_1^{\mathrm{T}}\,\mathrm{S}_N^{-1}\,\mathrm{T}_N$, where $e_1^{\mathrm{T}} = (1, 0, \ldots, 0)$ is a $(p+1)$-vector, S_N is a $(p+1) \times (p+1)$ matrix with $(k+1, \ell+1)$th element equal to $S_{N,k\ell}(x) = N^{-1}\, h_0^{-k-\ell} \sum_{j=1}^{J} \sum_{i=1}^{\nu_j} \mathcal{K}_{h_0}(X_{ij} - x)\,(X_{ij} - x)^{k+\ell}$, for $k, \ell = 0, \ldots, p$, and $\mathrm{T}_N = (T_{N,0}, \ldots, T_{N,p})^{\mathrm{T}}$ is a $(p+1)$-vector, with $T_{N,k}(x) = N^{-1}\, h_0^{-k} \sum_{j=1}^{J} \sum_{i=1}^{\nu_j} \widetilde{Y}_{ij}\, \mathcal{K}_{h_0}(X_{ij} - x)\,(X_{ij} - x)^k$, for $k = 0, \ldots, p$; see Fan and Gijbels (1996). For example, the local polynomial estimator of order $p = 0$ is equal to $\widehat{m}(x) = N^{-1} \sum_{j,i} \widetilde{Y}_{ij}\, \mathcal{K}_{h_0}(X_{ij} - x) / N^{-1} \sum_{j,i} \mathcal{K}_{h_0}(X_{ij} - x)$ and its denominator and numerator are consistent estimators of, respectively, $f_X(x)$ and $m(x)\, f_X(x)$.

In our case we can only observe vectors $(X_{1j}, \ldots, X_{\nu_j j}, W_j)$ $(j = 1, \ldots, J)$, where W_j follows the model at (2), and Y_j^* is obtained from W_j through (3). Since the $\widetilde{Y}_{ij}$s are not observed, we cannot compute the standard estimator $\widehat{m}$ derived above. Delaigle and Meister (2011) suggested a modified local polynomial estimator which can be computed from group testing data, but it cannot be applied here, since they assumed that the error of the test in each group depends only on the true status of the group. In this section we derive an oracle nonparametric estimator of m, where it is assumed that $q = 1 - \mathrm{pr}\bigl(\widetilde{Y}_{ij} = 1\bigr)$ and the specificities and sensitivities at (5) and (6) are known. The case where these quantities are unknown will be dealt with in Section 3·5.

To see how to construct a pth order local polynomial estimator in our context, recall the standard local polynomial estimator of order $p = 0$ derived above. By analogy, consider the estimator $m^{\dagger}(x) = N^{-1} \sum_{j,i} Y_j^*\, \mathcal{K}_{h_0}(X_{ij} - x) / N^{-1} \sum_{j,i} \mathcal{K}_{h_0}(X_{ij} - x)$ where the unobserved

$\widetilde{Y}_{ij}$s are replaced by the Y_j^*s. As in the standard case, the denominator of $m^\dagger$ converges to $f_X(x)$. However, as $N \to \infty$, the numerator converges to the limit of $f_X(x)\,N^{-1}\sum_{j,i} E\big(Y_j^* \,\big|\, X_{ij} = x\big)$, which is not equal to $f_X(x)m(x)$. Indeed, in the Supplementary Material, we prove that

$$E\big(Y_j^* \,\big|\, X_{ij}=x\big) = \mathrm{pr}\big(Y_j^*=1 \,\big|\, X_{ij}=x\big) = \mathcal{A}_j + m(x)\,\mathcal{B}_j\,, \tag{7}$$

with $\mathcal{A}_j = q^{\nu_j-1}(1-\mathrm{Sp}^{(j)}) + \mathrm{Se}_2^{(j)}$, $\mathcal{B}_j = \mathrm{Se}_1^{(j)} - \mathrm{Se}_2^{(j)} - q^{\nu_j-1}(1-\mathrm{Sp}^{(j)})$, and, recalling (6),

$$\mathrm{Se}_\ell^{(j)} = \sum_{k=1}^{\nu_j+1-\ell} \mathrm{Se}^{(j,k)} \binom{\nu_j-1}{k+\ell-2} (1-q)^{k+\ell-2}\, q^{\nu_j-k-\ell+1}\,, \ell=1,2\,; \tag{8}$$

here, for $\ell=2$, the sum over k is interpreted as zero when $\nu_j=1$, so that in that case $\mathrm{Se}_2^{(j)}=0$.

Therefore $m^\dagger$ does not converge to m. On the other hand, letting $Z_j^* = Y_j^*/\mathcal{B}_j$ and $\mathcal{D} = N^{-1}\sum_{j,i}\mathcal{A}_j/\mathcal{B}_j$, we deduce from those calculations that $\widetilde{m}(x) = N^{-1}\sum_{j,i} Z_j^*\,\mathcal{K}_{h_0}(X_{ij}-x)/N^{-1}\sum_{j,i}\mathcal{K}_{h_0}(X_{ij}-x) - \mathcal{D}$ converges to $m(x)$. More generally, a pth order local polynomial estimator of m can be defined by

$$\widetilde{m} = e_1^{\mathrm{T}}\,\mathrm{S}_N^{-1}\,\mathrm{V}_N - \mathcal{D}, \tag{9}$$

where S_N is the matrix defined above and $\mathrm{V}_N = (V_{N,0},\ldots,V_{N,p})^{\mathrm{T}}$, with

$$V_{N,k}(x) = N^{-1}\,h_0^{-k}\sum_{j=1}^{J}\sum_{i=1}^{\nu_j} Z_j^*\,\mathcal{K}_{h_0}(X_{ij}-x)\,(X_{ij}-x)^k\,. \tag{10}$$

3. ESTIMATORS WHEN q, SPECIFICITY AND SENSITIVITY ARE ESTIMATED

3·1. *Model and data for the biomarker*

In Section 2·2, we derived an estimator of m based on the assumption that q, $\mathrm{Sp}^{(j)}$ and $\mathrm{Se}^{(j,k)}$ were known. In Sections 3·2 to 3·4 we shall construct estimators of these quantities from additional data, and derive from there a modified estimator of m in Section 3·5. Recall from (5) and (6) that $\mathrm{Sp}^{(j)}$ and $\mathrm{Se}^{(j,k)}$ depend on the unknown distribution of B. Following Wein and Zenios (1996), Zenios and Wein (1998) and McMahan et al. (2013), the distribution of B_{ij} depends on the status of the individual. Let f_{B^+} and f_{B^-} denote the density of B_{ij} given that $\widetilde{Y}_{ij}=1$ and B_{ij} given that $\widetilde{Y}_{ij}=0$, respectively, and let $f_{\bar{B}_{j;k}}$ denote the density of $\bar{B}_j$ in the jth group, given that it contains exactly $k \le \nu_j$ positive individuals. As in Wein and Zenios (1996), we have

$$f_{\bar{B}_{j;k}}(x) = \nu_j\, f_{B^+}^{*k} * f_{B^-}^{*(\nu_j-k)}(\nu_j\, x)\,,$$

where $*$ denotes convolution product, f^{*k} is the k-fold convolution of f, and we use the convention that $f_{B^+}^{*0} * f_{B^-}^{*\nu_j} = f_{B^-}^{*\nu_j}$ and $f_{B^+}^{*\nu_j} * f_{B^-}^{*0} = f_{B^+}^{*\nu_j}$.

As in McMahan et al. (2013), in addition to the main sample in Section 2·1, we observe training samples of contaminated data $W_1^-,\ldots,W_{n_-}^-$ and $W_1^+,\ldots,W_{n_+}^+$ obtained from, respectively, negative and positive individuals, where

$$W_i^- = B_i^- + U_i^- \quad , \quad W_i^+ = B_i^+ + U_i^+\,, \tag{11}$$

with the B_i^-s, the U_i^-s, the B_i^+s and the U_i^+s totally independent, $B_i^- \sim f_{B^-}$, $B_i^+ \sim f_{B^+}$, $U_i^- \sim f_U$ and $U_i^+ \sim f_U$. See Wein and Zenios (1996) for an example with training data from the National HIV Reference Laboratory in Australia.

The analysis in our paper is based on the assumption that training sample size is of the same order of magnitude as the total number of individuals in the sample. Specifically, $n_- \asymp N$ and $n_+ \asymp N$, where here and below, for any real a and b we use the notation $a \asymp b$ when the ratio a/b is bounded away from zero and infinity as N diverges. The larger N, the better. There does not exist a specific value of training sample size such that performance decreases precipitously for lower values of training sample size, or performance increases dramatically for larger values.

3·2. *Nonparametric estimator of* $\mathrm{Sp}^{(j)}$

Next we show how to estimate $\mathrm{Sp}^{(j)}$ nonparametrically from data $W_1^-, \ldots, W_{n_-}^-$ from the model at (11). Since $\widetilde{Y}_j^* = 0$ implies that there is no true positive in the jth group, then conditional on $\widetilde{Y}_j^* = 0$, the density of $W_j = \bar{B}_j + U_j$ is $f_U * f_{\bar{B}_{j;0}}$. As in Wein and Zenios (1996), Zenios and Wein (1998) and McMahan et al. (2013) it follows from (5) that in the jth group,

$$\mathrm{Sp}^{(j)} = \int_{-\infty}^{t_0^{(j)}} f_U * f_{\bar{B}_{j;0}}(x)\, dx = \nu_j \int_{-\infty}^{t_0^{(j)}} \int_{-\infty}^{\infty} f_U(x-y)\, f_{B^-}^{*\nu_j}(\nu_j\, y)\, dy\, dx\,. \qquad (12)$$

When $\nu_j = 1$, $\mathrm{Sp}^{(j)} = \mathrm{pr}\big(W^- \le t_0^{(j)}\big)$, which can be estimated at the parametric rate $n_-^{-1/2}$ by $\widehat{\mathrm{Sp}}^{(j)} = n_-^{-1} \sum_{i=1}^{n_-} 1\big(W_i^- \le t_0^{(j)}\big)$. Next we consider the more difficult case where $\nu_j > 1$.

McMahan et al. (2013) proposed a parametric procedure for estimating $\mathrm{Sp}^{(j)}$, and in their numerical section they considered briefly a nonparametric estimator. There, they estimated f_{B^-} by the deconvolution estimator $\widehat{f}_{B^-}$ of Carroll and Hall (1988) and Stefanski and Carroll (1990) with a second-order kernel and the bootstrap bandwidth of Delaigle and Gijbels (2004), and approximated $\mathrm{Sp}^{(j)}$ from data they generated from $\widehat{f}_{B^-}$; see the Supplementary Material. While their suggestion is of interest, in the present setting the procedure is rather complex and suffers from slow convergence rates. For example, if f_U is normally distributed, the rates are logarithmic.

These difficulties can be avoided by constructing a simple kernel estimator that converges to $\mathrm{Sp}^{(j)}$ at a much faster rate, indeed sometimes reaching the parametric rate $n_-^{-1/2}$. To motivate this estimator, we first recall how to construct a kernel estimator of the density f_Z of a random variable Z using independent and identically distributed data $Z_1, \ldots, Z_n \sim f_Z$. Let ϕ_Z denote the characteristic function of Z. By the Fourier inversion theorem, if $|\phi_Z|$ is integrable we can write $f_Z(x) = (2\pi)^{-1} \int e^{-itx}\, \phi_Z(t)\, dt$. Let K be a kernel, $\phi_K(t) = \int e^{itx}\, K(x)\, dx$, $h > 0$ be a bandwidth and $\widehat{\phi}_Z$ denote the empirical characteristic function of Z. The kernel estimator of $f_Z(x)$ is defined by

$$\widehat{f}_Z(x) = \frac{1}{nh} \sum_{j=1}^{n} K\Big(\frac{x - Z_j}{h}\Big) = \frac{1}{2\pi} \int e^{-itx}\, \widehat{\phi}_Z(t)\, \phi_K(ht)\, dt\,.$$

Using Fourier inversion, we rewrite $\mathrm{Sp}^{(j)}$, at (12), as

$$\mathrm{Sp}^{(j)} = \frac{1}{2\pi} \int_{-\infty}^{t_0^{(j)}} \int_{-\infty}^{\infty} e^{-itx}\, \phi_U(t)\, \phi_{B^-}^{\nu_j}(t/\nu_j)\, dt\, dx\,. \qquad (13)$$

Then, parallelling the construction of $\widehat{f}_Z$ above, noting that $\phi_{B^-}^{\nu_j}(t/\nu_j) = \phi_{W^-}^{\nu_j}(t/\nu_j)/\phi_U^{\nu_j}(t/\nu_j)$ where ϕ_U is known since f_U is known, and recalling (4), we suggest estimating $\mathrm{Sp}^{(j)}$ by

$$\widehat{\mathrm{Sp}}^{(j)} = \frac{1}{2\pi} \int_{-\infty}^{t_0^{(j)}} \int_{-\infty}^{\infty} e^{-ity}\, \phi_U(t)\, \widehat{\phi}_{W^-}^{\nu_j}(t/\nu_j)\, \big\{\phi_U^{\nu_j}(t/\nu_j)\big\}^{-1}\, \phi_K(ht)\, dt\, dy\,. \qquad (14)$$

Here, $\widehat{\phi}^{\nu_j}_{W^-}$ denotes the unbiased estimator of $\phi^{\nu_j}_{W^-}$ defined by

$$\widehat{\phi}^{\nu_j}_{W^-}(t) = \binom{n_-}{\nu_j}^{-1} \sum_{\ell_1<\ldots<\ell_{\nu_j}} \exp\left\{it\left(W^-_{\ell_1}+\cdots+W^-_{\ell_{\nu_j}}\right)\right\},$$

and the kernel K is chosen so that $\phi_K(0)=1$ and the integral at (14) is well defined.

Remark 2. An estimator that is simpler to compute in practice, and which produces estimators $\widehat{\mathrm{Sp}}^{(j)}$ with the same convergence rates, is obtained by replacing $\widehat{\phi}^{\nu_j}_{W^-}$ in (14) by the empirical characteristic function of W^- raised to the power ν_j. We used this estimator in Section 5.

Remark 3. In the simpler error-free case where $U \equiv 0$, an estimator with the same parametric rates as the estimator at (14) can be computed without any smoothing. In that case, $W^- = B^-$ and $\phi_U \equiv 1$, and we can take $\widehat{\mathrm{Sp}}^{(j)} = \sum_{\ell_1<\ldots<\ell_{\nu_j}} 1\{(B^-_{\ell_1}+\cdots+B^-_{\ell_{\nu_j}})/\nu_j \le t_0^{(j)}\}\,/\binom{n_-}{\nu_j}$.

3·3. *Nonparametric estimator of* $\mathrm{Se}^{(j,k)}$

Next we construct a nonparametric estimator of $\mathrm{Se}^{(j,k)}$ from data $W_1^-,\ldots,W_{n_-}^-$ and $W_1^+,\ldots,W_{n_+}^+$, generated by the model at (11). When $\nu_j=1$, $\mathrm{Se}^{(j,1)} = \mathrm{pr}(W_j^+ > t_0^{(j)})$ can be estimated at the parametric rate $n_+^{-1/2}$ by $\widehat{\mathrm{Se}}^{(j,1)} = n_+^{-1}\sum_{i=1}^{n_+} 1\big(W_i^+ > t_0^{(j)}\big)$. When $\nu_j > 1$, we use the ideas employed in Section 3·2. We start by rewriting $\mathrm{Se}^{(j,k)}$ at (6) as

$$\mathrm{Se}^{(j,k)} = 1 - \frac{1}{2\pi}\int_{-\infty}^{t_0^{(j)}}\int_{-\infty}^{\infty} e^{-ity}\,\phi_U(t)\,\phi^k_{B^+}(t/\nu_j)\,\phi^{\nu_j-k}_{B^-}(t/\nu_j)\,dt\,dy\,. \qquad (15)$$

Next we note that $\phi^k_{B^+}(t/\nu_j)\,\phi^{\nu_j-k}_{B^-}(t/\nu_j) = \phi^k_{W^+}(t/\nu_j)\,\phi^{\nu_j-k}_{W^-}(t/\nu_j)/\phi^{\nu_j}_U(t/\nu_j)$, where $\phi^{\nu_j}_{W^\pm}(t) \equiv \phi^k_{W^+}(t)\,\phi^{\nu_j-k}_{W^-}(t)$ can be estimated unbiasedly by

$$\widehat{\phi}^{\nu_j}_{W^\pm}(t) = \frac{1}{\binom{n_+}{k}}\,\frac{1}{\binom{n_-}{\nu_j-k}} \sum_{\ell_1<\ldots<\ell_k\,,\,\ell_{k+1}<\ldots<\ell_{\nu_j}} \exp\left\{it\big(W^+_{\ell_1}+\cdots+W^+_{\ell_k}+W^-_{\ell_{k+1}}+\cdots+W^-_{\ell_{\nu_j}}\big)\right\}.$$

Finally, we define our estimator by

$$\widehat{\mathrm{Se}}^{(j,k)} = 1 - \frac{1}{2\pi}\int_{-\infty}^{t_0^{(j)}}\int_{-\infty}^{\infty} e^{-ity}\,\widehat{\phi}^{\nu_j}_{W^\pm}(t/\nu_j)\,\phi_K(ht)\,\phi_U(t)\,\left\{\phi^{\nu_j}_U(t/\nu_j)\right\}^{-1} dt\,dy\,.$$

Remark 4. As in the case of $\widehat{\mathrm{Sp}}^{(j)}$ a simpler estimator, with the same convergence rates as $\widehat{\mathrm{Se}}^{(j,k)}$, can be computed by replacing $\widehat{\phi}^{\nu_j}_{W^\pm}$ by the empirical characteristic function of W^- raised to the power $\nu_j - k$, multiplied by the empirical characteristic function of W^+ raised to the power k. We used this estimator in Section 5.

Remark 5. As in Remark 3, when $U \equiv 0$, noting that $W^- = B^-$, $W^+ = B^+$ and $\phi_U \equiv 1$, we can define a simpler estimator which has the same parametric rates as $\widehat{\mathrm{Se}}^{(j,k)}$ above, by $\widehat{\mathrm{Se}}^{(j,k)} = \sum_{\ell_1<\ldots<\ell_k\,,\,\ell_{k+1}<\ldots<\ell_{\nu_j}} 1\{(\sum_{i=1}^k B^+_{\ell_i} + \sum_{i=k+1}^{\nu_j} B^-_{\ell_i})/\nu_j \le t_0^{(j)}\}/\{\binom{n_+}{k}\binom{n_-}{\nu_j-k}\}$.

3·4. *Estimating q*

In the Supplementary Material we prove that the likelihood of the data $Y_1^*, \ldots, Y_J^*$ is

$$L(q) = \prod_{j=1}^{J} \left\{ \sum_{k=1}^{\nu_j} \mathrm{Se}^{(j,k)} \binom{\nu_j}{k} (1-q)^k \, q^{\nu_j - k} + q^{\nu_j} \left(1 - \mathrm{Sp}^{(j)}\right) \right\}^{Y_j^*}$$
$$\times \left\{ 1 - \sum_{k=1}^{\nu_j} \mathrm{Se}^{(j,k)} \binom{\nu_j}{k} (1-q)^k \, q^{\nu_j - k} - q^{\nu_j} \left(1 - \mathrm{Sp}^{(j)}\right) \right\}^{1-Y_j^*} .$$

We suggest estimating q by $\widehat{q} = \mathrm{argmax}_q \, \widehat{L}(q)$, where $\widehat{L}(q)$ is the version of $L(q)$ obtained by replacing $\mathrm{Sp}^{(j)}$ and $\mathrm{Se}^{(j,k)}$ by the estimators $\widehat{\mathrm{Sp}}^{(j)}$ and $\widehat{\mathrm{Se}}^{(j,k)}$ developed in Sections 3·2 and 3·3.

3·5. *Fully data-driven nonparametric estimator of m*

It remains to modify the oracle estimator $\widetilde{m}$ at (9), replacing there all unknown quantities by estimators. Let $\widehat{\mathcal{D}}$ and $\widehat{\mathcal{B}}_j$ denote estimators of $\mathcal{D}$ and $\mathcal{B}_j$, obtained by replacing q by $\widehat{q}$ defined in Section 3·4, and $\mathrm{Se}_1^{(j)}$ and $\mathrm{Se}_2^{(j)}$ by $\widehat{\mathrm{Se}}_1^{(j)}$ and $\widehat{\mathrm{Se}}_2^{(j)}$, where the latter are obtained by replacing, in their definition at (8), q by $\widehat{q}$ and $\mathrm{Sp}^{(j)}$ and $\mathrm{Se}^{(j,k)}$ by the estimators $\widehat{\mathrm{Sp}}^{(j)}$ and $\widehat{\mathrm{Se}}^{(j,k)}$ defined in Sections 3·2 and 3·3. Put $\widehat{Z}_j^* = Y_j^* / \widehat{\mathcal{B}}_j$. We suggest estimating $m(x)$ by

$$\widehat{m}(x) = e_1^{\mathrm{T}} \, \mathrm{S}_N^{-1} \, \widehat{\mathrm{V}}_N - \widehat{\mathcal{D}} \,, \tag{16}$$

where S_N is as in Section 2·2 and $\widehat{\mathrm{V}}_N = (\widehat{V}_{N,0}, \ldots, \widehat{V}_{N,p})^{\mathrm{T}}$, with

$$\widehat{V}_{N,k}(x) = N^{-1} \, h_0^{-k} \sum_{j=1}^{J} \sum_{i=1}^{\nu_j} \widehat{Z}_j^* \, \mathcal{K}_{h_0}(X_{ij} - x) \, (X_{ij} - x)^k \,. \tag{17}$$

The problems of estimating the variance of $\widehat{m}(x)$, and constructing confidence intervals for that quantity, are more complex than their counterparts for conventional local polynomial estimators; and even there the problems are awkward. The main difficulties centre around choice of smoothing parameters, which should be different in each of the cases of estimating $\widehat{m}(x)$, either itself or its variance, and constructing a confidence interval. For example, the bootstrap is an attractive approach to estimating variance or computing a confidence interval, but in the context of nonparametric function estimation the bootstrap fails to estimate bias accurately, with the result that the accuracy of the variance estimator, or the coverage of the confidence interval, is compromised seriously. As a result, special methods have to be developed for implementing the bootstrap, and that places the problem beyond the scope of the present paper.

4. THEORETICAL PROPERTIES OF ESTIMATORS

4·1. *Theoretical properties of $\widehat{\mathrm{Sp}}^{(j)}$ and $\widehat{\mathrm{Se}}^{(j,k)}$*

We saw in Sections 3·2 and 3·3 that when $\nu_j = 1$, $\mathrm{Sp}^{(j)}$ and $1 - \mathrm{Se}^{(j,k)}$ are cumulative distribution functions which can be estimated by empirical cumulative distribution functions. It is well known that these have parametric convergence rates. Here we focus on the case where $\nu_j > 1$, where estimating $\widehat{\mathrm{Sp}}^{(j)}$ and $\widehat{\mathrm{Se}}^{(j,k)}$ nonparametrically from the contaminated data W_i^- and W_i^+ is related to the deconvolution problem. In the standard deconvolution problem, the asymptotic behaviour of estimators depends on the rate of decay of ϕ_U in the tails. We make the usual distinction between two cases. In the first case, which encompasses Laplace distributions and their convolutions, we assume that there exists $M > 0$ so that

$$d_0\,|t|^{-\beta-\ell} \le |\phi_U^{(\ell)}(t)| \le d_1\,|t|^{-\beta-\ell}\,,\ |t| > M,\ \ell = 0,1,2, \tag{18}$$

for finite constants $\beta > 0$ and $0 < d_0 \le d_1$. In the second case, which includes normal distributions, we assume that there exists $M > 0$ such that

$$d_0\,|t|^{\beta_0+\ell\beta-\ell}\exp(-|t|^\beta/\gamma) \le |\phi_U^{(\ell)}(t)| \le d_1\,|t|^{\beta_1+\ell\beta-\ell}\exp(-|t|^\beta/\gamma),\ |t| > M,\ \ell = 0,1,2, \tag{19}$$

for finite constants $\beta > 0$, $0 < d_0 \le d_1$ and $\beta_0 \le \beta_1$.

In deconvolution problems, the faster ϕ_U decays to zero in the tails, the slower the rates of convergence of nonparametric estimators; see Carroll and Hall (1988) and Fan (1991). For example, if ϕ_U is as at (19) then the mean squared error of estimators converges to zero at logarithmic rates. The situation is different in our case because, unlike standard problems, the targets $\mathrm{Sp}^{(j)}$ and $\mathrm{Se}^{(j,k)}$ also depend on ϕ_U; see (13) and (15). As a consequence, if ϕ_U tends to zero very fast, then as long as we choose K carefully, the bias of $\widehat{\mathrm{Sp}}^{(j)}$ and $\widehat{\mathrm{Se}}^{(j,k)}$ is particularly small, and these estimators do not suffer from the slow deconvolution convergence rates.

We make the following assumptions:

(A1) For $\ell = 0,1,2$, and for $V = B^+$ and $V = B^-$, we have $\|\phi_V^{(\ell)}\|_\infty < \infty$; and for all t, $|\phi_V^{(\ell)}(t)| \le \mathrm{const.}\,|t|^{-\alpha-\ell}$, where $1 < \alpha < \infty$ and $\mathrm{const.}$ denotes a finite positive constant;

(A2) $\int x^2\,\{f_{B^-}(x) + f_{B^+}(x) + f_U(x)\}\,dx < \infty$;

(A3) for all t, $\phi_U(t) \ne 0$;

(A4) K is the sinc kernel, defined through its Fourier transform as $\phi_K(t) = 1_{[-1,1]}(t)$.

Conditions (A1) and (A2) are rather basic, and condition (A3) is satisfied by common error distributions such as the normal. We use the sinc kernel in (A4) to benefit from the fast rates discussed above. The next theorem states that, no matter whether the errors are as at (18) or as at (19), the mean squared error, MSE, of our nonparametric estimator $\widehat{\mathrm{Sp}}^{(j)}$ converges to zero at the parametric rate. See the Supplementary Material for a proof.

THEOREM 1. *Assume that* (A1) *to* (A4) *hold.* (i) *If* ϕ_U *satisfies* (18) *with* $\beta \ge 1$ *and* $h \asymp n_-^{-1/(2\beta+\eta)}$, *where* $\eta \in (0, 2+2\alpha\nu]$, *then* $\mathrm{MSE}\big(\widehat{\mathrm{Sp}}^{(j)}\big) = O\big(n_-^{-1}\big)$. (ii) *If* ϕ_U *satisfies* (19) *with* $\beta > 1$ *and* $h < \{(\gamma/2)\,\ln n_-\}^{-1/\beta}$, *then* $\mathrm{MSE}\big(\widehat{\mathrm{Sp}}^{(j)}\big) = O\big(n_-^{-1}\big)$.

The same arguments as those used to derive Theorem 1 can be employed to prove similar results for $\widehat{\mathrm{Se}}^{(j,k)}$. In this case the convergence rate depends on the relative sizes n_- and n_+. For simplicity we assume that $n_- \asymp n_+$, which is satisfied in most realistic applications.

THEOREM 2. *Assume that* (A1) *to* (A4) *hold and* $n_- \asymp n_+$. (i) *If* ϕ_U *satisfies* (18) *with* $\beta \ge 1$ *and* $h \asymp n_-^{-1/(2\beta+\eta)}$, *where* $\eta \in (0, 2+2\alpha\nu]$, *then* $\mathrm{MSE}(\widehat{\mathrm{Se}}^{(j,k)}) = O(n_-^{-1})$. (ii) *If* ϕ_U *satisfies* (19) *with* $\beta > 1$, *and if* $h < \{(\gamma/2)\ln n_-\}^{-1/\beta}$, *then* $\mathrm{MSE}(\widehat{\mathrm{Se}}^{(j,k)}) = O(n_-^{-1})$.

4·2. *Properties of the maximum likelihood estimator* $\widehat{q}$

In Theorem 3 we derive the convergence rate of the maximum likelihood estimator $\widehat{q}$ defined in Section 3·4. As in the case of Theorem 2 this rate depends on the relative sample sizes n_- and n_+ of the two biomarker samples, and like there, to make our discussion simpler we assume in Theorem 3 that $n_- \asymp n_+$. The next theorem establishes $n_-^{-1/2}$ consistency of $\widehat{q}$. See the Supplementary Material for a statement of its conditions and for a proof.

THEOREM 3. *Assume that the conditions of Theorems* 1 *and* 2 *hold, that* $\sup_j \nu_j < \infty$*, and that* (E32)–(E33) *in the Supplementary Material hold. Let* η_{n_-} *be any sequence converging to 0 and satisfying* $n_-^{1/2}\,\eta_{n_-} \to \infty$ *as* $n_- \to \infty$*, and assume that* q_0*, the true value of* q*, lies in* $(0,1)$*. Then, with probability converging to 1 as* $n_- \to \infty$*, the equation* $\widehat{L}'(q) = 0$ *has a solution* $\widehat{q}$ *in the interval* $[q_0 - \eta_{n_-}, q_0 + \eta_{n_-}]$*, and* $\widehat{q} - q_0 = O_p(n_-^{-1/2})$.

4·3. *Properties of local polynomial estimator*

Let $\mu_j = \int u^j \mathcal{K}(u)\,du$, $\mathrm{S} = \big(\mathrm{S}_{i,j}\big)_{i,j=0,\ldots,p}$, $\widetilde{\mathrm{S}} = \big(\widetilde{\mathrm{S}}_{i,j}\big)_{i,j=0,\ldots,p}$ and $\mathrm{S}^* = \big(\mathrm{S}^*_{i,j}\big)_{i,j=0,\ldots,p}$, where $\mathrm{S}_{i,j} = \mu_{i+j}$, $\widetilde{\mathrm{S}}_{i,j} = \mu_{i+j+1}$ and $\mathrm{S}^*_{i,j} = \int u^{i+j}\mathcal{K}^2(u)\,du$. Finally, put $\mu = (\mu_{p+1},\ldots,\mu_{2p+1})^{\mathrm{T}}$ and $\widetilde{\mu} = (\mu_{p+2},\ldots,\mu_{2p+2})^{\mathrm{T}}$. We make the following assumptions, which are standard conditions for local polynomial regression procedures:

(B1) $\mathcal{K}$ is real and symmetric, $\|\mathcal{K}\|_\infty < \infty$, $\int \mathcal{K}(u)\,du = 1$, $\int |u|^{2p+3}|\mathcal{K}(u)|\,du < \infty$, $\int \big(|u|^{3p+1} + u^{4p}\big)\mathcal{K}^2(u)\,du < \infty$;

(B2) $h_0 \to 0$ and $Nh_0 \to \infty$;

(B3) $f_X(x) > 0$ and f_X is twice differentiable and satisfies $\|f_X^{(j)}\|_\infty < \infty$ for $j = 0,1,2$;

(B4) m is $p+3$ times differentiable, and $\|m^{(j)}\|_\infty < \infty$ for $j = 0,\ldots,p+3$.

As usual with local polynomial regression, instead of studying the bias and variance of $\widehat{m}(x)$ itself, which are not necessarily well defined, in Theorem 4 below, we derive asymptotic expressions for the mean and variance of a random variable $Z_N(x)$ which is asymptotically equivalent to $\widehat{m}(x) - m(x)$. Abusing terminology a little, in the sequel we shall refer to the mean and variance of $Z_N(x)$ as the bias and the variance of $\widehat{m}(x)$, respectively, and we shall write them as $\mathrm{bias}\{\widehat{m}(x)\}$ and $\mathrm{var}\{\widehat{m}(x)\}$. See the Supplementary Material for a proof of the theorem, where we also prove that $\widehat{m}(x)$ enjoys the same rate of convergence as the oracle estimator $\widetilde{m}(x)$ at (9).

THEOREM 4. *Assume the conditions of Theorem* 3*, that conditions* (B1) *to* (B4) *hold, and that* $n_- \asymp n_+ \asymp N$*. Then, for each* x*, we have* $\widehat{m}(x) - m(x) = Z_N(x)\,\{1 + o_P(1)\}$*, where* $Z_N(x)$ *is a random variable with the following properties: if* p *is odd,* $E\{Z_N(x)\} = e_1^{\mathrm{T}}\,\mathrm{S}^{-1}\,\mu\,\{(p+1)!\}^{-1}\,m^{(p+1)}(x)\,h_0^{p+1} + o(h_0^{p+1})$*; if* p *is even,*

$$E\{Z_N(x)\} = e_1^{\mathrm{T}}\mathrm{S}^{-1}\widetilde{\mu}\frac{1}{(p+2)!}\Big\{m^{(p+2)}(x) + (p+2)\,m^{(p+1)}(x)\,\frac{f_X'(x)}{f_X(x)}\Big\}\,h_0^{p+2} +\ o(h_0^{p+2}).$$

Moreover, $\mathrm{var}\{Z_N(x)\} = \{Nh_0 f_X(x)\}^{-1}\,e_1^{\mathrm{T}}\,\mathrm{S}^{-1}\,\mathrm{S}^*\mathrm{S}^{-1}\,e_1\,\big\{\mathcal{C}_B + m(x)\,\mathcal{D}_B - m^2(x)\big\}\,\{1 + o(1)\}$*, where* $\mathcal{C}_B = N^{-1}\sum_{j=1}^{J}(\nu_j\,\mathcal{A}_j/\mathcal{B}_j^2) - \mathcal{D}^2$ *and* $\mathcal{D}_B = N^{-1}\sum_{j=1}^{J}(\nu_j/\mathcal{B}_j) - 2\,\mathcal{D}$ *are to be interpreted as computed for* $q = q_0$.

It follows from this theorem that the best convergence rate is obtained by choosing h_0 so that $[E\{Z_N(x)\}]^2 \asymp \mathrm{var}\{Z_N(x)\}$. When p is odd, this bandwidth satisfies $h_0 \asymp N^{-1/(2p+3)}$, and with this bandwidth, we have $\widehat{m}(x) - m(x) = O_P(N^{-(p+1)/(2p+3)})$. When p is even, $h_0 \asymp N^{-1/(2p+5)}$, and with this bandwidth, we have $\widehat{m}(x) - m(x) = O_P(N^{-(p+2)/(2p+5)})$.

Although these results indicate that increasing the value of p improves the convergence rate of the estimator, in standard local polynomial regression it is well known that, in finite samples, increasing p tends to make the estimator too variable, and the most commonly used values are $p = 0$ and $p = 1$, with a preference for $p = 1$; see Fan and Gijbels (1996). This is also the version of our estimator that we recommend using in practice.

5. SIMULATIONS AND REAL DATA EXAMPLE

5·1. *Choosing the bandwidths for* $\widehat{\mathrm{Sp}}^{(j)}$ *and* $\widehat{\mathrm{Se}}^{(j,k)}$

We propose a simulation extrapolation approach for choosing the bandwidths needed to compute $\widehat{\mathrm{Sp}}^{(j)}$ and $\widehat{\mathrm{Se}}^{(j,k)}$. Simulation extrapolation bandwidth selectors were introduced by Delaigle and Hall (2008) in a related context of errors-in-variables regression estimators. See Cook and Stefanski (1994) for their introduction in the parametric context. We show the details only for $\widehat{\mathrm{Sp}}^{(j)}$, but it is straightforward to adapt our method for choosing a bandwidth for $\widehat{\mathrm{Se}}^{(j,k)}$. As indicated by our theory, the estimator is not very sensitive to h, and the same bandwidth can be used for all k, using for example the bandwidth computed for $\widehat{\mathrm{Se}}^{(j,2)}$ or $\widehat{\mathrm{Se}}^{(j,3)}$.

To understand how simulation extrapolation can be used in our context, write $\mathrm{Sp}^{(j)}$ as $\mathrm{Sp}^{(j)} = \int f_U(t_0^{(j)} - u)\, F_{\bar{B}_{j;0}}(u)\, du$, where $F_{\bar{B}_{j;0}}$ is the distribution function of $\sum_{i=1}^{\nu_j} B_i^-/\nu_j$. Suppose that, instead of estimating $\mathrm{Sp}^{(j)}$, our goal was to estimate $\mathrm{Sp}_*^{(j)} = \int f_U(t_0^{(j)} - u)\, F_{\bar{W}_{j;0}}(u)\, du$, where $\bar{W}_{j;0}$ has the distribution of $\sum_{i=1}^{\nu_j} W_i^-/\nu_j$. Then, since we have a sample of W_i^-s, we could estimate $F_{\bar{W}_{j;0}}(u)$ and $\mathrm{Sp}_*^{(j)}$ by $\widehat{F}_{\bar{W}_{j;0}}(u) = \sum_{\ell_1<\ldots<\ell_{\nu_j}} 1\big(\sum_{i=1}^{\nu_j} W_{\ell_i}^-/\nu_j \le u\big)/\binom{n_-}{\nu_j}$ and $\widetilde{\mathrm{Sp}}_*^{(j)} = \int f_U(t_0^{(j)} - u)\, \widehat{F}_{\bar{W}_{j;0}}(u)\, du = \sum_{\ell_1<\ldots<\ell_{\nu_j}} F_U\big(t_0^{(j)} - \sum_{i=1}^{\nu_j} W_{\ell_i}^-/\nu_j\big)/\binom{n_-}{\nu_j}$.

In practice, to speed up the calculations, we would compute this sum only for a large number, $M = 10^4$ say, of randomly chosen indices. On the other hand, we could also generate a new sample $W_{*i}^- = W_i^- + U_{*i}$, where U_{*i} is generated from f_U. Using that new sample, we could estimate $\mathrm{Sp}_*^{(j)}$ by $\widehat{\mathrm{Sp}}_*^{(j)}$, the version of $\widehat{\mathrm{Sp}}^{(j)}$ obtained by replacing each W_i^- by W_{*i}^-. Clearly, $\widetilde{\mathrm{Sp}}_*^{(j)}$ is a better estimator of $\mathrm{Sp}_*^{(j)}$ than is $\widehat{\mathrm{Sp}}_*^{(j)}$. Therefore, to choose the bandwidth h^* for computing $\widehat{\mathrm{Sp}}_*^{(j)}$ we could take $h_* = \mathrm{argmin}_h \big|\widehat{\mathrm{Sp}}_*^{(j)}(h) - \widetilde{\mathrm{Sp}}_*^{(j)}\big|$.

Similarly, if our goal was to estimate $\mathrm{Sp}_{**}^{(j)} = \int f_U(t_0^{(j)} - u)\, F_{\bar{W}_{*j;0}}(u)\, du$, where $\bar{W}_{*j;0}$ has the distribution of $\sum_{i=1}^{\nu_j} W_{*i}^-/\nu_j$, we could use $\widetilde{\mathrm{Sp}}_{**}^{(j)}$, the version of $\widetilde{\mathrm{Sp}}_*^{(j)}$ obtained by replacing each W_i^- by W_{*i}^-. We could also generate a new sample $W_{**i}^- = W_{*i}^- + U_{**i}$, where $U_{**i} \sim f_U$, and compute $\widehat{\mathrm{Sp}}_{**}^{(j)}$, the version of $\widehat{\mathrm{Sp}}_*^{(j)}$ obtained by replacing each W_{*i}^- by W_{**i}^-. To choose the bandwidth h^{**} for $\widehat{\mathrm{Sp}}_{**}^{(j)}$ we could take $h_{**} = \mathrm{argmin}_h \big|\mathrm{Sp}_{**}^{(j)}(h) - \widetilde{\mathrm{Sp}}_{**}^{(j)}\big|$. To reduce variability we can repeat the process, say R times, and replace $\big|\widehat{\mathrm{Sp}}_*^{(j)}(h) - \widetilde{\mathrm{Sp}}_*^{(j)}\big|$ and $\big|\mathrm{Sp}_{**}^{(j)}(h) - \widetilde{\mathrm{Sp}}_{**}^{(j)}\big|$ by the average of these quantities computed from R samples generated as described above. Here R does not need to be large; for example, we can take $R = 10$ or $R = 5$, which is what we used in our numerical work.

Let $h^\dagger = \mathrm{argmin}_h \big|\widehat{\mathrm{Sp}}^{(j)}(h) - \widetilde{\mathrm{Sp}}^{(j)}\big|$, where $\widetilde{\mathrm{Sp}}^{(j)}$ denotes the version of $\widetilde{\mathrm{Sp}}_*^{(j)}$ obtained by replacing each W_i^- by B_i^-, which we would compute if we had access to the B_i^-s. Clearly, $h^\dagger$ would be a good bandwidth for $\widehat{\mathrm{Sp}}^{(j)}$, but it cannot be computed since we do not observe the B_i^-s. As in Delaigle and Hall (2008), since W_{**i}^- measures W_{*i}^- in the same way as W_{*i}^- measures W_i^-, we can expect that the relationship between $\mathrm{Sp}^{(j)}$ and $\mathrm{Sp}_*^{(j)}$ is mimicked well by that between $\mathrm{Sp}_*^{(j)}$ and $\mathrm{Sp}_{**}^{(j)}$. Specifically, we can expect that the relationship between $h^\dagger$, and h_* is well approximated by that between h_* and h_{**}. That is, $h^\dagger/h_* \approx h_*/h_{**}$. This motivates us to choose the bandwidth $\widehat{h}$ for computing $\widehat{\mathrm{Sp}}^{(j)}$ by taking $\widehat{h} = (h_*)^2/h_{**}$.

5·2. *Plug-in bandwidth for $\widehat{m}$*

In this section we suggest a plug-in approach to bandwidth selection for our local linear estimator $\widehat{m}$, based on the ideas used by Ruppert et al. (1995) in the standard local linear estimation problem with independent and identically distributed data.

Recalling the notation $\mathrm{bias}\{\widehat{m}(x)\}$ and $\mathrm{var}\{\widehat{m}(x)\}$ introduced above Theorem 4, we deduce from that theorem that, in the local linear case, $\mathrm{bias}\{\widehat{m}(x)\} = \mathrm{Abias}\{\widehat{m}(x)\} + o(h_0^2)$ and $\mathrm{var}\{\widehat{m}(x)\} = \mathrm{Avar}\{\widehat{m}(x)\} + o\big\{(Nh_0)^{-1}\big\}$, where $\mathrm{Abias}\{\widehat{m}(x)\} = h_0^2\mu_2\, m''(x)/2$, $\mathrm{Avar}\{\widehat{m}(x)\} = R(\mathcal{K})\,\{Nh_0 f_X(x)\}^{-1}\,\big\{\mathcal{C}_B + m(x)\,\mathcal{D}_B - m^2(x)\big\}$, and $R(\mathcal{K}) = \int \mathcal{K}^2(u)\,du$. As in more standard nonparametric regression problems, we base our bandwidth selection method on the following weighted theoretical criterion:

$$\begin{aligned}\mathrm{AMISE}_w \equiv & \int_a^b \big[\mathrm{Abias}\{\widehat{m}(x)\}\big]^2 f_X(x)\,dx + \int_a^b \mathrm{Avar}\{\widehat{m}(x)\}\, f_X(x)\,dx \\ = & h_0^4\,\mu_2^2\,\theta_2/4 + R(\mathcal{K})(Nh_0)^{-1}\,v\,,\end{aligned}$$

where $\theta_2 = \int_a^b \{m''(x)\}^2 f_X(x)\,dx$, $v = \int_a^b \big\{\mathcal{C}_B + m(x)\,\mathcal{D}_B - m^2(x)\big\}\,dx$, and a and b are the empirical quantiles $0{\cdot}05$ and $0{\cdot}95$ of the distribution of X.

Of course, AMISE_w depends on unknown quantities which have to be estimated. In practice, we choose h_0 by minimising the following estimator of AMISE_w:

$$\widehat{\mathrm{AMISE}}_w = h_0^4\,\mu_2^2\,\widehat{\theta}_2/4 + R(\mathcal{K})(Nh_0)^{-1}\,\widehat{v}\,,$$

where $\widehat{v} = \int_a^b \big\{\widehat{\mathcal{C}}_B + \widehat{m}_0(x)\,\widehat{\mathcal{D}}_B - \widehat{m}_0^2(x)\big\}\,dx$, with $\widehat{m}_0$ denoting a pilot estimator of m, $\widehat{\mathcal{C}}_B$ and $\widehat{\mathcal{D}}_B$ denoting the estimators of $\mathcal{C}_B$ and $\mathcal{D}_B$ obtained by replacing q, $\mathrm{Sp}^{(j)}$, $\mathrm{Se}_1^{(j)}$ and $\mathrm{Se}_2^{(j)}$ by their estimators, and where, using the notation $w(x) = 1_{[a,b]}(x)$, $\widehat{\theta}_2 = N^{-1}\,\sum_{j=1}^J \sum_{i=1}^{\nu_j} \{\widehat{m}''_{(-j)}(X_{ij})\}^2\, w(X_{ij})$, with $\widehat{m}''_{(-j)}$ denoting a local polynomial estimator of m'' of order $p = 3$, constructed without employing the observations from the jth group, and using a bandwidth $h_2 \neq h_0$; see the Supplementary Material for details.

Since the pilot estimator $\widehat{m}_0$ appears inside an integral, it can be rather crude. We use a quadratic spline estimator with a small number, $\kappa = 5$ say, of knots. Let $\beta = (\beta_0, \beta_1, \beta_2, \beta_{21}, \ldots, \beta_{2\kappa})^{\mathrm{T}}$, and let $m_0(x \,|\, \beta) = \sum_{j=0}^2 \beta_j\, x^j + \sum_{k=1}^{\kappa} \beta_{2k}\,(x - \xi_k)_+^2$ be a quadratic spline with κ knots. Recalling (7) we take $\widehat{m}_0(x) = m_0(x \,|\, \widehat{\beta})$, where

$$\widehat{\beta} = \mathrm{argmin}_\beta \sum_{j=1}^J \sum_{i=1}^{\nu_j} \big\{Y_j^* \widehat{\mathcal{B}}_j^{-1} - \widehat{\mathcal{A}}_j\, \widehat{\mathcal{B}}_j^{-1} - m_0(X_{ij} \,|\, \beta)\big\}^2\,.$$

5·3. *Choosing $t_0^{(j)}$ from the data*

The value of $t_0^{(j)}$ affects the quality of the estimator $\widehat{m}$. It follows from Theorem 4 that $t_0^{(j)}$ affects the asymptotic variance term of $\widehat{m}(x)$, but not its asymptotic bias term. Therefore, to optimise asymptotic performance of $\widehat{m}(x)$ we can choose $t_0^{(j)}$ so as to minimise an estimator of its asymptotic variance term. Motivated by the fact that, for each x, this term depends on $t_0^{(j)}$ only through $\mathcal{C}_B + m(x)\,\mathcal{D}_B$, if the value of $t_0^{(j)}$ is open to choice, we suggest choosing it by minimising an estimator of $\mathcal{C}_B + E\{m(X)\}\,\mathcal{D}_B$. Recalling that $q = 1 - E\{m(X)\}$, we choose $t_0^{(j)}$ that minimises $\widehat{\mathcal{C}}_B + (1 - \widehat{q})\,\widehat{\mathcal{D}}_B$ under the constraint that $\widehat{\mathrm{Sp}}^{(j)} \geq 0{\cdot}5$ and $\widehat{\mathrm{Se}}^{(j,\nu_j)} \geq 0{\cdot}5$ and where $\widehat{q}$, $\widehat{\mathcal{C}}_B$ and $\widehat{\mathcal{D}}_B$ are as in Sections 3·4 and 5·2.

Table 1. *Simulation results for models (i) to (iv), in the case where* $\tau = \{var(B^+)\}^{1/2} = 0{\cdot}2$. *The numbers show* $10^3\times$ *median integrated squared error (interquartile range) calculated from 200 simulated samples. In each group of three rows, the first row is for our estimator* $\widehat{m}$ *with* $t_0^{(j)}$ *as in Section 5·3, the second is for Delaigle and Meister's (2011)* $\widehat{m}_{\rm DM}$ *estimator and the third is for McMahan et al.'s (2013) parametric estimator with* $t_0^{(j)} = 0{\cdot}2$.

		$\nu=5$			$\nu=10$			$\nu=15$	
Model	$J=200$	$J=500$	$J=1000$	$J=200$	$J=500$	$J=1000$	$J=200$	$J=500$	$J=1000$
(i)	4·0(4·7)	2·1(2·4)	1·1(0·9)	5·0(6·5)	2·3(3·4)	1·5(1·6)	4·7(6·0)	2·5(2·8)	1·3(1·5)
	4·2(5·6)	1·8(1·7)	1·0(1·1)	19(13)	20(8)	20(5)	29(13)	29(9)	28(6)
	1·5(2·4)	0·5(1·1)	0·2(0·5)	1·5(3·2)	0·7(1·6)	0·4(0·8)	1·11(2)	0·4(1·1)	0·3(0·6)
(ii)	4·5(5·8)	2·0(2·1)	1·2(1·0)	4·9(5·9)	2·5(3·0)	1·5(1·2)	4·5(5·1)	2·5(2·2)	1·4(1·5)
	3·7(4·4)	2·0(2·2)	1·0(0·9)	13(9)	13(5)	12(4)	19(8)	18(5)	18(4)
	1·6(2·6)	0·7(1·0)	0·4(0·1)	1·1(3·8)	1·2(1·7)	0·6(0·7)	1·9(3·7)	1·1(1·5)	0·5(0·8)
(iii)	26(16)	14(9)	9(7)	26(18)	15(9)	10(7)	25(17)	17(11)	10(6)
	26(17)	16(9)	11(5)	55(24)	50(16)	43(12)	67(25)	59(21)	54(16)
	40(5)	38(2)	37(1)	40(7)	38(3)	37(2)	40(8)	38(3)	38(2)
(iv)	18(11)	10(7)	6(3)	19(14)	10(7)	7(4)	20(16)	11(7)	7(4)
	20(16)	11(8)	7(5)	43(17)	36(14)	36(10)	52(24)	49(16)	44(12)
	40(62)	38(35)	38(2)	40(7)	38(3)	38(2)	40(9)	38(4)	38(3)

5·4. *Simulated models*

Following McMahan et al. (2013) we took $B^- \sim N(0{\cdot}1, 0{\cdot}02^2)$ and $B^+ \sim N(1, \tau^2)$, where $\tau = 0{\cdot}1$, $0{\cdot}2$ or $0{\cdot}3$, and generated $W_1^-, \ldots, W_{n_-}^-$ and $W_1^+, \ldots, W_{n_+}^+$ as at (11), where $U_i^+ \sim N(0, 0{\cdot}01^2)$ and $n_- = n_+ = 200$. The true status $\widetilde{Y}_{ij}$ was generated from a Bernoulli distribution, specifically $\widetilde{Y}_{ij} \mid X_{ij} \sim \mathrm{Be}\{m(X_{ij})\}$, as in (1). We used four models:
(i) $m(x) = a(x)/\{1 + a(x)\}$ where $a(x) = \exp(-3 + 2x)$, and $X_{ij} \sim N(0, 0{\cdot}75^2)$;
(ii) $m(x) = b(x)/\{1 + b(x)\}$ where $b(x) = \exp(-3 + x + 0{\cdot}5\, x^2)$, and $X_{ij} \sim N(0, 0{\cdot}75^2)$;
(iii) $m(x) = c(x)/\{1 + c(x)\}$ with $c(x) = a(x)\{2 + \cos(4x - 0{\cdot}5)\}^2$ and $X_{ij} \sim N(0, 0{\cdot}75^2)$;
(iv) $m(x) = c(x)/\{1 + c(x)\}$ and $X_{ij} \sim U[-1{\cdot}6, 1{\cdot}6]$.
Models (i) and (ii) are simple and were considered by McMahan et al. (2013). In models (iii) and (iv), we added a level of complexity by introducing a cosine function.

In the main sample, defined in Section 2·1, we partitioned the data into J groups of sizes $\nu = 5$, 10 or 15, for $J = 200$, 500 or 1000. We generated the W_js as in (2), where $U_j \sim N(0, 0{\cdot}01^2)$. Finally, we took $Y_j^* = 1(W_j > t_0^{(j)})$. For our estimator $\widehat{m}$ we took $t_0^{(j)}$ as in Section 5·3, and for the other estimators, we took $t_0^{(j)} = 0{\cdot}2$ as in McMahan et al. (2013).

In each case we generated 200 samples. For each sample we computed our local linear estimator of m, that is, $\widehat{m}$ at (16) with $p = 1$, using the bandwidths introduced in Sections 5·1 and 5·2, and where $\widehat{\mathrm{Sp}}^{(j)}$ and $\widehat{\mathrm{Se}}^{(j,k)}$ were computed as in Remarks 2 and 4. We also computed the local linear estimator $\widehat{m}_{\rm DM}$ of Delaigle and Meister (2011), using their plug-in bandwidth. This estimator ignores the dilution effect and assumes that we observe the true statuses. Since m is a probability, in each case we truncated these estimators to 0 or 1. We also considered the estimator in Delaigle and Meister's (2011) Section 5, where an error correction is applied without taking dilution into account, assuming that specificity and sensitivity do not depend on group sizes. With data from our model, these could be estimated by an average of specificity and sensitivity over various group sizes. However, in the examples we considered, this corrected estimator performed too poorly to be considered here. Finally, we computed the parametric estimator $\widehat{m}_{\rm P}$ of McMahan et al. (2013), using the correct parametric model for m, f_U and the biomarker distribution in cases (i) and (ii), and using the correct parametric model for f_U and the biomarker distribution, but the incorrect first order logistic regression in cases (iii) and (iv).

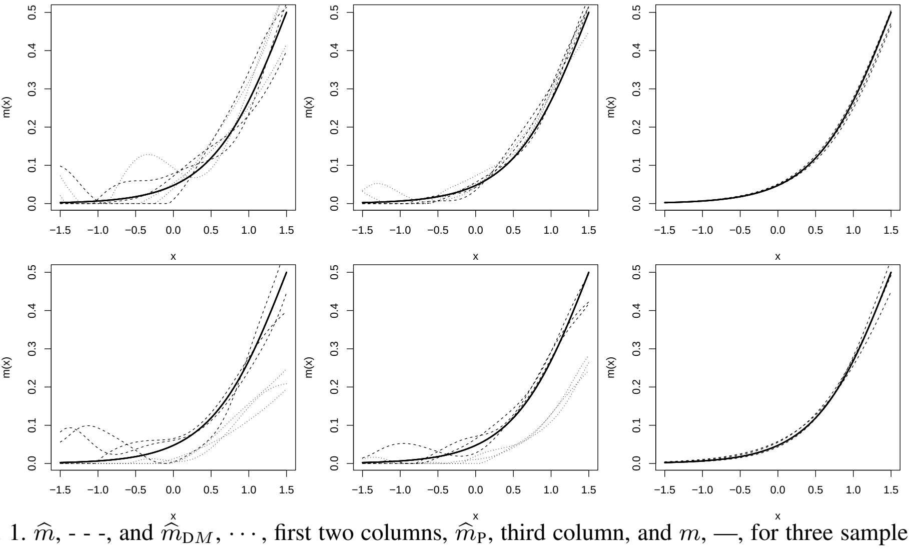

Fig. 1. $\widehat{m}$, - - -, and $\widehat{m}_{DM}$, $\cdots$, first two columns, $\widehat{m}_{\rm P}$, third column, and m, —, for three samples coming from model (i) with $\tau = 0{\cdot}2$, and corresponding to the first three quartiles of $\text{ISE}_{\rm NEW}$, $\text{ISE}_{\rm DM}$ and $\text{ISE}_{\rm P}$, respectively, when $\nu = 5$, row 1, and $\nu = 15$, row 2, and $J = 200$, first column, and $J = 1000$, last two columns.

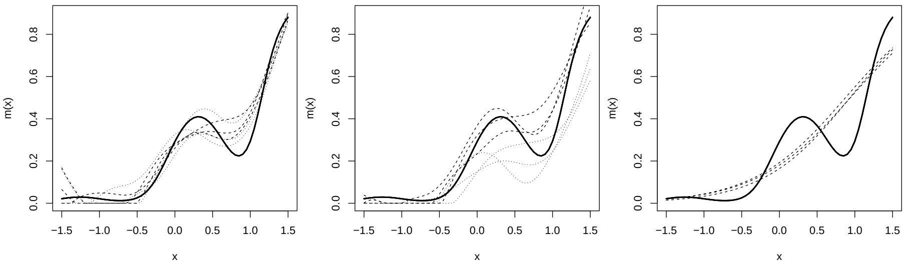

Fig. 2. $\widehat{m}$, - - -, and $\widehat{m}_{DM}$, $\cdots$, first two columns, $\widehat{m}_{\rm P}$, third column, and m, —, for three samples coming from model (iv) with $\tau = 0{\cdot}2$, and corresponding to the first three quartiles of $\text{ISE}_{\rm NEW}$, $\text{ISE}_{\rm DM}$ and $\text{ISE}_{\rm P}$, respectively, when $J = 500$, and $\nu = 5$, first column, and $\nu = 15$, last two columns.

To assess performance, in each case we computed 200 values of the integrated squared errors $\text{ISE}_{\rm NEW} = \int_c^d (\widehat{m} - m)^2$, $\text{ISE}_{\rm DM} = \int_c^d (\widehat{m}_{\rm DM} - m)^2$ and $\text{ISE}_{\rm P} = \int_c^d (\widehat{m}_{\rm P} - m)^2$, where $[c, d] = [-1{\cdot}5, 1{\cdot}5]$. Table 1 reports, for each method and each model, the median and the interquartile range of the 200 integrated squared errors in the case where $\tau = \{\text{var}(B^+)\}^{1/2} = 0{\cdot}2$. The results in other cases were similar; see the Supplementary Material. The table indicates that, in these examples, when the group size was less than $\nu = 5$, our new method worked slightly better or slightly worse than the one that ignores dilution and errors. However, as ν increased, the advantage of using our method became clearer. For example, when $\nu = 15$ the median $\text{ISE}_{\rm NEW}$ was up to 25 times smaller than the median $\text{ISE}_{\rm DM}$. The comparison with the parametric estimator $\widehat{m}_{\rm P}$ is as expected: in cases (i) and (ii), the parametric model is correct and $\widehat{m}_{\rm P}$ outperforms $\widehat{m}$; in cases (iii) and (iv), $\widehat{m}_{\rm P}$ targets the wrong curve and provides a strongly biased estimator, and our new, consistent, estimator $\widehat{m}$ performs considerably better; see also Fig. 1 and Fig. 2.

Figure 1 depicts, for the three estimators, the estimated curves corresponding to the samples, coming from model (i), that resulted in the first three quartiles of the 200 values of $\text{ISE}_{\rm NEW}$, $\text{ISE}_{\rm DM}$ and $\text{ISE}_{\rm P}$ when $\nu = 5$ and 15, and $J = 200$ and 1000. When $\nu = 5$, we can see that the two nonparametric methods gave very similar results, but when $\nu = 15$ our new method significantly outperformed $\widehat{m}_{\rm DM}$. The same conclusions can be drawn from Fig. 2, where we show estimated curves for model (iv) when $J = 500$, and $\nu = 5$ and 15. The figures also depict

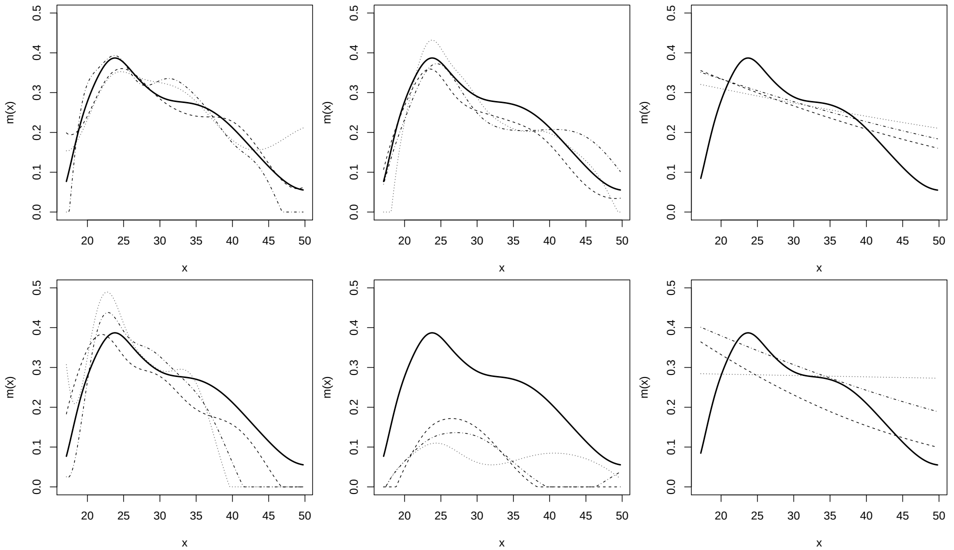

Fig. 3. Estimated curves for hepatitis C using the estimator $\widehat{m}$, left, $\widehat{m}_{\rm DM}$, middle, or $\widehat{m}_{\rm P}$, right, for three samples of the Irish prisons data, corresponding to the first, — — —, second, — · —, and third, · · ·, quartiles of ISE$_{\rm NEW}$, ISE$_{\rm DM}$ and ISE$_{\rm P}$, respectively, when $\nu = 2$, first row, and $\nu = 6$, second row. The thick lines depict $\widehat{m}_{\rm LL}$.

the estimator $\widehat{m}_{\rm P}$. As noted above, it performs better than $\widehat{m}$ in Fig. 1 where the parametric model is correct, but is strongly biased in Fig. 2, where the parametric model is incorrectly specified.

5·5. *Illustration with Irish prisons data*

As in McMahan et al. (2013) we applied our technique to data collected on prisoners from nine Irish prisons; see Allwright et al. (2000) for a description of the data and collection method. The dataset contains information about the presence of antibodies to hepatitis B core antigen and to hepatitis C virus, biomarker readings and individual covariates. As in McMahan et al. (2013), our goal is to estimate the prevalence of antibodies to hepatitis B core antigen given age, but we also wish estimate the conditional prevalence of antibodies to hepatitis C virus. Let X be the age of a prisoner; for each type of hepatitis, let $\widetilde{Y}$ be the true status, here absence, $\widetilde{Y} = 0$, or presence, $\widetilde{Y} = 1$, of antibodies; and let B denote the biomarker, which, for both types of hepatitis, was obtained through oral fluid testing procedures.

For hepatitis B, we have a sample of $N = 1098$ prisoners for which $\widetilde{Y}$, X and B were each available, with $N_+ = 60$ positive individuals. For hepatitis C, we have $N = 1009$ observations with $N_+ = 290$ positive individuals. Following McMahan et al. (2013), for both types of hepatitis, we pooled the data randomly in groups of size 2, 4, 6 and 10. Although this paper is about random pooling, we should note that, when possible, homogeneous pooling of the type considered by Delaigle and Hall (2012) is preferable because, as demonstrated there, it usually results in better estimators. To assess the performance of our procedure, we did this 200 times and created in each case 200 samples of pooled data. For each sample, we computed our estimator $\widehat{m}$ and Delaigle and Meister's (2011) $\widehat{m}_{\rm DM}$ estimator which ignores error and dilution. In each case we used the bandwidths described in Section 5·4 and took $t_0^{(j)}$ as in Section 5·3. We also computed the standard local linear estimator $\widehat{m}_{\rm LL}$ from the non-grouped data $(X_i, \widetilde{Y}_i)$, for $i = 1, \ldots, N$, using the true statuses and a plug-in bandwidth. Since this estimator uses non-grouped data and the true statuses, it is much closer to the true curve and we shall treat $\widehat{m}_{\rm LL}$ as the truth. We estimated specificities and sensitivities as in Remarks 3 and 5, using a subsample of n_- negative Bs and n_+ positive Bs, drawn randomly from the N_- and the N_+ available individuals, respectively. We took n_- equal to 100 and 20 in the cases of hepatitis C and B, respectively, and we took $n_+ = 100$ in both cases.

As usual, it is difficult to compare parametric and nonparametric estimators. Parametric estimators converge fast to the truth, but only when one can determine the correct parametric model. Nonparametric estimators are consistent regardless of parametric assumptions, but they have slower convergence rates. Thus, in practice, when it is possible to make good parametric approximations to the truth, parametric estimators usually work better, but otherwise, they are biased and nonparametric estimators are expected to perform better. Of course, guessing a correct parametric model is challenging for binary response variables, especially when the only information available comes from grouped data. To illustrate this, we computed the parametric first order logistic estimator $\widehat{m}_{\mathrm{P}}$ of McMahan et al. (2013) with $t_0^{(j)}$ chosen as at their page 293 and estimating specificities and sensitivities parametrically like them, using the subsample of size $n_- + n_+$ defined above and assuming, like them, that the biomarkers have a lognormal distribution. The gamma and Weibull distributions they also considered lead to similar conclusions.

We summarize the results for two values of ν in Fig. 3 for hepatitis C, and in the Supplementary Material for hepatitis B. There, each figure shows $\widehat{m}_{\mathrm{LL}}$, and three estimated curves obtained using $\widehat{m}$, $\widehat{m}_{\mathrm{DM}}$ or $\widehat{m}_{\mathrm{P}}$. These curves were chosen among the 200 estimated curves as those corresponding to the first three quartiles of the 200 values of the integrated squared error $\mathrm{ISE}_{\mathrm{NEW}} = \int_c^d (\widehat{m} - \widehat{m}_{\mathrm{LL}})^2$ in the case of $\widehat{m}$, $\mathrm{ISE}_{\mathrm{DM}} = \int_c^d (\widehat{m}_{\mathrm{DM}} - \widehat{m}_{\mathrm{LL}})^2$ for $\widehat{m}_{\mathrm{DM}}$ and $\mathrm{ISE}_{\mathrm{P}} = \int_c^d (\widehat{m}_{\mathrm{P}} - \widehat{m}_{\mathrm{LL}})^2$ for $\widehat{m}_{\mathrm{P}}$, where c and d were, respectively, the 0·025 and 0·975 empirical quantiles of the X_is. For both types of hepatitis, the shape of $\widehat{m}_{\mathrm{LL}}$ confirms Allwright et al.'s (2000) findings: for hepatitis B, prevalence increases to reach a peak around age 35, and then decreases, whereas for hepatitis C, the peak is reached around age 25, then decreases to reach a low value at age 45. Overall, $\widehat{m}$ is able to capture the right shape, but $\widehat{m}_{\mathrm{P}}$ is not.

The figures show that the improvement of our estimator over the parametric estimator $\widehat{m}_{\mathrm{P}}$ is substantial in all cases, and indicate that the logistic model is not a good approximation to the truth. In the cases of hepatitis C, as ν increases, the specificities $\widehat{\mathrm{Se}}^{(j,k)}$ differ widely over k, with some of them being far from 1, and so our estimator improves on $\widehat{m}_{\mathrm{DM}}$ substantially. However, in the case of hepatitis B, $\widehat{\mathrm{Sp}}^{(j)}$ and $\widehat{\mathrm{Se}}^{(j,k)}$ are all very close to 1, so that the improvement of our estimator over $\widehat{m}_{\mathrm{DM}}$ is marginal. For $\nu = 10$, our estimator $\widehat{m}$ also outperformed $\widehat{m}_{\mathrm{P}}$ significantly, but the number of groups, J, was too small and all three methods performed poorly.

Acknowledgement

Research supported by the Australian Research Council. We thank Professor Shane Allwright for kindly providing the Irish prisons data.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs as well as additional methodological development and practical results.

References

Allwright, S., Bradley, F., Long, J., Barry, J., Thornton, L. & Parry, J. (2000). Prevalence of antibodies to hepatitis B, hepatitis C, and HIV and risk factors in Irish prisoners: results of a national cross sectional survey. *Brit. Med. J.*, **321**, 78-82.

CARROLL, R.J. & HALL, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, **83**, 1184–1186.

CHEN, P., TEBBS, J.M. & BILDER, C.R. (2009). Group testing regression models with fixed and random effects. *Biometrics*, **65**, 1270–1278.

CHEN, C.L. & SWALLOW, W.H. (1990). Using group testing to estimate a proportion, and to test the binomial model. *Biometrics*, **46**, 1035–1046

COOK, J. R., AND STEFANSKI, L.A. (1994), Simulation-extrapolation estimation in parametric measurement error models. *J. Amer. Statist. Assoc.*, **89**, 1314–1328.

DELAIGLE, A., FAN, J. & CARROLL, R.J. (2009). A design-adaptive local polynomial estimator for the errors-in-variables problem. *J. Amer. Statist. Assoc.*, **104**, 348–359.

DELAIGLE, A. & GIJBELS, I. (2004). Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Ann. Inst. Statist. Math.*, **56**, 19–47.

DELAIGLE, A. & HALL, P. (2008). Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *J. Amer. Statist. Assoc.*, **103**, 280–287.

DELAIGLE, A. & HALL, P. (2012). Nonparametric regression with homogeneous group testing data. *Ann. Statist.*, **40**, 131–158.

DELAIGLE, A. & HALL, P. (2015). Methodology for non-parametric deconvolution when the error distribution is unknown. *J. Roy. Statist. Soc.* Ser. B, DOI: 10.1111/rssb.12109, in press.

DELAIGLE, A., HALL, P. & MEISTER, A. (2008). On deconvolution with repeated measurements. *Ann. Statist.*, **36**, 665–685.

DELAIGLE, A., HALL, P. & WISHART, J. (2014). New approaches to non- and semi-parametric regression for univariate and multivariate group testing data. *Biometrika*, **101**, 567–585.

DELAIGLE, A. & MEISTER, A. (2011). Nonparametric regression analysis for group testing data. *J. Amer. Statist. Assoc.*, **106**, 640–650.

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

FAN, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, **19**, 1257–1272.

FAN, J., AND GIJBELS, I. (1996). *Local Polynomial Modelling and its Applications*, Chapman & Hall, London.

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

GASTWIRTH, J.L. & HAMMICK, P.A. (1989). Estimation of prevalence of a rare disease, preserving anonymity of subjects by group testing: application to estimating the prevalence of AIDS antibodies in blood donors. *J. Statist. Plann. Inf.*, **22**, 15–27.

LI, M. & XIE, M. (2012). Nonparametric and semiparametric regression analysis of group testing samples. *Int. J. Stats. Med. Res.*, **1** 60–72.

MCMAHAN, C.S., TEBBS, J.M. & BILDER, C.R. (2013). Regression models for group testing data with pool dilution effects. *Biostatistics*, **14**, 284–298.

RUPPERT, D., SHEATHER, S.J. & WAND, M.P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.*, **90**, 1257–1270.

STEFANSKI, L. & CARROLL, R.J. (1990). Deconvoluting kernel density estimators. *Statistics*, **21**, 169–184.

VANSTEELANDT, S., GOETGHEBEUR, E. & VERSTRAETEN, T. (2000). Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics*, **56**, 1126–1133.

XIE, M. (2001). Regression analysis of group testing samples. *Statist. Med.*, **20**, 1957–1969.

WEIN, L.M. & ZENIOS, S. (1996). Pooled testing for HIV screening: capturing the dilution effect. *Oper. Res.*, **44**, 543–569.

ZENIOS, S. & WEIN, L.M. (1998). Pooled testing for HIV seroprevalence estimation: exploiting the dilution effect. *Statist. Med.*, **17**, 1447–1467.