Nonparametric regression analysis for group testing data

Aurore Delaigle and Alexander Meister *

Abstract: Group testing is a procedure employed to reduce the cost and increase the speed of large screening studies where infection or contamination of individuals is detected by a test carried out on a sample of, for example, blood, urine, water, etc. Instead of testing the sample of each individual, the method consists in pooling samples of groups of several individuals, and test those pooled samples. We construct a nonparametric procedure for estimating the conditional probability of contamination given an explanatory variable, when the observations are pooled data of this type. We investigate asymptotic theoretical properties of the estimator and establish its consistency. The procedure requires the selection of an important smoothing parameter, and we suggest a way for choosing it automatically from the data. We illustrate the numerical performance of the method on some simulated examples and on data from the National Health and Nutrition Examination Survey. We discuss extensions of the procedure to cases where the test is imprecise and the covariates are observed inaccurately, and to the multivariate setting. Supplemental materials including proofs, R codes and additional simulation results are available from the online JASA website.

Keywords: bandwidth selection; binary regression; Dorfman screening; kernel estimator; local polynomial estimator; pooled data; prevalence.
AMS subject classification: 62G08

*Aurore Delaigle is a principal researcher and Queen Elizabeth II fellow, Department of Mathematics and Statistics, University of Melbourne, VIC, 3010, Australia (E-mail: A.Delaigle@ms.unimelb.edu.au). Alexander Meister is Professor, Institut für Mathematik, Universität Rostock, D-18051 Rostock, Germany. Delaigle's research was supported by grants and a fellowship from the Australian Research Council. The authors thank the editor, the associate editor, and two referees for their valuable comments that helped improve a previous version of the manuscript. Address for correspondence: Aurore Delaigle, Department of Mathematics and Statistics, University of Melbourne, VIC, 3010, Australia

1 Introduction

We consider nonparametric estimation of a conditional probability from group testing data, where observations are pooled in groups before a detection test (with results $Y = 0$ for negative and $Y = 1$ for positive) is applied. Pooling data in groups was originally suggested by Dorfman (1943) for detecting syphilis in U.S. soldiers during the Second World War, and has since been applied to a great many studies. With this group screening technique, in order to detect a disease in a population, instead of testing separately samples of blood or urine of each individual, the test is applied to pooled samples of groups of several individuals, which permits to reduce the number of tests to apply and hence to save time and money. Pooling observations in groups is also employed to detect pollution (for example in water, milk, etc.), where samples (e.g. of water or milk) are pooled together before being tested for contamination by a toxic substance (see for example Nagi and Raggi, 1972, Wahed et al., 2006, Lennon, 2007 and Fahey et al., 2007).

This grouping method has stimulated a great deal of research activity, and one of the interesting problems studied in the literature concerns the estimation of the conditional probability $p(x) = P(Y = 1|X = x)$, where X is an explanatory variable. For example, if we are interested in the prevalence of a disease, X could be the age, the weight, the absorption of fat etc., whereas pollution in water could be explained by variables such as the size of surrounding crop areas or livestock. The problem is not a standard regression estimation one because the variable Y is not observed for each individual, as only the disease or pollution status of groups is available. Research on this and related problems includes the work of Farrington (1992), Vansteelandt et al. (2000), Xie (2001), Bilder and Tebbs (2009) and Chen et al. (2009). Other related work includes Gastwirth and Hammick (1989), Chen and Swallow (1990), Gastwirth and Johnson (1994), Hardwick et al. (1998) and Hung and Swallow (2000). However, these papers suggest methods based on parametric models, where the shape of the

regression curve is known in advance and only finitely many real-valued parameters have to be estimated. In this paper, we construct a nonparametric procedure for estimating $p(x)$ from data pooled in groups which are not necessarily of equal size.

This paper is organised as follows. We introduce our nonparametric procedure for estimating p in Section 2, and we investigate its theoretical asymptotic properties in Section 3. Section 4 is devoted to numerical properties of the estimator. In Section 4.1 we suggest a data-driven selector of the bandwidth, a smoothing parameter required to calculate the estimator. In Section 4.2 we illustrate the finite sample performance of our method on simulated data, and in Section 4.3 we apply our procedure to data from the National Health and Nutrition Examination Survey (NHANES). In Section 5 we show how to modify our estimator when the detection test is not perfectly accurate. Finally, we discuss two extensions of our work in Section 6: the case where the covariate is measured inaccurately, and the multivariate setting. Proofs of our results are derived in a supplemental file available online from the JASA website.

2 Model and methodology

Let N be the total number of individuals in the study. The statistical model for group testing procedures can be described as follows. First, the N individuals are divided randomly into J groups of sizes n_j, for $j = 1, \ldots, J$. Then, the specimen (blood, urine, water, etc.) that we want to test for contamination or infection are pooled together in groups before the test is applied. In other words, the result, $Y_{i,j} = 0$ or 1, of the test for the ith individual in the jth group is not observed, and instead we observe the result Y_j^* of the test carried out on the entire jth group. That is, we observe

$$Y_j^* = \max_{i=1,\ldots,n_j} Y_{i,j} \,. \tag{2.1}$$

In addition, we observe an explanatory variable X on each individual, that is, we observe $X_{i,j}$ for $i = 1, \ldots, n_j$, $j = 1, \ldots, J$. Here, X does not usually come from a test and the $X_{i,j}$'s are not usually pooled. We assume that the $(X_{i,j}, Y_{i,j})$'s are independent and identically distributed (i.i.d.). Since $Y_{i,j} = 0$ or 1, $Y_{i,j}|X_{i,j}$ has a Bernoulli distribution with parameter $p(X_{i,j})$, where p is a function taking values on the interval $[0, 1]$. We wish to estimate the function p from the sample of observations $(X_{i,j}, Y_j^*)$, $i = 1, \ldots, n_j$, $j = 1, \ldots, J$. This problem has been considered by other authors before, but only in the case where p is estimated parametrically; see the introduction for a list of references. Our goal is to take a nonparametric approach for estimating p.

In order to do this, first note that if the observations $(X_{i,j}, Y_{i,j})$, $i = 1, \ldots, n_j$, $j = 1, \ldots, J$, were available, we could estimate p nonparametrically by the local polynomial regression estimator; see e.g. Fan and Gijbels (1996) for an introduction. The idea of this technique is to approximate the regression curve of interest, here $p(x) = E(Y|X = x)$, locally at each point x by an ℓth order polynomial, and to fit the curve locally at each x by weighted least squares. The locality is controlled by a smoothing parameter $h > 0$ called the bandwidth, and the weights are values of $K_h(\cdot) = h^{-1}K(\cdot/h)$, where K is a smooth and symmetric function called the kernel. For $\ell \geq 0$ an integer, the ℓth order local polynomial estimator of p can be written as $\widehat{p}_\ell(x) = \mathbf{e}_0^T \mathcal{S}^{-1} \mathcal{T}$, where $\mathbf{e}_0 = (1, 0, \ldots, 0)^T$, $\mathcal{S} = \left(S_{k,k'}\right)_{0 \leq k,k' \leq \ell}$ and $\mathcal{T} = (T_0, \ldots, T_\ell)^T$, with, for $k, k' = 0, \ldots, \ell$, $S_{k,k'} = (Nh^{k+k'})^{-1} \sum_{j=1}^{J} \sum_{r=1}^{n_j} K_h(x - X_{r,j})(x - X_{r,j})^{k+k'}$ and $T_k = (Nh^k)^{-1} \sum_{j=1}^{J} \sum_{r=1}^{n_j} Y_{i,j} K_h(x - X_{r,j})(x - X_{r,j})^k$. See Fan and Gijbels (1996).

Of course, in our case, we cannot calculate this estimator since we do not observe the $Y_{i,j}$'s. Instead of estimating p directly, we suggest estimating another function g which is empirically accessible from the type of data we have, and from which we can

reconstruct p. To derive such a function, we first note that

$$\begin{aligned} E(Y_j^*|X_{1,j},\ldots,X_{n_j,j}) &= P(Y_j^* = 1|X_{1,j},\ldots,X_{n_j,j}) \\ &= 1 - P(Y_{1,j} = 0,\ldots,Y_{n_j,j} = 0|X_{1,j},\ldots,X_{n_j,j}) \\ &= 1 - \prod_{i=1}^{n_j} P(Y_{i,j} = 0|X_{1,j},\ldots,X_{n_j,j}) \\ &= 1 - \prod_{i=1}^{n_j}\{1 - p(X_{i,j})\}. \end{aligned}$$

Using the notations $Z_j^* = 1 - Y_j^*$ and $q = E\{1 - p(X_{1,1})\}$, and taking, respectively, the expectation and the conditional expectation given $X_{i,j}$, of the last expression, we deduce that $EZ_j^* = q^{n_j}$ and $E(Z_j^*|X_{i,j}) = q^{n_j-1}\{1 - p(X_{i,j})\}$ for $i = 1,\ldots,n_j$. Hence we have

$$\sum_{j=1}^{J}\sum_{r=1}^{n_j} E\big(Z_j^*|X_{r,j} = x\big) \,/\, \sum_{j=1}^{J} n_j EZ_j^* = \{1 - p(x)\}/q\,.$$

Let $\mu_Z^* = N^{-1}\sum_{j=1}^{J} n_j EZ_j^* = N^{-1}\sum_{j=1}^{J} n_j q^{n_j} = qM/N$, with $M = \sum_{j=1}^{J} n_j q^{n_j-1}$, and remember that N is the sample size, that is, $N = \sum_{j=1}^{J} n_j$. The above calculations show that

$$p(x) = 1 - q \cdot g(x)/\mu_Z^*\,, \tag{2.2}$$

where

$$g(x) = N^{-1}\sum_{j=1}^{J}\sum_{r=1}^{n_j} E\big(Z_j^*|X_{r,j} = x\big) = (M/N)\cdot\big\{1 - p(x)\big\}\,.$$

Therefore, to estimate p, it suffices to construct estimators of g, q and μ_Z^*. Now, the advantage of reformulating the problem in this way is that, contrary to p, the function g depends directly on the data that we have, and this enables us to estimate g by a simple nonparametric procedure that we shall describe shortly. We will also see that we can estimate μ_Z^* and q from our data, and from there we will derive a nonparametric estimator of p.

First we show how to estimate g nonparametrically from the data $(X_{i,j}, Y_j^*)$, $i = 1,\ldots,n_j$, $j = 1,\ldots,J$, using a local polynomial technique. A priori, the task

seems complex because our target function is an average of several regression curves. Moreover, our sample is of a rather unusual type since pooled data are not independent and not identically distributed, and, in each group, the X observations share the same Y^*. Despite these difficulties, we prove in the next section that the function g is estimated consistently by the following simple modification of the standard formula for ℓth order local polynomial estimators:

$$\widehat{g}(x) \;=\; \mathbf{e}_0^T \widehat{\mathbf{S}}^{-1} \widehat{\mathbf{T}}\,, \tag{2.3}$$

where $\widehat{\mathbf{S}} = \big(\widehat{S}_{k,k'}\big)_{0\le k,k'\le \ell}$ and $\widehat{\mathbf{T}} = (\widehat{T}_0, \ldots, \widehat{T}_\ell)^T$, with,

$$\widehat{S}_{k,k'} = \frac{1}{Nh^{k+k'}} \sum_{j=1}^{J} \sum_{r=1}^{n_j} K_h(x - X_{r,j})(x - X_{r,j})^{k+k'}\,,$$

$$\text{and} \quad \widehat{T}_k = \frac{1}{Nh^{k}} \sum_{j=1}^{J} Z_j^* \sum_{r=1}^{n_j} K_h(x - X_{r,j})(x - X_{r,j})^{k}\,,$$

for $k, k' = 0, \ldots, \ell$. To calculate this estimator in practice, we need to select the bandwidth h and the kernel K. As usual in nonparametric techniques, the latter does not matter much; essentially, we can take K equal to any smooth and symmetric density. We will see the exact conditions imposed on the choice of K in the next section. By contrast, the success of the estimator depends crucially on the value of h, which has to be chosen with a lot of care. We describe how to do that in Section 4.1. Finally, to calculate the estimator, we also need to choose the order ℓ of the polynomial. Although, it could in principle be any non negative integer, the most commonly used values are $\ell = 0$ or 1; see the discussion at the end of Section 3.

It remains to estimate the parameters μ_Z^* and q. For μ_Z^*, we can simply use the unbiased estimator

$$\widehat{\mu}_Z^* \;=\; N^{-1} \sum_{j=1}^{J} n_j Z_j^*\,. \tag{2.4}$$

Estimating q seems more difficult since it depends on the unknown p. We suggest using the maximum likelihood estimator. For this, first note that the (unobserved)

$Y_{i,j}$'s are Bernoulli$(1-q)$. From there, we can deduce that the likelihood function for the observed data, that is, the joint distribution of the independent non-identically distributed variables Z_j^*, $j = 1, \ldots, J$, is given by

$$L(z_1, \ldots, z_J; q) \;=\; \prod_{j=1}^{J} \left\{ z_j q^{n_j} + (1 - z_j)(1 - q^{n_j}) \right\}, \quad z_1, \ldots, z_J \in \{0, 1\}.$$

The maximum likelihood estimator for q is obtained by solving the equation $\partial \log L / \partial q = 0$ for q, and in Section A.1 of the supplemental file, we show that this is equivalent to solving $\Phi(z_1, \ldots, z_J; q) = 0$, where

$$\Phi(z_1, \ldots, z_J; q) \;=\; \sum_{j=1}^{J} n_j (z_j - q^{n_j}) \Big/ \Big(\sum_{k=0}^{n_j - 1} q^k \Big).$$

Here we use the notation $\sum_{k=0}^{n_j-1} q^k$ instead of $(1-q^{n_j})/(1-q)$ as the latter is numerically more unstable in cases where q is close to 1, which arises frequently in disease prevalence applications. In Section A.1 of the supplemental file, we show that, for any $z_1, \ldots, z_J \in \{0, 1\}$, $\Phi(z_1, \ldots, z_J; \cdot)$ has exactly one zero in the interval $[0, 1]$. Therefore we can define our estimator $\widehat{q}$ of q to be the unique zero of the function $\Phi(Z_1^*, \ldots, Z_J^*, \cdot)$ on the domain $[0, 1]$.

Finally, combining the estimators of g, μ_Z^* and q we just discussed, we define our nonparametric estimator of p by

$$\widehat{p}(x) \;=\; 1 - \widehat{q} \cdot \widehat{g}(x) / \widehat{\mu}_Z^*. \tag{2.5}$$

3 Asymptotic properties

In this section we investigate the theoretical asymptotic behaviour of our estimator. In settings where more conventional samples are observed, this is usually done by examining properties of the estimator as the sample size increases. In our context, it is natural to derive the asymptotic behaviour of the estimator when the number of groups,

J, tends to infinity, while the group sizes n_j remain bounded. Indeed, in applications, group sizes are often small compared to J; see for example Xie (2001), where $J = 9575$ and $n_j \equiv 10$. Moreover, the World Health Organization does not recommend using $n_j > 6$, although larger values of n_j could be reasonable in cases of low prevalence; see e.g. `http://www.gsb.stanford.edu/news/research/healthcare_donors.shtml`. To derive asymptotic properties of our local polynomial estimator of order ℓ at a point x, we impose the following assumptions:

Condition A

(A1) $q \geq q_0 > 0$ and $\sup_{j\in\mathbb{N}} n_j < \infty$.

(A2) The density f_X of the $X_{i,j}$'s is bounded and continuous in a neighbourhood of x, and is such that $f_X(x) > 0$. Moreover, if ℓ is even, f_X is continuously differentiable in that neighbourhood.

(A3) p has β continuous derivatives in a neighbourhood of x and $\sup_{k=0,\ldots,\beta} |p^{(k)}(y)| \leq B$ for all y located inside that neighbourhood, where B is a finite constant and where $\beta \geq \ell + 1$ if ℓ is odd and $\beta \geq \ell + 2$ if ℓ is even.

(A4) K is a symmetric, bounded and continuous density and $\int y^{2\ell+4} K(y)\, dy < \infty$.

(A5) $h \to 0$ and $Nh \to \infty$ as $N \to \infty$.

The first part of (A1) is natural in our applications as $1 - q$ denotes the probability that an individual in a population suffers from a disease (or that water samples are contaminated by a pollutant), and in general the latter is significantly smaller than 1. Conditions (A2) to (A5) are standard in nonparametric regression. Intuitively (A2) ensures that we have enough observations in the neighborhood of x for a local estimator to make sense. Similarly, we need p to be smooth enough to be able to fit locally around x a polynomial of order ℓ, whence condition (A3). Conditions (A4) and (A5) are easy to satisfy since we can choose K and h.

Let $\mu_j = \int u^j K(u)\, du$, $\nu_j = \int u^j K^2(u)\, du$, $\boldsymbol{S} = \left(\boldsymbol{S}_{j,k}\right)_{0\leq j,k\leq \ell}$, $\boldsymbol{S}^* = \left(\boldsymbol{S}^*_{j,k}\right)_{0\leq j,k\leq \ell}$

and $\widetilde{\boldsymbol{S}} = \big(\widetilde{\boldsymbol{S}}_{j,k}\big)_{0\le j,k\le \ell}$, where $\boldsymbol{S}_{j,k} = \mu_{j+k}$, $\boldsymbol{S}^*_{j,k} = \nu_{j+k}$ and $\widetilde{\boldsymbol{S}}_{j,k} = \mu_{j+k+1}$, and let $\boldsymbol{\mu} = (\mu_{\ell+1}, \ldots, \mu_{2\ell+1})^T$ and $\widetilde{\boldsymbol{\mu}} = (\mu_{\ell+2}, \ldots, \mu_{2\ell+2})^T$. The next theorem describes asymptotic properties of our estimator $\widehat{p}(x)$.

Theorem 3.1. *Under condition A and if $N \to \infty$, the estimator $\widehat{p}$ at (2.5) satisfies*

$$\big|\widehat{p}(x) - p(x)\big|^2 \;=\; \mathrm{ASE}_P(x) \;+ O_P\{(Nh)^{-3/2}\} + O_P(N^{-1}h^{-1/2})\,,$$

where $\mathrm{ASE}_P(x)$ is a nonnegative random variable such that

$$\mathrm{E}\{\mathrm{ASE}_P(x)\} \;=\; N^2M^{-2}\{V(x) + B^2(x)\}\cdot\{1 + o(1)\}\,,$$

with $V(x) = (Nh)^{-1}\boldsymbol{e}_0^T\boldsymbol{S}^{-1}\boldsymbol{S}^*\boldsymbol{S}^{-1}\boldsymbol{e}_0\{g(x) - g^2(x)\}/f_X(x)$, and

$$\begin{aligned}
B(x) =& \boldsymbol{e}_0^T\boldsymbol{S}^{-1}\boldsymbol{\mu}\,\frac{1}{(\ell+1)!}g^{(\ell+1)}(x)h^{\ell+1} && \text{if } \ell \text{ is odd}\,;\\
B(x) =& \boldsymbol{e}_0^T\boldsymbol{S}^{-1}\widetilde{\boldsymbol{\mu}}\frac{1}{(\ell+2)!}\Big[g^{(\ell+2)}(x) + (\ell+2)g^{(\ell+1)}(x)\frac{f_X'(x)}{f_X(x)}\Big]h^{\ell+2}\\
&- \boldsymbol{e}_0^T\boldsymbol{S}^{-1}\widetilde{\boldsymbol{S}}\boldsymbol{S}^{-1}\boldsymbol{\mu}\frac{1}{(\ell+1)!}g^{(\ell+1)}(x)\frac{f_X'(x)}{f_X(x)}h^{\ell+2} && \text{if } \ell \text{ is even.}
\end{aligned}$$

This theorem shows that, asymptotically the squared error $\big|\widehat{p}(x) - p(x)\big|^2$ equals $\mathrm{ASE}_P(x)$, the "asymptotic squared error", plus a term that can asymptotically be neglected in probability, and the mean of $\mathrm{ASE}_P(x)$ is of larger order than the remainder O_P terms. Note that, in general, the dominating part of $E\{\mathrm{ASE}_P(x)\}$ is not equal to the asymptotic mean squared error of $\widehat{p}$. Indeed, for this to be true, we need to ensure that the mean of the O_P terms exists and is of lower order than $E\{\mathrm{ASE}_P(x)\}$, which requires an extra regularization parameter, for example a ridge. The arguments for proving this are quite complex but can be derived using calculations similar to those used in, for example, Delaigle and Meister (2011). However, it is not very difficult to prove, using techniques similar to Delaigle et al. (2009), that, even without this extra parameter, B (for bias) and V (for variance) are such that $g + B$ and V are the mean and variance of the asymptotic distribution of $\widehat{g}$. Later,

abusing notations a little, we refer to $V(x) + B^2(x)$ as the asymptotic mean squared error of $\widehat{g}$.

As usual for local polynomial estimators, the expression of the bias is different for ℓ even and ℓ odd. It is easy to prove that the optimal bandwidth, that is, the bandwidth producing the smallest asymptotic "mean squared error", is of order $h \asymp N^{-1/(2\ell+3)}$ if ℓ is odd and $h \asymp N^{-1/(2\ell+5)}$ if ℓ is even, and that, with such bandwidths, $|\widehat{p}(x) - p(x)|^2 = O_P\big(N^{-(2\ell+2)/(2\ell+3)}\big)$ if ℓ is odd and $O_P\big(N^{-(2\ell+4)/(2\ell+5)}\big)$ if ℓ is even. Although this implies that the higher ℓ the faster the convergence rates, high values of ℓ can only be used when p has enough derivatives. Since the smoothness of a target curve is usually unknown in practice, in standard nonparametric estimation problems it is common to use local polynomial estimators of order $\ell = 0$ or 1. It is also well known that, in finite sample size, taking ℓ higher often does not improve the results. This is because, even though the order of the variance V does not depend on ℓ, the constant term of V (i.e. the term that does not depend on N) increases with ℓ. Although not discussed here in details, it is also well known that the local linear estimator ($\ell = 1$) behaves better than the local constant estimator ($\ell = 0$) in cases where the target curve (here p) is not continuous at the end of its support (see Fan and Gijbels, 1996), and for this reason, the local linear estimator is usually preferred to the local constant one.

The asymptotic "bias" of our estimator $\widehat{p}(x)$, that is, the bias component of $\mathrm{ASE}_P(x)$, is equal to $-(N/M)B(x)$. Using the fact that, for all integers $k > 0$, $g^{(k)}(x) = -(M/N)p^{(k)}(x)$, it is straightforward to see that the asymptotic "bias" of $\widehat{p}(x)$ can be written as $B(x)$, with g replaced by p. We conclude that the bias of our estimator is the same no matter what the pool sizes are. In particular, since, when $n_j = 1$, our estimator reduces to the standard regression estimator of p, the bias of the estimator for grouped data is the same as the bias of the standard regression estimator constructed from non-pooled data. The asymptotic "variance" of $\widehat{p}(x)$, that

is, the variance component of $\mathrm{ASE}_P(x)$, is equal to $(N/M)^2 V(x)$. Simple calculations show that this variance depends on the group sizes n_j. In particular, the variance of our estimator for grouped data differs from that of the standard regression estimator; the variance of the latter is obtained by letting $n_j = 1$. Nevertheless, the two variances are of the same order $1/(Nh)$, and it is only the multiplicative constant factor that is larger for our estimator than for the standard regression estimator. It follows from this discussion that the two estimators share the same rates of convergence, and, under condition A, grouping the data does not lead to any deterioration of the standard nonparametric rates.

4 Numerical properties

As discussed earlier, in standard regression problems it is common to use the local linear estimator ($\ell = 1$), and for the same reasons as there, in our context too, our preference is to use $\ell = 1$. Therefore, we only investigate numerical properties of our local linear estimator.

4.1 Bandwidth selection

Bandwidth selection in the standard regression setting has been studied by many authors. The methods we suggest here are similar to those suggested by Ruppert et al. (1995). See also Fan and Gijbels (1996) and Simonoff (1996). Note that our problem is more difficult because the data are grouped and because we have to deal with heteroscedastic regression, which makes our variance term more complicated to estimate. To choose h, we suggest minimising with respect to h the weighted "asymptotic mean integrated squared" expression $\mathrm{AMISE} = \int \{V(x)+B^2(x)\}w(x)\,dx$, where w is a weight function and B and V are given in Theorem 3.1; remember that B and V are the bias and variance resulting from the asymptotic distribution of $\widehat{g}$. To put

less emphasize on areas where we have few observations, we take $w(x) = f_X(x)$, as in more standard regression problems. In the local linear case ($\ell = 1$), this gives

$$\text{AMISE} = b\,\mu_2^2\,\frac{h^4}{4} + v\,\frac{R(K)}{Nh}\,,$$

where $R(K) = \int K^2$, $b = \int \{g''(x)\}^2 f_X(x)\,dx$ and $v = \int g(x)\{1 - g(x)\}\,dx$. The theoretical bandwidth obtained by minimising this expression w.r.t. h is given by

$$h^* = \left\{R(K)v/(\mu_2^2\,b)\right\}^{1/5} N^{-1/5},$$

but it cannot be calculated in practice since it depends on b and v, which themselves depend on g, g'' and f_X, which are unknown. Instead, we can only use

$$\widehat{h}^* = \left\{R(K)\widehat{v}/(\mu_2^2\,\widehat{b})\right\}^{1/5} N^{-1/5},$$

where $\widehat{b}$ and $\widehat{v}$ are estimators of b and v. In what follows, we construct such estimators.

4.1.1 Estimation of v

We start by deriving a consistent nonparametric estimator of v. For $j = 1, \ldots, J$, let $T_j^* = \widehat{\mu}_Z^* \widehat{q}^{-n_j} Z_j^*$. If the groups all have the same size $n_j = n_1$, we estimate v by

$$\widehat{v} = \frac{1}{n_1} \sum_{i=1}^{n_1} \sum_{j=1}^{J-1} T_{i,[j]}^* \big(1 - T_{i,[j+1]}^*\big)\big(X_{i,(j+1)} - X_{i,(j)}\big), \tag{4.1}$$

where, for each i fixed, $X_{i,(1)} < \cdots < X_{i,(J)}$ are the order statistics of $X_{i,1} \ldots, X_{i,J}$ and $T_{i,[j]}^*$ denotes the response variable T_k^* which corresponds to $X_{i,(j)}$.

If the n_j's are not all equal, we proceed as follows. For $i = 1, \ldots, \max_j n_j$, let J_i denote the number of groups of size larger or equal to i. To simplify the notation, suppose, without loss of generality, that the groups are indexed in such a way that, for all $j < k$, we have $n_j \geq n_k$. Then, for each fixed i, calculate the order statistics $X_{i,(1)} < \cdots < X_{i,(J_i)}$ of $X_{i,1} \ldots, X_{i,J_i}$, and let $T_{i,[j]}^*$ denote the response variable

T_k^* which corresponds to $X_{i,(j)}$. For $i = 1, \ldots, \max_j n_j$, let $\widehat{v}_i = \sum_{j=1}^{J_i-1} T_{i,[j]}^* \big(1 - T_{i,[j+1]}^*\big)\big(X_{i,(j+1)} - X_{i,(j)}\big)$. We estimate v by

$$\widehat{v} = \sum_{i=1}^{\max_j n_j} w_i \widehat{v}_i, \tag{4.2}$$

where the w_is are weights summing to one, whose value depends on the group sizes. More precisely, since $\widehat{v}_i$ is less accurate when estimated from fewer observations, we take w_i large (resp., small) if J_i is large (resp., small). It can be proved that a choice that optimizes asymptotic order is

$$w_i = \sqrt{J_i} \Big/ \sum_{\ell=1}^{\max_j n_j} \sqrt{J_\ell}\,. \tag{4.3}$$

Note that if all group sizes are equal, (4.2) reduces to (4.1). In section B.1 of the supplemental file, we show that $\widehat{v}_i$ is a consistent estimator of v for each i such that $J_i \to \infty$ as $N \to \infty$. Since, in addition, for each i such that $J_i/J \to 0$ as $N \to \infty$, we have $w_i \to 0$, the estimator $\widehat{v}$ is consistent.

4.1.2 Parametric estimation of b

As we have just seen, we can estimate the variance term v nonparametrically, and without using any additional bandwidth parameter. As in standard regression problems, estimating the bias term b nonparametrically is much more complex; we will show how to do that in Section 4.1.3. However, a reasonable bandwidth can often be obtained by estimating b with a simple parametric method. Even though the resulting estimator of b is usually not consistent (since we do not know the right parametric model), it produces a bandwidth of the right asymptotic order, which in turns provides a consistent estimator for g. In the nonparametric terminology, a bandwidth based on such parametric estimation of the bias is called a rule of thumb (ROT).

We can estimate b parametrically by $\widetilde{b} = N^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \{\widetilde{g}''(X_{i,j})\}^2$, where $\widetilde{g}''$ denotes the second derivative of a parametric estimator $\widetilde{g}$ of g. For bandwidth selection

purposes, to estimate a derivative, $g^{(k)}$ say, of a regression curve g parametrically, it is quite standard to approximate g by a polynomial

$$\widetilde{g}(x) = \sum_{j=0}^{\nu} \beta_j \, x^j, \tag{4.4}$$

where the parameters β_j are estimated from the data and where $\nu \geq k$. To make the distinction with the local polynomial procedure, we call this the "global polynomial procedure".

We estimate the parameter $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_\nu)^T$ from the data as follows. Let $\mathcal{T} = (T_1^*, \ldots, T_1^*, \ldots, T_J^*, \ldots, T_J^*)^T$, with T_j^* as in Section 4.1.1 and where each T_j^* is repeated n_j times. Then, let $\mathcal{X}$ be the $N \times (\nu + 1)$ matrix with rows equal to $\big(1, X_{i,j}, \ldots, X_{i,j}^\nu\big)$, for $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$, and where the jth group of n_j rows corresponds to $i = 1 \ldots, n_j$. We estimate $\boldsymbol{\beta}$ by least squares, that is we take

$$\widehat{\boldsymbol{\beta}} = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathcal{T}. \tag{4.5}$$

See section B.2 of the supplemental file for properties of this estimator. To estimate g'', we take $\nu = 3$, which gives $\widetilde{g}''(x) = 2\widehat{\beta}_2 + 6\widehat{\beta}_3 \, x$.

4.1.3 Nonparametric estimation of b

Next, we show how to estimate b nonparametrically. Here, unlike $\widetilde{b}$, the resulting estimator $\widehat{b}$ is consistent, but the procedure is more complex. One of the main difficulties is that the estimator depends on a nonparametric estimator $\widehat{g}''$ of g'', which itself requires a smoothing parameter. In other words, with this method, in order to calculate the bandwidth $\widehat{h}^*$ for estimating p, we need to choose a second bandwidth h_2 for estimating b. In the literature, the resulting bandwidth $\widehat{h}^*$ is called a plug-in bandwidth (PI).

There are various ways to estimate g'' nonparametrically, and we take an approach for which it is relatively easy to derive a data-driven bandwidth selector. Let $h_2 > 0$

and, for $i = 1, \ldots, \max_j n_j$, let $\widetilde{\mathbf{S}}_i = \big(\widetilde{S}_{i,k,k'}\big)_{0 \le k,k' \le \ell}$ and $\widetilde{\mathbf{T}}_i = (\widetilde{T}_{i0}, \ldots, \widetilde{T}_{i\ell})^T$, where

$$\widetilde{S}_{i,k,k'} = \frac{1}{J_i h_2^{k+k'}} \sum_{j=1}^{J_i} K_{h_2}(x - X_{i,j})(x - X_{i,j})^{k+k'} ,$$

$$\text{and} \quad \widetilde{T}_{i,k} = \frac{1}{J_i h_2^{k}} \sum_{j=1}^{J_i} T_j^* K_{h_2}(x - X_{i,j})(x - X_{i,j})^{k} ,$$

with T_j^* and J_i as in Section 4.1.1. Using arguments similar to these used in standard nonparametric regression problems, it is possible to show that for each i, a consistent ℓth order local polynomial estimator of g'' can be defined by $\widehat{g}_i''(x) = 2h_2^{-2}\, \mathbf{e}_2^T \widetilde{\mathbf{S}}_i^{-1} \widetilde{\mathbf{T}}_i$, where $\ell \ge 2$ and $\mathbf{e}_2$ is the vector of length $\ell + 1$ defined by $\mathbf{e}_2 = (0, 0, 1, 0 \ldots, 0)^T$. We suggest estimating b by

$$\widehat{b} = \sum_{i=1}^{\max_j n_j} \frac{w_i}{J_i} \sum_{j=1}^{J_i} \{\widehat{g}_i''(X_{i,j})\}^2 ,$$

with the weights w_i as defined in (4.3). Note that, when the groups all have the same size $n_j = n_1$, this estimator reduces to $n_1^{-1} \sum_{i=1}^{n_1} J^{-1} \sum_{j=1}^{J} \{\widehat{g}_i''(X_{i,j})\}^2$.

For reasons similar to those discussed in Section 3, when estimating a derivative, $g^{(k)}$ say, in practice it is standard to choose $\ell = k + 1$. Therefore, to estimate g'' we take $\ell = 3$. Of course, this estimator depends on a bandwidth h_2, which should be chosen for $\widehat{b}$ to be a good estimator of b. In section B.3 of the supplemental file, using arguments of Ruppert et al. (1995), we suggest taking this second bandwidth equal to an estimator of

$$h_2^* = C_2(K)\big(v/|\theta_{24}|\big)^{1/7}\Big(\sum_{i=1}^{\max_j n_j} w_i/J_i \Big)^{1/7},$$

where $C_2(K) = (\alpha_K \mathbf{e}_2^T \mathbf{S}^{-1} \mathbf{S}^* \mathbf{S}^{-1} \mathbf{e}_2 / \mathbf{e}_2^T \mathbf{S}^{-1} \boldsymbol{\mu})^{1/7}$, with $\alpha_K = 60$ if $\theta_{24} > 0$ and $\alpha_K = 24$ if $\theta_{24} < 0$. For example, for a Gaussian kernel, $C_2(K) = \{3/(8\sqrt{\pi})\}^{1/7}$ if $\theta_{24} < 0$ and $C_2(K) = \{15/(16\sqrt{\pi})\}^{1/7}$ if $\theta_{24} > 0$.

We estimate the bandwidth h_2^* by replacing v by the estimator of Section 4.1.1

and θ_{24} by

$$\widehat{\theta}_{24} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\widetilde{g}''(X_{i,j})\widetilde{g}^{(4)}(X_{i,j}),$$

where $\widetilde{g}''$ is the global polynomial estimator discussed in Section 4.1.2, and $\widetilde{g}^{(4)}$ is the fourth derivative of a fourth order global polynomial estimator of g. That is, we take $\widetilde{g}$ as in (4.4), with $\nu = 4$, and then we put $\widetilde{g}^{(4)}(x) = 24\widehat{\beta}_4$. In general $\widehat{\theta}_{24}$ is not a consistent estimator of θ_{24}. However, at this pilot stage, we are sufficiently far from the original problem of estimating p, so that replacing unknown quantities by rough approximations does not seriously impact the quality of the estimator of p.

4.1.4 Weighted versions

In practice, estimating the derivatives of a regression curve is not easy. Estimators of the derivatives, especially if they are nonparametric, can be too variable near the boundary. In standard regression problems, to overcome this issue, it is common to replace estimators of quantities such as b and θ_{24} by weighted versions; see for example Gasser et al. (1991). In our case, such weighted estimators can be defined by $\widehat{b} = N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\{\widehat{g}''(X_{i,j})\}^2\omega(X_{i,j})$, $\widetilde{b} = N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\{\widetilde{g}''(X_{i,j})\}^2\omega(X_{i,j})$ and $\widehat{\theta}_{24} = N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\widetilde{g}''(X_{i,j})\widetilde{g}^{(4)}(X_{i,j})\omega(X_{i,j})$, where ω is a weight function.

For the PI method, we considered two weights, namely $\omega = \omega_0$ and $\omega = \omega_1$, where $\omega_0(x) = 1\{q_{0.1}, q_{0.9}\}(x)$ and $\omega_1(x) = 1\{q_{0.2}, q_{0.8}\}(x)$, with q_α denoting the α quantiles of the X_{ij}'s. We denote by PI_{ω_0} and PI_{ω_1} the two corresponding weighted PI bandwidth selectors. For the ROT, using a weight to estimate b is less crucial, and we considered the unweighted ROT and the ROT that uses the weight $\omega = \omega_0$. We denote these two versions of the ROT bandwidth selector by ROT and ROT_{ω_0}.

Table 1: Simulation results for models (i) to (iv), when the $X_{i,j}$'s are uniform. The numbers show $10^4\times$ MISE (stdev) calculated from 200 simulated samples.

		$n_j = 1$	$n_j = 2$		$n_j = 5$		$n_j = 10$	
Model	N	MW	ROT	ROT_{ω_0}	ROT	ROT_{ω_0}	ROT	ROT_{ω_0}
(i)	5.10^3	4.58(2.57)	6.22(3.44)	6.42(3.41)	12.6(6.93)	12.3(6.49)	22.5(13.1)	20.9(11.7)
	10^4	2.47(1.34)	3.80(1.97)	4.06(2.08)	7.21(3.90)	7.53(3.90)	12.6(6.87)	12.6(6.45)
(ii)	5.10^3	3.52(2.83)	4.98(3.51)	4.44(3.34)	10.4(7.44)	9.22(7.03)	20.4(15.3)	18.0(14.0)
	10^4	2.08(1.66)	2.95(1.83)	2.58(1.72)	5.16(3.29)	4.44(3.05)	9.33(6.68)	7.89(5.97)
(iii)	5.10^3	.557(.559)	.561(.461)	.595(.438)	.921(.827)	.932(.814)	1.72(1.66)	1.67(1.59)
	10^4	.321(.312)	.359(.221)	.402(.213)	.521(.376)	.554(.364)	.781(.631)	.791(.610)
(iv)	5.10^3	1.59(1.11)	2.10(1.21)	2.21(1.22)	3.94(2.55)	4.12(2.55)	6.82(4.00)	7.06(4.06)
	10^4	.858(.650)	1.30(.659)	1.50(.674)	2.31(1.20)	2.67(1.28)	4.19(2.39)	4.74(2.54)
Model	N	MW	PI_{ω_0}	PI_{ω_1}	PI_{ω_0}	PI_{ω_1}	PI_{ω_0}	PI_{ω_1}
(i)	5.10^3	4.58(2.57)	6.32(3.33)	6.78(3.39)	12.0(6.70)	12.0(5.86)	22.5(13.7)	20.2(12.0)
	10^4	2.47(1.34)	3.96(1.95)	4.22(2.04)	7.02(3.70)	7.63(3.65)	12.4(6.80)	11.9(6.04)
(ii)	5.10^3	3.52(2.83)	4.97(3.56)	4.38(3.30)	11.7(7.82)	9.72(7.23)	25.8(16.7)	21.1(14.8)
	10^4	2.08(1.66)	2.77(1.80)	2.50(1.67)	5.57(3.54)	4.55(3.10)	11.5(7.33)	9.05(6.28)
(iii)	5.10^3	.557(.559)	.632(.636)	.591(.610)	1.47(1.17)	1.33(1.11)	3.37(2.38)	3.04(2.27)
	10^4	.321(.312)	.329(.278)	.308(.265)	.694(.516)	.625(.491)	1.46(1.07)	1.28(1.00)
(iv)	5.10^3	1.59(1.11)	1.91(1.31)	1.84(1.24)	3.82(2.84)	3.60(2.73)	7.61(4.19)	6.90(3.95)
	10^4	.858(.650)	1.14(.736)	1.11(.700)	2.04(1.24)	1.96(1.19)	4.13(2.45)	3.80(2.35)

4.2 Simulations

To study the practical performance of the estimator $\widehat{p}$ in (2.5), we applied it to samples generated from the following models:

(i) $p(x) = \{\sin(\pi x/2) + 1.2\}/[20 + 40x^2\{\text{sign}(x) + 1\}]$ and $X \sim U[-3, 3]$ or $X \sim N(0, 1.5^2)$;

(ii) $p(x) = \exp(-4 + 2x)/\{8 + 8\exp(-4 + 2x)\}$ and $X \sim U[-1, 4]$ or $X \sim N(2, 1.5^2)$;

(iii) $p(x) = x^2/8$ and $X \sim U[0, 1]$ or $X \sim N(0.5, 0.5^2)$;

(iv) $p(x) = x^2/8$ and $X \sim U[-1, 1]$ or $X \sim N(0, 0.75^2)$.

We chose these curves so as to get examples with various features. We considered several values of N and n_j. For each combination of p, N, n_j and distribution of X, we generated 200 random samples of size $N = \sum_{j=1}^{J} n_j$ from (X, Y), where

Table 2: Simulation results for models (i) to (iv), when the $X_{i,j}$'s are normal. The numbers show $10^4\times$ MISE (stdev) calculated from 200 simulated samples.

		$n_j = 1$	$n_j = 2$		$n_j = 5$		$n_j = 10$	
Model	N	MW	ROT	ROT_{ω_0}	ROT	ROT_{ω_0}	ROT	ROT_{ω_0}
(i)	5.10^3	5.69(4.12)	13.4(6.24)	17.6(6.44)	17.2(8.98)	21.0(8.60)	25.6(14.8)	28.3(14.1)
	10^4	3.38(2.08)	9.67(4.41)	14.5(5.29)	13.3(5.60)	17.7(6.05)	16.6(8.19)	20.4(8.37)
(ii)	5.10^3	2.52(2.26)	3.42(2.36)	3.41(2.26)	7.5(8.00)	6.61(7.23)	16.4(20.6)	13.8(18.1)
	10^4	1.31(.955)	2.18(1.56)	2.41(1.65)	4.77(4.19)	4.51(3.83)	10.3(14.9)	8.84(13.1)
(iii)	5.10^3	.368(.331)	1.33(.763)	1.59(.833)	1.93(1.04)	2.21(1.10)	2.78(1.62)	3.07(1.66)
	10^4	.202(.191)	.851(.400)	1.06(.449)	1.26(.650)	1.50(.698)	1.83(.963)	2.11(.999)
(iv)	5.10^3	1.23(.768)	2.86(1.54)	3.07(1.60)	5.36(2.79)	5.61(2.82)	9.84(6.45)	10.1(6.38)
	10^4	.693(.445)	1.82(.955)	2.23(1.06)	3.27(1.75)	3.78(1.84)	5.78(3.06)	6.37(3.13)
Model	N	MW	PI_{ω_0}	PI_{ω_1}	PI_{ω_0}	PI_{ω_1}	PI_{ω_0}	PI_{ω_1}
(i)	5.10^3	5.69(4.12)	9.88(4.64)	11.3(4.92)	14.7(8.97)	16.0(8.80)	23.2(15.9)	23.6(15.2)
	10^4	3.38(2.08)	6.87(3.00)	8.14(3.32)	10.6(4.53)	12.1(4.59)	13.5(7.40)	14.5(7.16)
(ii)	5.10^3	2.52(2.26)	3.33(2.31)	3.41(2.30)	7.14(7.79)	6.74(7.45)	17.7(21.3)	16.1(19.9)
	10^4	1.31(.955)	2.27(1.59)	2.44(1.62)	4.65(4.39)	4.57(4.17)	10.5(15.7)	9.80(15.2)
(iii)	5.10^3	.368(.331)	.663(.509)	.701(.528)	1.18(.890)	1.21(.885)	2.24(1.83)	2.16(1.71)
	10^4	.202(.191)	.335(.228)	.360(.239)	.638(.456)	.665(.471)	1.18(.913)	1.18(.890)
(iv)	5.10^3	1.23(.768)	2.30(1.40)	2.41(1.41)	4.74(2.79)	4.83(2.75)	9.24(6.92)	9.18(6.70)
	10^4	.693(.445)	1.38(.822)	1.48(.847)	2.74(1.67)	2.84(1.67)	5.27(3.10)	5.30(3.06)

$Y|X \sim \text{Bernoulli}\{p(X)\}$. Then we divided the data randomly into J groups, each with n_j observations, to produce 200 samples of size J from $(X_{1j}, \ldots, X_{n_j,j}, Y_j^*)$, following (2.1), and calculated the 200 corresponding estimators $\widehat{p}$ of p. For each, we then calculated the integrated squared error $\text{ISE} = \int_a^b \{\widehat{p}(x) - p(x)\}^2\, dx$, where $[a, b]$ is the x-range of the figures. We then ordered the 200 estimators from the best to the worst according to their ISE values. In the figures, we show the three estimators corresponding to the first, second and third quartile values of these 200 ISEs. The true curve p is shown in uninterrupted line. Throughout, we used the two versions of the ROT and of the PI bandwidth selectors described in Section 4.1.4. We took the kernel K equal to a standard normal density. Since p is a probability, in each case we truncated our estimator $\widehat{p}$ to $[0, 1]$; this does not change asymptotic properties of our estimator.

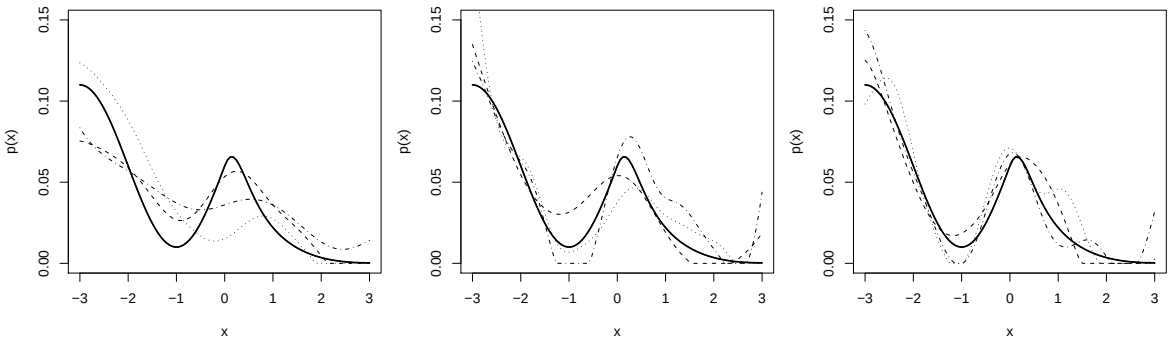

Figure 1: 1st quartile ($- - -$), 2nd quartile ($-\cdot-\cdot-$) and 3rd quartile ($\cdots$) curves for model (i) using ROT_{ω_0}, when $X \sim U[-3,3]$ and $n_j = 10$, with $N = 5000$ (left), $N = 10000$ (center) and $N = 20000$ (right).

Before we discuss the results, it is important to realise that the information contained in a sample of size N from our type of observations is not comparable to that contained in a sample of size N from (X, Y). The latter is considerably more informative as, for example, with our type of data we only observe J values of Y^*, and it is really J that plays the role of the traditional sample size. Moreover, the Y^*'s are only summary statistics of the unobserved Y's, and thus the problem we consider is quite complex. Finally, we should note that, to our knowledge, there does not exist in the literature any other nonparametric method for estimating p in the group testing context. Therefore we cannot compare our method with any other procedure since, as usual, it would not make sense to compare a nonparametric estimator with a parametric one; the latter can only be applied when the shape of p is known, whereas the former can be applied without such information on p.

We show the full simulation results for $N = 5000$ and 10000, with $n_j = 2$, 5 or 10 in Table 1, where the $X_{i,j}$'s are uniform, and in Table 2, where the $X_{i,j}$'s are normal. In each case, the numbers shown in the tables are 10^4 times the MISE and the standard deviation of the ISE, both calculated from 200 samples. To quantify the loss encountered by grouping the data, we also show the results for $n_j = 1$, i.e. the

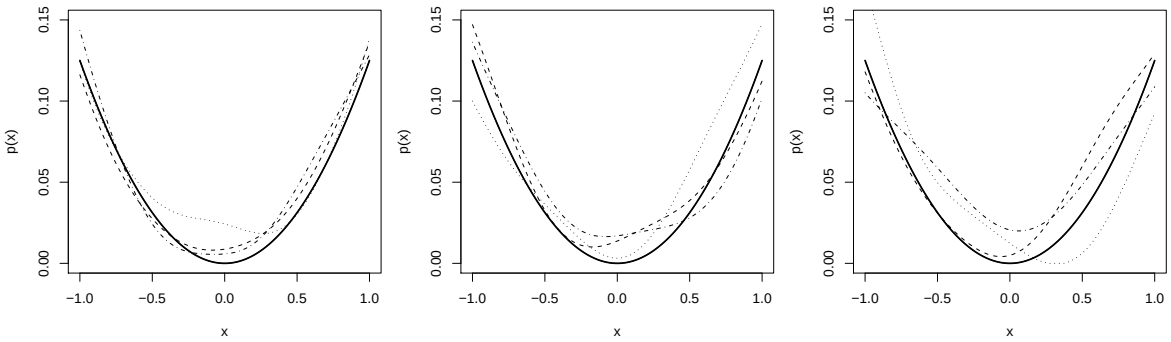

Figure 2: 1st quartile $(- - -)$, 2nd quartile $(- \cdot - \cdot -)$ and 3rd quartile $(\cdots)$ curves for model (iv) using PI_{ω_1}, when $X \sim N(0.5, 0.5^2)$ and $N = 5000$, with $n_j = 2$ (left), $n_j = 5$ (center) and $n_j = 10$ (right).

ideal case where the data are not grouped. There, we used the standard local linear estimator with Matt Wand's R `dpill` bandwidth, denoted by MW in the tables. See Wand and Ripley (2010). Overall, we see that when the $X_{i,j}$'s were uniform, the ROT gave the best results, and when the $X_{i,j}$'s were normal, it is the PI method that gave the best results. In general, our preference is to use the ROT method when the data are relatively uniformly distributed over their range, and to use the PI method otherwise. Our results also indicate that, overall, when the group sizes are small, the best results are obtained by the ROT and the PI_{ω_0}, and when the group sizes are larger, ROT_{ω_0} and PI_{ω_1} give the best results. This can be explained from the fact that the estimator is more variable when group sizes are large, and hence methods giving a larger bandwidth, such as these two, often work better in practice. Unsurprisingly, the performance of our estimator degraded as the group size increased and the results improved as the total sample size increased.

Next we depict these results by presenting the quartile estimated curves obtained in several cases. Additional simulation results for unequal group sizes are provided in Section C of the supplemental file. In Figure 1, we illustrate the effect of increasing the sample size by showing the results for curve (i) when $X \sim U[-3, 3]$ and $n_j = 10$

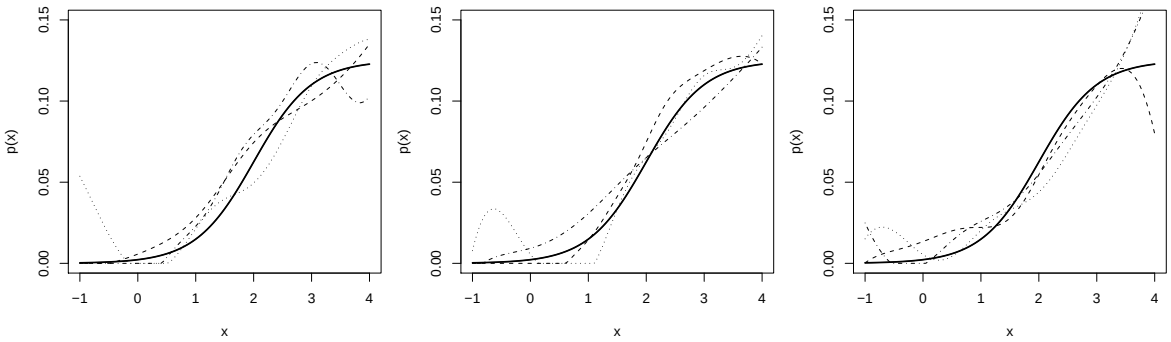

Figure 3: Row 1: 1st quartile $(- - -)$, 2nd quartile $(-\cdot-\cdot-)$ and 3rd quartile $(\cdots)$ curves for model (ii) using ROT_{ω_0}, when $X \sim U[-1, 4]$, with $n_j = 10$ and $N = 10000$ (left), $n_j = 10$ and $N = 20000$ (center) and $n_j = 20$ and $N = 20000$ (right).

with $N = 5000$, $N = 10000$, and $N = 20000$. Figure 2 shows the deterioration of the estimator when the group size increases; we report the results for curve (iv) when $X \sim N(0.5, 0.5^2)$ with $n_j = 2$, 5 or 10 and $N = 5000$. Finally, in Figure 3, we show that when the sample size is large, the estimator also works well with very large group sizes; we present the quartile curves for model (ii) with $X \sim U[-1, 4]$ and $n_j = 10$ or 20, when $N = 10000$ or $N = 20000$.

4.3 Application to the NHANES study

In the group testing literature, real data illustration is usually based on a sample for which the individual observations on (X, Y) are available, and groups are created artificially to compare the performance of group testing methods with the performance of the ideal estimator based on non grouped sample of individual observations, where $n_j = 1$. The illustration provided in this section is of this type. We applied our technique to data from the NHANES study, a very large nutrition and health study carried out in the US. We used data from 1999-2000, which are available at `www.cdc.gov/nchs/nhanes/nhanes1999-2000/nhanes99_00.htm`. These data were collected between 1999 and 2000. The survey design was a stratified, multistage prob-

ability sample of the civilian noninstitutionalized US population. For more details, see `www.cdc.gov/nchs/data/nhanes/nhanes_99_00/general_data_release_doc.pdf`. Our goal is to illustrate on these data the performance of our nonparametric estimator. We estimated two conditional probabilities:

(a) $p_{\mathrm{HBc}}(x) = P(Y_{\mathrm{HBc}} = 1|X = x)$, where X was the age of the patient and Y_{HBc} was a binary variable taking values 0 and 1 indicating, respectively, the absence or presence of antibody to hepatitis B virus core antigen in the patient's serum or plasma. Values of Y_{HBc} were obtained by the ORTHO HBc ELISA test system. After removing the individuals with missing values of X or Y_{HBc}, the sample size N was 7121 and the age X ranged from 1 month to 84 years;

(b) $p_{\mathrm{CL}}(x) = P(Y_{\mathrm{CL}} = 1|X = x)$, where X was the age of the patient and Y_{CL} was a binary variable taking values 0 and 1 indicating, respectively, the absence or presence of genital Chlamydia trachomatis infection in the urine of the patient. Note that C. trachomatis is responsible for many sexual transmittable diseases. Values of Y_{CL} were obtained by the LCx C. trachomatis assay. After removing the individuals with missing values of X or Y_{CL}, the sample size N was 2042 and the age X ranged from 12 to 40 years.

In both cases, we first calculated the local linear estimator of p based on the individual observations of (X, Y). We denote this estimator by $\widehat{p}_{\mathrm{ideal}}$; it is of much higher quality than estimators calculated from grouped data, but it cannot be calculated when the only available data are grouped. We take $\widehat{p}_{\mathrm{ideal}}$ as our reference curve and represent it by an uninterrupted curve in the figures. To assess the performance of our estimator on these data, we artificially created groups of sizes n_j from the original data, where n_j was equal to 2, 5 or 10. In each case we created 200 grouped samples in this way, and calculated our estimator for these grouped data. To assess the loss incurred by pooling the data, for each sample we calculated the integrated squared difference $\mathrm{ISD} = \int \{\widehat{p} - \widehat{p}_{\mathrm{ideal}}\}^2$; in the graphs, we show the three estimators that we

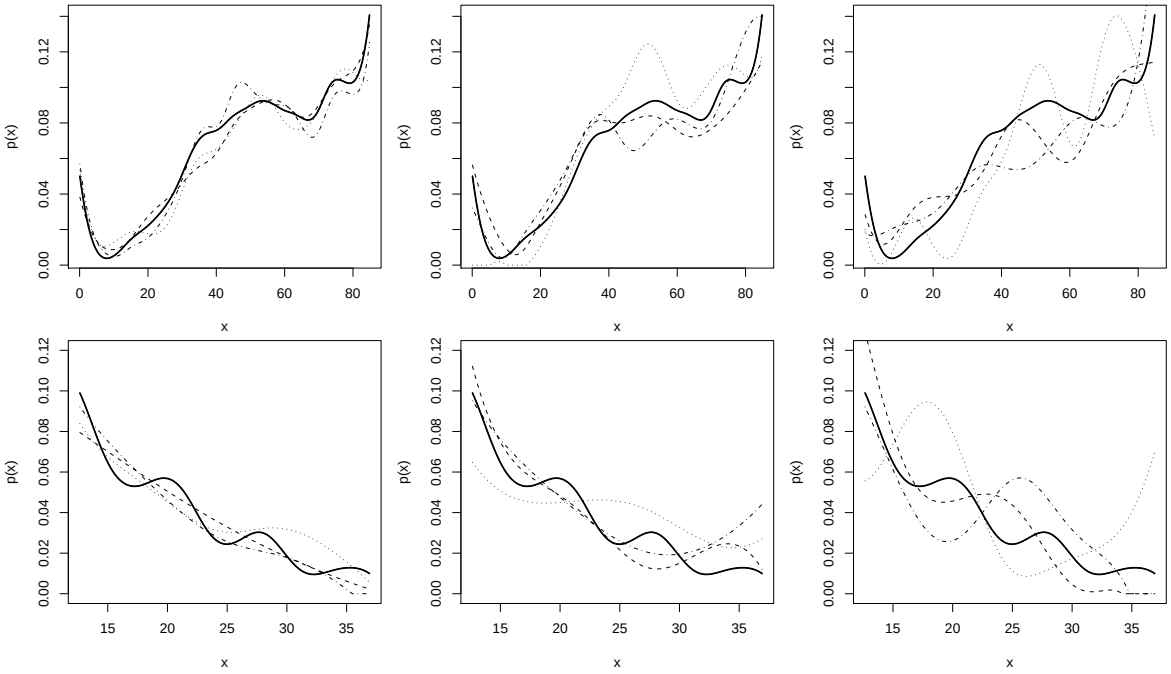

Figure 4: Nhanes study: 1st quartile $(- - -)$, 2nd quartile $(-\cdot-\cdot-)$ and 3rd quartile $(\cdots)$ curves, and ideal estimator $\widehat{p}_{\text{ideal}}$ (—) when $Y = Y_{\text{HBc}}$ (row 1) or $Y = Y_{\text{CL}}$ (row 2), and when $n_j = 2$ (left), $n_j = 5$ (center) or $n_j = 10$ (right), using the ROT_{ω_0}.

call quartile curves, and which correspond to the first, second and third quartiles of these 200 calculated ISDs. Since, in this case, the observations $X_{i,j}$ are relatively evenly widespread over their range, we used the ROT_{ω_0} bandwidth. The results are shown in Figure 4, where we see that our estimator worked reasonably well, and performed the best when estimating p_{HBc}. This is because the hepatitis sample was much larger than the Chlamydia sample. Unsurprisingly, and as already illustrated by our simulated examples, we can see that the quality of our estimator degrades as the group sizes n_j increase.

5 Test with imperfect accuracy

A detection test is not always perfectly accurate, and two types of errors can be encountered: the test on a group is positive when no one in the group is positive, and the test is negative when at least one member of the group is positive. If we let $D_{i,j}$ be the true status (0/1) of the ith individual from the jth group and D_j^* be the corresponding group status, then the probability of the first error is $p_1 \equiv P(Y_j^* = 1|D_j^* = 0)$, and the probability of the second error is $p_2 \equiv P(Y_j^* = 0|D_j^* = 1)$; note that $1 - p_1$ and $1 - p_2$ are usually referred to as, respectively, the specificity and sensitivity of the test. In that case, the function of interest is no longer $p(x) = P(Y = 1|X = x)$, but rather

$$\breve{p}(x) = P(D = 1|X = x) = 1 - \breve{q}\breve{g}(x)/\breve{\mu}_Z^*, \tag{5.1}$$

where D is the true status of an individual, $\breve{q} = E\{1-\breve{p}(X)\}$, $\breve{\mu}_Z^* = N^{-1}\sum_{j=1}^J n_j E\breve{Z}_j^*$, $\breve{Z}_j^* = 1 - D_j^*$ and $\breve{g}(x) = N^{-1}\sum_{j=1}^J \sum_{r=1}^{n_j} E\big(\breve{Z}_j^*|X_{r,j} = x\big)$.

As in Vansteelandt et al. (2000) and Xie (2001), we assume that p_1 and p_2 are known and that p_1 and p_2 are unaffected by the group size n_j. Assume that the result of the test depends only on the true status, and that p_1 and p_2 are less than 0.5; this would normally be the case since p_1 and p_2 are the probabilities of errors of the test. Then it is not very hard to prove that $P(D_j^* = 1|X_{i,j}) = (1 - p_2 - p_1)^{-1}P(Y_j^* = 1|X_{i,j}) - p_1(1 - p_2 - p_1)^{-1}$, that is,

$$\breve{g}(x) = (1 - p_2 - p_1)^{-1}g(x) - p_2(1 - p_2 - p_1)^{-1}. \tag{5.2}$$

See Vansteelandt et al. (2000) for an analogous formula in the parametric context. We deduce an estimator $\widehat{\breve{g}}(x)$ of $\breve{g}(x)$ by replacing the unknown g by $\widehat{g}$ of Section 2. To estimate $\breve{\mu}_Z^*$, note that $\breve{\mu}_Z^* = \big(\mu_Z^* - p_2\big)/(1 - p_2 - p_1)$, which we can estimate by $\widehat{\breve{\mu}}_Z^*$, obtained by replacing μ_Z^* in this equation by $\widehat{\mu}_Z^*$ defined in Section 2. We obtain an estimator $\widehat{\breve{q}}$ of $\breve{q}$ by solving for $\breve{q}$ the equation $\widehat{\breve{\mu}}_Z^* = \breve{M}/N\breve{q} = N^{-1}\sum_{j=1}^J n_j\breve{q}^{n_j}$, where $\breve{M} = \sum_{j=1}^J n_j\breve{q}^{n_j-1}$. Finally, we can estimate $\breve{p}$ by $\widehat{\breve{p}}(x) = 1 - \widehat{\breve{q}}\widehat{\breve{g}}(x)/\widehat{\breve{\mu}}_Z^*$.

Table 3: Perfect tests versus imperfect tests. The numbers show the mean (stdev) of the ratio ISE_I / ISE_P calculated from 200 simulated samples.

	$N = 5000$			$N = 10000$		
Case	$n_j = 2$	$n_j = 5$	$n_j = 10$	$n_j = 2$	$n_j = 5$	$n_j = 10$
Ex1	1.00 (.057)	.997 (.073)	.995 (.028)	.997 (.047)	.999 (.043)	.999 (.036)
Ex2	1.19 (.617)	1.16 (.544)	1.23 (.770)	1.19 (.407)	1.14 (.438)	1.31 (.873)
Ex3	1.43 (.857)	1.34 (.827)	1.41 (1.05)	1.26 (.672)	1.20 (.586)	1.35 (.755)
Ex4	2.35 (2.10)	1.71 (3.90)	1.53 (1.19)	2.03 (2.12)	1.57 (1.85)	1.38 (.946)

Asymptotic properties of $\widehat{\breve{p}}$ follow directly from Theorem 3.1. In particular, we have $\left|\widehat{\breve{p}}(x)-\breve{p}(x)\right|^2 = \mathrm{ASE}_P(x) + O_P\{(Nh)^{-3/2}\} + O_P(N^{-1}h^{-1/2})$, where $\mathrm{ASE}_P(x)$ is a nonnegative random variable such that $\mathrm{E}\{\mathrm{ASE}_P(x)\} = N^2\breve{M}^{-2}(1-p_2-p_1)^{-2}\{V(x)+B^2(x)\}\cdot\{1+o(1)\}$, with B and V as in the theorem. In particular, since $g^{(\ell+1)}(x) = (1-p_2-p_1)\breve{g}^{(\ell+1)}(x) = -(\breve{M}/N)(1-p_2-p_1)\breve{p}^{(\ell+1)}(x)$, then here too the bias component of $\mathrm{ASE}_P(x)$ is exactly equal to the bias of the standard regression estimator, and the inaccuracy of the test only affects the variance of the estimator.

We applied our procedure to four examples where the test was not perfectly accurate. In the first we took $p_1 = 0.0003$ and $p_2 = 0$ as in Vansteelandt et al. (2000); in the second we took $p_1 = 0.004$ and $p_2 = 0.077$ as in Xie (2001), in the third we took $p_1 = 0.01$ and $p_2 = 0.1$ and in the fourth we took $p_1 = 0.1$ and $p_2 = 0.01$. To choose the bandwidth, we can use the exact same procedures as before. Indeed, the bandwidth is used to compute the estimator of g, which is calculated in exactly the same way as when the test is accurate. The fact that the Y_i^*'s do not actually represent the true disease status has no impact on the estimation on g; it only impacts the way in which we construct an estimator of $\breve{p}$ from $\widehat{g}$. The results of simulations for model (ii) when $X \sim N(2, 1.5^2)$, using ROT_{ω_0} are shown in Table 3. The numbers show the mean (stdev) of the ratio ISE_I / ISE_P calculated from 200 simulated samples, where ISE_P and ISE_I denote the integrated squared error of the estimator for perfect and imperfect tests, respectively. These results illustrate that the loss caused by imperfect tests remains quite moderate.

6 Extensions

Our method can be extended to a variety of related problems which have been considered in the literature. In this section we introduce two possible generalisations: the case where the explanatory variable is not observed with perfect accuracy, and the case where several covariates are available. In both cases we show how to consistently estimate the unknown curve. We leave details of implementation for future research.

6.1 Covariate measured inaccurately

A problem that often arises in practice is that the covariate X can only be observed with non-negligible measurement error, which is often referred to as an errors-in-variables problem. There, the only data we can observe are $(W_{i,j}, Y_j^*)$, $j = 1, \ldots, J$, $i = 1, \ldots, n_j$, where

$$W_{i,j} \;=\; X_{i,j} + \delta_{i,j}\,, \quad j = 1, \ldots, J,\, i = 1, \ldots, n_j\,, \tag{6.1}$$

and where the error variables $\delta_{i,j}$ are i.i.d. and independent of the $X_{i,j}$'s and of the $Y_{i,j}$'s. Errors-in-variables regression problems have received considerable attention over the last two decades, see Carroll et al. (2006) for an introduction. Huang (2009) and Huang and Tebbs (2009) considered the errors-in-variables problem for parametric estimation of p with group testing data.

In this section we show how to construct a consistent nonparametric estimator of p by combining our procedure in Section 2 with the local polynomial regression method for contaminated data constructed in Delaigle et al. (2009). Throughout, we assume for simplicity that the density f_δ of the $\delta_{i,j}$'s is known, as is commonly done in the literature. This is not very restrictive as, when f_δ is unknown, it can be estimated, parametrically as well as nonparametrically, by replicated measurements, and this does not change first order asymptotic properties of estimators. See Delaigle et al. (2008) for details and conditions. It is easy to see by inspecting our estimator

$\widehat{p}$ defined in Section 2 that the only quantity that needs to be modified is $\widehat{g}$, as $\widehat{\mu}_Z^*$ and $\widehat{q}$ depend only on the Y_j^*'s. Using the same ideas as in Delaigle et al. (2009), we suggest taking

$$\widehat{g}(x) \;=\; \mathbf{e}_0^T \widehat{\mathbf{S}}_{\rm eiv}^{-1} \widehat{\mathbf{T}}_{\rm eiv}\,, \tag{6.2}$$

where $\widehat{\mathbf{S}}_{\rm eiv} = \big(\widehat{S}_{k,k'}^{\rm eiv}\big)_{0\le k,k'\le \ell}$ and $\widehat{\mathbf{T}}_{\rm eiv} = (\widehat{T}_0^{\rm eiv}, \ldots, \widehat{T}_\ell^{\rm eiv})^T$, with

$$\widehat{S}_{k,k'}^{\rm eiv} = \frac{1}{Nh} \sum_{j=1}^{J} \sum_{r=1}^{n_j} \frac{(-i)^{k+k'}}{2\pi} \int e^{-it(x-W_{r,j})/h} \phi_K^{(k+k')}(t)/\phi_\delta(t/h)\, dt$$

$$\text{and} \qquad \widehat{T}_k^{\rm eiv} = \frac{1}{Nh} \sum_{j=1}^{J} Z_j^* \sum_{r=1}^{n_j} \frac{(-i)^{k}}{2\pi} \int e^{-it(x-W_{r,j})/h} \phi_K^{(k)}(t)/\phi_\delta(t/h)\, dt\,,$$

and where ϕ_K and ϕ_δ denote, respectively, the Fourier transform of K and f_δ. Combining our techniques of proofs with those of Delaigle et al. (2009), it can be proved that, under a combination of our condition A and the conditions in Delaigle et al. (2009), this estimator is consistent and has the same rates as the regression estimator studied in Delaigle et al. (2009).

Our results of Section 5, where we considered tests with imperfect accuracy, can be extended to this case too. To estimate $\breve{p}(x)$ defined at (5.1), we can use the estimators of $\breve{q}$ and $\breve{\mu}_Z^*$ as those derived in Section 5, since these do not depend on the mismeasured covariates. The only change concerns the estimator of $\breve{g}$. To estimate $\breve{g}$, we take the formula at (5.2), where we replace g by the estimator $\widehat{g}$ defined at (6.2).

6.2 Multivariate setting

Our procedure can also be extended to the multivariate setting where, for the ith individual of the jth group, we observe a vector $\mathbf{X}_{i,j} = (X_{i,j,1}, \ldots, X_{i,j,d})^T$ of d covariates. Here, for $\mathbf{x} = (x_1, \ldots, x_d)^T$, we have $p(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}) = 1 - q \cdot g(\mathbf{x})/\mu_Z^*$, where

$$g(\mathbf{x}) \;=\; N^{-1} \sum_{j=1}^{J} \sum_{r=1}^{n_j} E\big(Z_j^* | \mathbf{X}_{r,j} = \mathbf{x}\big)$$

and with q and μ_Z^* as in the univariate case. Since $\widehat{q}$ and $\widehat{\mu}_Z^*$ depend only on the Y_j^*'s, they can be estimated as before, and the only difference with the univariate case comes from estimating the function g. For the latter, since the local linear estimator is the most widely used in practice, we extend to the multivariate context only our local linear estimator. This extension is done in a way similar to the standard regression context. See for example Wand and Jones (1995) and Fan and Gijbels (1996). More precisely, we define $\widehat{g}(\mathbf{x}) = \mathbf{e}_0^T \widehat{\mathbf{S}}^{-1} \widehat{\mathbf{T}}$, where $\mathbf{e}_0 = (1, 0, \ldots, 0)^T$ is a vector of length $d+1$, $\widehat{\mathbf{S}} = \left(\widehat{S}_{k,k'}\right)_{0 \le k,k' \le d}$ and $\widehat{\mathbf{T}} = (\widehat{T}_0, \ldots, \widehat{T}_d)^T$, with,

$$\widehat{S}_{k,k'} = \sum_{j=1}^{J} \sum_{r=1}^{n_j} \mathbf{K}_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_{r,j})(x_k - X_{r,j,k})^{\delta_k}(x_{k'} - X_{r,j,k'})^{\delta_{k'}} ,$$

$$\text{and} \quad \widehat{T}_k = \sum_{j=1}^{J} Z_j^* \sum_{r=1}^{n_j} \mathbf{K}_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_{r,j})(x_k - X_{r,j,k})^{\delta_k} ,$$

for $k, k' = 0, \ldots, d$, and where, for each k, $\delta_k = 1\{k > 0\}$. Here, $\mathbf{H}$ is a symmetric positive definite $d \times d$ bandwidth matrix, $\mathbf{K}$ is a d-variate kernel, and we used the standard notation $\mathbf{K}_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2}\mathbf{K}(\mathbf{H}^{-1/2}\mathbf{x})$. For example, $\mathbf{K}$ could be the d-dimensional standard normal density, and it is common to take $\mathbf{H}$ to be a diagonal matrix. Like in the standard regression case, simplified multivariate estimators can also be defined, using for example additive models or their variants. See Fan and Gijbels (1996).

7 Conclusion

In large screening studies, data are often pooled in groups to help significantly reduce cost and increase speed. We have shown how to construct a nonparametric estimator of a conditional probability in this context. Our kernel method relies on a smoothing parameter, and we have derived two automatic ways for choosing this parameter: a simple rule of thumb and a more complex plug-in method. A numerical investiga-

tion has shown that our curve estimator works well in practice. Finally, we have suggested extensions of the procedure to multivariate settings and to cases where the observations are measured with errors.

8 Supplemental Materials

Appendix: The proofs can be found in Sections A and B of the file `Supplem.pdf` which is available online from the JASA website. This file also contains additional simulations for groups of unequal sizes, see Section C.

R code: R codes can be found in the files `Equal.R` and `Unequal.R` available from the JASA website. `Equal.R` includes routines for calculating the estimator in the case of equal group sizes; `Unequal.R` includes routines for calculating the estimator in the case of unequal group sizes.

References

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

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

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.

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

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

Delaigle, A., Hall, P. and Meister, A. (2008). On deconvolution with repeated measurements. *Ann. Statist.* **36**, 665–685.

Delaigle, A. and Meister, A. (2011). Rate-optimal nonparametric estimation in classical and Berkson errors-in-variables problems. *J. Statist. Plann. Inf.* **141**, 102–114.

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

Fahey, J.W., Ourisson, P.J. and Degnan, F.H. (2006). Pathogen detection, testing, and control in fresh broccoli sprouts. *Nutrition J.* **5**:13.

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.

Gasser, T., Kneip, A., and Kohler, W. (1991). A flexible and fast method for automatic smoothing. *J. Amer. Statist. Assoc.* **86**, 643–652.

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.

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.

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.

Huang, X. (2009). An improved test of latent-variable model misspecification in structural measurement error models for group testing data. *Statist. Med.* **28**, 3316–3327.

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

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.

Lennon, J.T. (2007). Diversity and metabolism of marine bacteria cultivated on dissolved DNA. *Applied and Environmental Microbiology* **73**, 2799–2805.

Meister, A. (2010). Nonparametric Berkson regression under normal measurement error and bounded design. *J. Multivariate Anal.* **101**, 1179–1189.

Nagi, M.S. and Raggi, L.G. (1972). Importance to "airsac" disease of water supplies contaminated with pathogenic escherichia coli. *Avian Diseases* **16**, 718–723.

Ruppert, D., Sheather, S.J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90**, 1257–1270.

Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. Springer-Verlag, New York.

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.

Wahed, M.A., Chowdhury, D., Nermell, B., Khan, S.I., Ilias, M., Rahman, M., Persson, L.A. and Vahter, M. (2006). A modified routine analysis of arsenic content in drinking-water in Bangladesh by hydride generation-atomic absorption spectrophotometry. *J. Health, Population and Nutrition* **24**, 36–41.

Wand, M. P and Jones, M.C. (1995). *Kernel Smoothing*. Chapman & Hall, London.

Wand, M. and Ripley, B. (2010). KernSmooth: Functions for kernel smoothing for Wand and Jones (1995). R package version 2.23-4, available at `http://CRAN.R-project.org`.

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