Submitted to the Annals of Statistics
arXiv: math.PR/0000000

NONPARAMETRIC REGRESSION WITH HOMOGENEOUS GROUP TESTING DATA

By Aurore Delaigle*, and Peter Hall*

University of Melbourne

We introduce new nonparametric predictors for homogeneous pooled data in the context of group testing for rare abnormalities, and show that they achieve optimal rates of convergence. In particular, when the level of pooling is moderate then, despite the cost savings, the method enjoys the same convergence rate as in the case of no pooling. In the setting of "over-pooling" the convergence rate differs from that of an optimal estimator by no more than a logarithmic factor. Our approach improves on the random-pooling nonparametric predictor, which is currently the only nonparametric method available, unless there is no pooling, in which case the two approaches are identical.

1. Introduction.

In large screening studies where infection is detected by testing a fluid (e.g. blood, urine, water, etc), data are often pooled in groups before the test is carried out, which permits savings in time and money. This technique, known as group testing, dates back at least to the Second World War, where Dorfman (1943) suggested using it to detect syphilis in U.S. soldiers. It has been used in a variety of large screening studies, for example to detect human immunodeficiency virus, or HIV (Gastwirth and Hammick, 1989), but pooling is also employed to detect pollution, e.g. in water or milk (see Nagi and Raggi, 1972; Wahed et al., 2006; Lennon, 2007; Fahey et al., 2006). Often in these studies, one or several explanatory variables are available, in which case it is generally of interest to estimate the conditional probability of infection. This problem has received considerable attention in the group testing literature, where most suggested techniques are parametric. See, for example, Vansteelandt et al. (2000), Bilder and Tebbs (2009) and Chen et al. (2009). Related work includes that of Chen and Swallow (1990), Gastwirth and Johnson (1994), Hardwick et al. (1998), and Xie (2001).

Thus, although the original purpose of group testing was merely to identify infected individuals more economically, the idea has since been expanded

*Research supported by grants and fellowships from the Australian Research Council.

AMS 2000 subject classifications: Primary 62G08

Keywords and phrases: bandwidth, local polynomial estimator, pooling, prevalence, smoothing

extensively to include more general statistical methodology when the data have to be gathered through grouping. Our paper contributes in this context, developing and describing a particularly effective approach to nonparametric regression. Obtaining information in this way can be useful on its own, or for planning a subsequent study.

Recently, Delaigle and Meister (2011) suggested a nonparametric estimator of the conditional probability of infection. Their method enjoys optimal convergence rates when pooling is random, but it is not consistent in the case of nonrandom, homogeneous pooling, which can be defined as a setting where the covariates of individuals in a group take similar values. In the parametric context it is well known that homogeneous grouping improves the quality of estimators, but the potential gains of homogeneous grouping are even greater in the nonparametric context, where random grouping in moderate to large groups can seriously degrade the quality of estimators.

We demonstrate that, when the data are grouped homogeneously, one can construct more accurate nonparametric estimators of the conditional probability of infection. We show that these improved estimators enjoy faster, and optimal, convergence rates in a variety of contexts. Having reliable estimators of the conditional probability of infection enables more accurate identification of vulnerable categories of people, and can lead to subsequent studies that can assist individuals who are particularly vulnerable to infection. We illustrate the practical performance of our procedure via simulated examples and an application to the National Health and Nutrition Examination Survey (NHANES) study, a large health and nutrition survey collected in the US; see `www.cdc.gov/nchs/nhanes.htm` for more about the NHANES research program.

2. Model and methodology.

2.1. *Main group testing model.* We observe independent and identically distributed (i.i.d.) data $X_1, \ldots, X_N$, where X is a covariate observed on each of N respective objects (e.g. items or individuals), each of which is subject to a potential relatively rare "abnormality". For example, X could be the age or weight of an individual, and the abnormality could be contamination by HIV. Let Y_i denote the result of a test on the ith object, such as blood or urine test. That is, Y_i takes the value 1 or 0 according as the abnormality is detected or not, respectively. In large screening studies, where N is very large, testing each individual for contamination can be too expensive or take too much time, and to overcome this difficulty, it is common to pool data on several individuals before performing the detection test.

Pooling is performed by partitioning the original dataset $\mathcal{X}$, comprised of the values $X_1, \ldots, X_N$, into J subsets, or groups, $\mathcal{X}_1, \ldots, \mathcal{X}_J$, say, where $\mathcal{X}_j$ is of size n_j and $n_1 + \ldots + n_J = N$. We denote the elements of $\mathcal{X}_j$ by $X_{1j}, \ldots, X_{n_j j}$. Each X_{ij} corresponds to an X_k, and each X_k has a concomitant Y_k. If the ith element X_{ij} of $\mathcal{X}_j$ is X_k, then the concomitant of X_{ij} is $Y_{ij} = Y_k$. Instead of trying to determine the value of Y_{ij} directly, each group $\mathcal{X}_j$ is tested to discover whether the abnormality is present in the group, i.e. to determine the value of

$$Y_j^* = \max_{1 \le i \le n_j} Y_{ij} = \begin{cases} 1 & \text{if } Y_{ij} = 1 \text{ for some } i \text{ in the range } 1 \le i \le n_j \\ 0 & \text{otherwise} \ . \end{cases}$$

Of course, Y_j^* is obtained without observing the Y_{ij}s directly; for example, when the abnormality is detected by a blood test, the bloods of all individuals in a group are mixed together, and this mixed blood is tested for contamination. From the data pairs $(\mathcal{X}_j, Y_j^*)$ we wish to estimate the probability function $p(x) = P(Y_i = 1 \,|\, X_i = x) = E(Y_i = 1 \,|\, X_i = x)$.

Since p is a regression curve, then if the sample $(X_i, Y_i), i = 1, \ldots, N$ were observed, we could use standard nonparametric regression techniques such as, for example, local polynomial estimators. Let $\ell \ge 0$ be an integer, $h > 0$ a bandwidth, K a kernel function and $K_h(x) = h^{-1} K(x/h)$. The standard ℓth degree local polynomial estimator of p is defined by

$$\widehat{p}_S(x) = (1, 0, \ldots, 0) \mathbf{Q}^{-1} \mathbf{R}, \tag{2.1}$$

where $\mathbf{R} = \big(R_0(x), \ldots, R_\ell(x)\big)^T$, $\mathbf{Q} = \big(\mathbf{Q}_{ij}\big)_{1 \le i,j \le \ell+1}$, with $\mathbf{Q}_{ij} = Q_{i+j-2}(x)$, and where $Q_k(x) = \sum_{i=1}^N (X_i - x)^k K_h(X_i - x)$ and $R_k(x) = \sum_{i=1}^N Y_i (X_i - x)^k K_h(X_i - x)$. See, for example, Fan and Gijbels (1996). Of course, when the data are pooled the Y_is are not available and we cannot calculate such estimators. Therefore we need to develop specific ways to estimate p from pooled data.

2.2. *Method for homogeneous pools.* Depending on the study, it is not always possible to observe the X_is before pooling the data, so that the individuals are pooled randomly. This is the context of the work of Delaigle and Meister (2011), who constructed a nonparametric estimator for the case where data X_i are assigned randomly to the groups $\mathcal{X}_j$. See appendix A.1 of the supplemental article (Delaigle and Hall, 2011) for a summary of properties of their estimator. In other studies, the X_is are observed beforehand; see, for example, the study of hepatitis C infection among 10,654 healthcare workers in Scotland, carried out by Thorburn et al. (2001). In such cases, it has already been demonstrated in the parametric context that it can be greatly advantageous to pool the data non randomly; see Vansteelandt et al. (2000).

Unfortunately, the only nonparametric estimator available for group testing data (see Delaigle and Meister, 2011) crucially relies on random grouping and is not valid when homogeneous groups are created. Below we suggest a new nonparametric approach which is valid with homogeneous pooling. We introduce our procedure in the case of a single covariate and equal-sized groups. Generalisations of our method to unequal group sizes, and multiple covariates, will be treated in section 5. These generalisations are similar in most respects.

To create homogeneous pools we divide the data into equal-number groups, taking the jth group to be $\mathcal{X}_j = \{X_{((j-1)\nu+1)}, \ldots, X_{(j\nu)}\}$, where $\nu = n_j$, in this case not depending on j, is the number of data in each group, and $X_{(1)} \leq \ldots \leq X_{(N)}$ denotes an ordering of the data in $\mathcal{X}$. We assume that ν divides N; the case where it does not is a particular case of our generalisation in section 5. Note that, with $Z_j^* = 1 - Y_j^*$,

$$E(Z_j^* \,|\, \mathcal{X}) = \prod_{i=1}^{\nu} \{1 - p(X_{ij})\} \,. \tag{2.2}$$

The right-hand side here is generally close to $\{1 - p(\bar{X}_j)\}^\nu$, where $\bar{X}_j = \nu^{-1} \sum_i X_{ij}$ denotes the average value of the X_{ij}s in the jth group, and that closeness motivates the definition of $\widehat{p}(x)$ at (2.4), below. Let

$$\mu(x) = \{1 - p(x)\}^\nu \,. \tag{2.3}$$

Reflecting (2.2) and the above discussion, we suggesting estimating $p(x)$ by

$$\widehat{p}(x) = 1 - \widehat{\mu}(x)^{1/\nu} \,, \tag{2.4}$$

where $\widehat{\mu}$ is a nonparametric estimator of μ.

It remains to estimate μ. We begin by giving motivation for our methodology. Since, by construction, the groups are homogeneous, the observations in a given group are similar. In particular, $p(X_{((j-1)\nu+1)}), \ldots, p(X_{(j\nu)})$ are well approximated by $p(\bar{X}_j)$. Together, this and the identity (2.2) suggest that $\mu(\bar{X}_j)$ can be approximated by $E(Z_j^* \,|\, \bar{X}_j)$, so that $\mu(x)$ is approximately equal to the average of the $E(Z_j^* \,|\, \bar{X}_j)$s over the $\bar{X}_j$s close to x, which can be estimated by standard nonparametric regression estimators calculated from the data $(\bar{X}_j, Z_j^*)$, $j = 1, \ldots, J$. Motivated by these considerations, we define an ℓth order local polynomial estimator of μ, constructed from the data $(\bar{X}_j, Z_j^*)$, by

$$\widehat{\mu}(x) = (1, 0, \ldots, 0)\mathbf{S}^{-1}\mathbf{T}, \tag{2.5}$$

where $\mathbf{T} = \big(T_0(x), \ldots, T_\ell(x)\big)^T$ and $\mathbf{S} = \big(\mathbf{S}_{ij}\big)_{1\le i,j\le \ell+1}$, with $\mathbf{S}_{ij} = S_{i+j-2}(x)$, $S_k(x) = \sum_j (\bar{X}_j - x)^k K_h(\bar{X}_j - x)$, and $T_k(x) = \sum_j Z_j^* (\bar{X}_j - x)^k K_h(\bar{X}_j - x)$.

We shall show in section 3 that this approach is well founded, by proving consistency of the resulting estimator $\widehat{p}$ of p. We shall develop our theoretical results for a larger class of estimators which encompasses the estimator at (2.5).

3. Theoretical properties.

To study properties of our estimator it is convenient to express the probability p, at a particular x, as

$$p(x) = \delta(N)\pi(x)\,, \tag{3.1}$$

where $\delta = \delta(N)$ denotes a sequence of positive numbers that potentially depend on N, and π is a fixed, nonnegative function. To be as general as possible, we permit the group size $\nu = \nu(N) \ge 1$ to increase, and $\delta = \delta(N)$ to decrease, as N diverges.

In large screening studies the abnormalities under investigation are invariably rare, i.e. p is small. To understand the limitations of our estimator, we shall study properties in the extreme situation where $\delta \to 0$ (and hence $p \to 0$) as $N \to \infty$. More precisely, we shall consider the "low prevalence" situation where $\nu\delta \to 0$ as $N \to \infty$, which is an asymptotic representation of the case where the group size ν is relatively small and infection is rare. In practice, groups are rarely taken larger than 10 to 20. One reason for this is that, depending on the proportion of positive individuals in the population, some tests (e.g. HIV tests) become too unreliable if the pool size is too large (larger than $\nu = 5$ to 10 in the HIV example). To reflect this fact, we shall also consider the standard "moderate pooling" situation where $\nu\delta \to c > 0$ as $N \to \infty$. However, there are tests for which groups could be taken as large as $\nu = 40$ to 50. From the viewpoint of economics, large groups would be beneficial, and might even be the only possible way to screen individuals in poor countries. Hence we need to understand their effects on the quality of estimators. We shall do this by investigating asymptotic properties of our estimator in the extreme "over-pooling" situation where $\nu\delta \to \infty$ as $N \to \infty$.

3.1. *Conditions.*

We shall derive theoretical properties of the estimator $\widehat{p}$ defined at (2.4), where for $\widehat{\mu}$ we shall generalise the local polynomial estimators introduced at (2.5), by considering a whole class of linear smoothers, defined by

$$\widehat{\mu}(x) = \sum_j w_j(x)\, Z_j^* \Big/ \sum_j w_j(x)\,, \tag{3.2}$$

where the weights w_j depend on $\mathcal{X}$ but not on the variables Z_j^*. The local polynomial estimator defined at (2.5) can be rewritten easily in this form, and other popular nonparametric estimators (e.g. smoothing splines) can be expressed in this form too; see, for example, Ruppert et al. (2003).

Recall that $\bar{X}_j = \nu^{-1} \sum_i X_{ij}$ and let $h = h(N)$ denote a sequence of constants decreasing to zero as $N \to \infty$. We can interpret $h(N)$ as the bandwidth in a kernel-based construction of the weight functions w_j in (3.2). Typically, the weights w_j would depend on $\bar{X}_j$, and we assume that, for each $x \in \mathcal{I}$, where $\mathcal{I}$ is a given compact, nondegenerate interval:

Condition S:
(S1) $\sum_j w_j(x)\,(\bar{X}_j - x)/\sum_j w_j(x) = 0\,,$
(S2) $\sum_j w_j(x)\,(\bar{X}_j - x)^2/\sum_j w_j(x) = h^2\,b(x) + o_p\big(h^2\big)\,,$
(S3) $\sum_j w_j(x)^2/\{\sum_j w_j(x)\}^2 = \nu\,v(x)/(Nh) + o_p\{\nu/(Nh)\}\,,$
(S4) for each integer $k \geq 1$, $\sum_j |w_j(x)|^k/\{\sum_j w_j(x)\}^k = O_p\big[\{\nu/(Nh)\}^{k-1}\big]\,,$

where the functions b and v are continuous on $\mathcal{J}$ and are related to the type of estimator. We also assume that:

Condition T:
(T1) the distribution of X has a continuous density, f, that is bounded away from zero on an open interval $\mathcal{J}$ containing $\mathcal{I}$;
(T2) $p = \delta\pi$ is bounded away from 1 uniformly in $x \in \mathcal{I}$ and in $N \geq 1$;
(T3) the function π in (3.1) has two Hölder-continuous derivatives on $\mathcal{J}$;
(T4) for some $\epsilon > 0$, $h + \nu\delta h + (\nu^2/N^{1-\epsilon}h\delta) \to 0$ as $N \to \infty$;
(T5) the weights $w_j(x)$ vanish for $|\bar{X}_j - x| > C\,h$, where $C > 0$ is a constant.

The assumption, in (T1), that f is bounded away from zero on a compact interval allows us to avoid pathological issues that arise when too few values of X are available in neighbourhoods of zeros of f. Finally, when describing the size of $\widehat{p}(x) - p(x)$ simultaneously in many values x we shall ask that for some $C, \epsilon > 0$,

$$\sup_{x,\,x' \in \mathcal{I}\,:\,|x-x'| \leq N^{-C}} \left\{ \frac{1}{|x-x'|^\epsilon} \sum_k \left| \frac{w_k(x)}{\sum_j w_j(x)} - \frac{w_k(x')}{\sum_j w_j(x')} \right| \right\} = O_p(1)\,. \quad (3.3)$$

For example, if the weights w_j correspond to the local polynomial estimator in (2.5) with $\ell = 1$ (i.e. the local linear estimator), with bandwidth h and a compactly supported, symmetric, Hölder continuous, nonnegative kernel K satisfying $\int K = 1$; if $h + (Nh)^{-1} = O(N^{-\epsilon_1})$ for some $\epsilon_1 > 0$, and (T1) holds; then (T5), S and (3.3) hold with, in (S2) and (S3), $b = \int u^2\,K(u)\,du$

(not depending on x) and $v(x) = f(x)^{-1} \int K^2$. Furthermore, S holds uniformly in $x \in \mathcal{I}$. More generally it is easy to see that when $\ell > 1$, the ℓth order local polynomial estimator in (2.5) satisfies $\sum_j w_j(x)\,(\bar{X}_j - x)^k = 0$ for $k = 0, \ldots, \ell-1$, and hence conditions (S1) and (S2) are trivially satisfied. Conditions (S3) and (S4) too are satisfied in this case, under mild conditions on the kernel. Note that condition (S1) is not satisfied in the local constant case ($\ell = 0$ in (2.5)). Although this instance can be easily accommodated by modifying our conditions slightly, we simply omit it from our theory because in practice the local linear estimator is almost invariably preferred to the local constant one.

Remark 1. Instead of linear smoothers, such as local polynomial estimators, we could use alternative procedures which are sometimes preferred in the context of binary dependent variables. For example, Fan, Heckman and Wand (1995) suggest modelling the regression curve m by $m(x) = g^{-1}\{\eta(x)\}$, where g is a known link function and η is an unknown curve. These methods have theoretical properties similar to those of local polynomial estimators; the two methods differ mostly through their bias, and, depending on the shapes of m and g, one method has a smaller bias than the other. We prefer local polynomial estimators because they are easier to implement in practice.

3.2. *Low prevalence and moderate pooling.* Our first result establishes convergence rates and asymptotic normality for the estimator $\widehat{p}$ defined at (2.4), with $\widehat{\mu}$ at (3.2). Note that we do not insist that ν and δ vary with N; the regularity conditions for Theorem 3.1 hold in many cases where ν and δ are both fixed. Below we use the notation $A(x)$ to denote the value taken by a function A at a point x, and the notation A when referring to the function itself. However, in some places, for example in result (3.4) where it is necessary to refer explicitly to the point x mentioned in the statement "for all $x \in \mathcal{I}$"; and in definitions (3.5) and (3.6), where we are defining functions; the two notations may appear a little ambiguous.

Theorem 3.1. *Assume that Conditions S and T hold, and that $\nu\delta = O(1)$. Then, for each $x \in \mathcal{I}$,*

$$\widehat{p}(x) - p(x) = A(x)\,V(x) + B(x) + o_p\big\{\delta h^2 + (\delta/Nh)^{1/2}\big\}\,, \tag{3.4}$$

where the distribution of $V(x)$ converges to the standard normal law as $N \to \infty$, and the functions A and B are given by

$$A \;=\; \Big[(\nu N h)^{-1}\,(1-p)^{2-\nu}\,\big\{1-(1-p)^{\nu}\big\}\,v\Big]^{1/2} = O\big\{(\delta/Nh)^{1/2}\big\}, \tag{3.5}$$

$$B \;=\; \tfrac{1}{2}\, h^2 \,\big\{p'' - (\nu - 1)\,(1 - p)^{-1}\,(p')^2\big\}\, b = O\big(\delta\, h^2\big)\,, \tag{3.6}$$

where b and v are as in (S2) and (S3). If, in addition, condition S holds uniformly in $x \in \mathcal{I}$, if (3.3) *holds, and if the functions b and v are bounded and continuous, then*

$$\int_{\mathcal{I}} (\widehat{p} - p)^2 = \int_{\mathcal{I}} (A^2 + B^2) + o_p\big\{\delta^2 h^4 + (\delta/Nh)\big\}\,. \tag{3.7}$$

Note that A and B represent, to first order, the standard deviation of the error about the mean, and the main effect of bias, which arise from the asymptotic distribution. For simplicity we shall call A^2 and B the asymptotic variance and bias of the estimator. From the theorem we see that, when $B(x) \neq 0$ (e.g. for the local polynomial estimator with $\ell = 1$), if $N\delta \to \infty$ as $N \to \infty$, then the rate of the estimator is optimised when h is of size $(N\delta)^{-1/5}$, in which case the estimator satisfies:

$$\text{for each } x \in \mathcal{I}\,, \quad \widehat{p}(x) - p(x) = O_p\big\{(\delta^3/N^2)^{1/5}\big\}\,. \tag{3.8}$$

Note that when $\nu = 1$ (no grouping), $\mu = 1-p$ and our estimator of p reduces to a standard local linear smoother of $1 - \mu$. For example, the estimator at (2.5) coincides with $1 - \widehat{p}_S$ in (2.1). Taking $\nu = 1$ in the theorem, we deduce that the convergence rate of our estimator for $\nu > 1$, given at (3.8), coincides with the rate for conventional linear smoothers employed with non-grouped data. By standard arguments it is straightforward to show that this rate is optimal when π has two derivatives, and hence our estimator is rate optimal. Although, in (T3), we assume that π has two continuous derivatives, continuity is imposed only so that the dominant term in an expansion of bias can be identified relatively simply, and the convergence rate at (3.8) can be derived without the assumption of continuity. In addition, note that when $\nu\delta = o(1)$ our estimator has the same asymptotic bias and variance expressions, B and A, as the estimator when $\nu = 1$, which in that case reduce to $A = (\delta/Nh)^{1/2}\,(\pi\, v)^{1/2}$ and $B = \frac{1}{2}\,\delta\, h^2\, \pi''\, b + o_p\big(\delta\, h^2\big)$. In other words, in that case the statistical cost of pooling is virtually zero.

The results discussed above also apply if performance is measured in terms of integrated squared error (ISE), as at (3.7). In particular, if h is of size $(N\delta)^{-1/5}$, provided that $\nu\delta$ is bounded, the estimator $\widehat{p}$ achieves the minimax optimal convergence rate:

$$\int_{\mathcal{I}} (\widehat{p} - p)^2 = O_p\big\{(\delta^3/N^2)^{2/5}\big\}\,. \tag{3.9}$$

Remark 2. Similar conclusions can be drawn in the case of estimators for which $B(x) = 0$, but this requires us to assume that the function π has enough derivatives so that an explicit, asymptotic, dominating, non zero bias term can be derived. For example, for our local polynomial estimator of order $\ell > 1$, we have $B(x) = 0$ and the term $o_P(\delta h^2)$ is only an upper bound to the bias of the estimator. A non vanishing asymptotic expression for the bias can easily be obtained for $\ell > 1$ if we assume that π has $\ell + 1$ continuous derivatives. This can be done in a straightforward manner, but to keep presentation simple, and since in practice local linear estimators are almost invariably preferred to other local polynomial estimators, we omit such expansions.

Remark 3. In the case where $\delta \to 0$ it could be argued that the rates are meaningless since we are trying to estimate a function that tends to zero, and that it is more appropriate to consider the non zero part π of p in the model at (3.1), and see how fast $\widehat{\pi} = \widehat{p}/\delta$ converges to π. The convergence rate of $\widehat{\pi}$ is easily deducible from (3.8):

$$\text{for each } x \in \mathcal{I}\,, \quad \widehat{\pi}(x) - \pi(x) = O_p\big\{(N\delta)^{-2/5}\big\}\,. \tag{3.10}$$

Provided that $N\delta \to \infty$ as $N \to \infty$, $\widehat{\pi}(x)$ is consistent for $\pi(x)$ and the convergence rate evinced by (3.10) is optimal.

3.3. *Over-pooling.* The situation is quite different when $\nu\delta \to \infty$ as $N \to \infty$, which can be interpreted as an asymptotic representation of the situation where the data are pooled in groups of relatively large size ν. In practical terms the results in this section serve as a salutary warning not to skimp on the testing budget. The work in section 3.2 shows that the performance of estimators is robust, up to a point, against increasing group size, but in the present section we demonstrate that, after the dividing line between moderate pooling and overpooling has been crossed, performance decreases sharply.

When $\nu\delta \to \infty$, properties of the estimator of $p(x)$ depend on x, because there the order of magnitude of $\mu(x)$, at (2.3), depends critically on the rate at which $\{1 - p(x)\}^\nu$ converges to zero. The following condition captures this aspect:

$$\text{for some } \epsilon > 0\,, \quad \nu/h = o\big[N^{1-\epsilon}\,\{1 - \delta\,\pi(x)\}^\nu\big]\,, \tag{3.11}$$

and the following theorem replaces Theorem 3.1.

THEOREM 3.2. *Assume that $\nu\delta \to \infty$ as $N \to \infty$, conditions S, T and* (3.11) *hold, and π, b and v are all nonzero at x. Then $\widehat{p}(x) - p(x) = A(x)\, V(x) + \{1 + o_p(1)\}\, B(x)$, where $V(x)$ is asymptotically distributed as a normal $N(0,1)$ as $N \to \infty$, and A and B are given by the first identities in each of* (3.5) *and* (3.6)*.*

Note that the orders of magnitude given by the second identities in each of (3.5) and (3.6) are not valid in this case, and neither does result (3.7) necessarily hold under the conditions of Theorem 3.2. Note too that the theorem can be extended to cases where $b = 0$, along the lines discussed in Remark 2. To elucidate the implications of Theorem 3.2, assume that $\pi'(x)$ is nonzero and define $\lambda_N(x)^5 = \{1 - \delta\, \pi(x)\}^{-\nu}$, which, when $\nu\delta \to \infty$, diverges exponentially fast as a function of $\nu\delta$. Given a sequence of constants c_N and a sequence of random variables V_N, write $V_N \asymp_p c_N$ to indicate that both $V_N = O_p(c_N)$ and $c_N = O_p(V_N)$ as $N \to \infty$. Theorem 3.2 implies that, if $\nu\delta \to \infty$ and h is a constant multiple of $\lambda_N (N\delta^4\nu^3)^{-1/5}$, then

$$\{\widehat{p}(x) - p(x)\}^2 \asymp_p (\delta^3/N^2)^{2/5}\, (\nu\delta)^{-2/5}\, \lambda_N(x)^4, \tag{3.12}$$

and in particular diverges at a rate that is exponentially slower, as a function of $\nu\delta$, than in the case where $\nu\delta = O(1)$ treated in section 3.2. Result (3.12) follows from the fact that $A(x)^2 \asymp (\nu N h)^{-1}\, \lambda_N(x)^5$ and $|B(x)| \asymp h^2\, \nu\, \delta^2$, where $a_1(N) \asymp a_2(N)$ means that $a_1(N)/a_2(N)$ is bounded away from zero and infinity. Note that (3.12) includes the case where p (and hence δ) is held fixed, and $\nu \to \infty$ as $N \to \infty$.

The result at (3.12) shows that when $\nu \to \infty$ as $N \to \infty$, $\widehat{p}$ suffers from a clear degradation of rates compared to the case where $\nu\delta = O(1)$. Next we show that this degradation is intrinsic to the problem, not to our estimator $\widehat{p}$; any estimator based on the pooled data in section 2.2 will experience an exponentially rapid decline in performance as $\nu\delta \to \infty$. More precisely we show in Theorem 3.3 that, when $\nu\delta \to \infty$ as $N \to \infty$, $\widehat{p}$ is near rate-optimal among all such estimators. Recall that, under our model (3.1), $p = \delta\, \pi$, where $\delta = \delta(N)$ potentially converges to zero. If $\nu\delta \to \infty$ then, by (3.12), we have:

$$|\widehat{p}(x) - p(x)| = O_p\Big[\big(\delta^3/N^2\big)^{1/5}\, (\nu\delta)^{-1/5}\, \{1 - p(x)\}^{-2\nu/5}\Big]\,. \tag{3.13}$$

Although this result was derived under the assumption that π is a fixed function with two continuous derivatives, since (3.13) is only an upper bound then it is readily established under the following more general assumption:

the nonnegative function $\pi = \pi_N$ can depend on N and satisfies $\pi_N(x) + |\pi'_N(x)| + |\pi''_N(x)| \le C_1$, for all N and all x, where the constant $C_1 > 0$ does not depend on N or x. (3.14)

Take the explanatory variables X_i to be uniformly distributed on the interval $\mathcal{M} = [-\frac{1}{2}, \frac{1}{2}]$, and let $\mathcal{I} \subset \mathcal{J} \subset \mathcal{M}$ where 0 is an interior point of $\mathcal{I}$. Let $p^1 = \delta\, \pi_N$, where π_N satisfies (3.14), let $p^0 \equiv \delta$ denote the version of p^1 when $\pi_N \equiv 1$, and consider the condition:

$$(\nu^3 \delta)^{1/2} = o\big\{N\,(1-\delta)^\nu\big\}\,. \tag{3.15}$$

This assumption permits $\nu\delta$ to diverge with N, but not too quickly. Indeed, using arguments similar to those in section 6.3 it can be shown that if (3.15) fails then no estimator of p is consistent. Let $\mathcal{P}$ be the class of measurable functions $\check{p}$ of the pooled data pairs $(\mathcal{X}_j, Y_j^*)$ introduced in section 2.2.

Theorem 3.3. *Assume that p^0 and p^1 are bounded below 1, that* (3.15) *holds and that $\nu\delta \to \infty$. Let x be an interior point of the support, $[-\frac{1}{2}, \frac{1}{2}]$, of the uniformly distributed explanatory variables X_i. Then $C_2 > 0$, and π_N, satisfying* (3.14)*, can be chosen such that*

$$\liminf_{n\to\infty} \max_{p=p^0,\,p^1} \inf_{\check{p}\in\mathcal{P}} P\Big[|\check{p}(x) - p(x)| > C_2\, \delta^{3/5} (N\nu\delta)^{-2/5}\, \{1-p(x)\}^{-2\nu/5}\Big] > 0\,. \tag{3.16}$$

Except for the fact that $(\nu\delta)^{-2/5}$, rather than $(\nu\delta)^{-1/5}$, appears in (3.16), the latter result represents a converse to (3.13). The difference in powers here is of minor importance since the main issue is the factor $\{1-p(x)\}^{-2\nu/5}$, which (in the context $\nu\delta \to \infty$ of over-pooling), diverges faster than any power of $\nu\delta$, and this feature is represented in both (3.13) and (3.14).

3.4. *Comparison with the approach of Delaigle and Meister.* Arguments similar to those of Delaigle and Meister (2011) can be used to show that, under conditions similar to those used in our Theorem 3.1, their estimator $\tilde{p}$ (see (A.1) in the supplemental article, Delaigle and Hall, 2011) satisfies $\tilde{p}(x) - p(x) = A_1(x)\, V_1(x) + B_1(x) + o_p\big\{\delta h^2 + (\nu\delta/Nh)^{1/2}\big\}$, where the random variable $V_1(x)$ has an asymptotic standard normal distribution and

$$A_1 = \Big[(Nh)^{-1}\,(1-p)\, q^{1-\nu}\, \big\{1 - (1-p)\, q^{\nu-1}\big\}\, v\Big]^{1/2} = O(\nu\delta/Nh)\,, \tag{3.17}$$

$$B_1 = \tfrac{1}{2}\, h^2\, p''\, b = O\big(\delta h^2\big)\,, \tag{3.18}$$

with $q = E\{1-p(X)\}$. Likewise, the analogue of (3.7) can be derived: $\int_{\mathcal{I}} (\tilde{p}-p)^2 = \int_{\mathcal{I}} (A_1^2 + B_1^2) + o_p\{\delta^2 h^4 + (\nu\delta/Nh)\}$. To simplify the comparison, assume that we use estimators for which b and v do not vanish, and that $\pi > 0$. We see when comparing (3.17)–(3.18) with (3.5)–(3.6) that the asymptotic variance term A^2 of our estimator is an order of magnitude ν times smaller than A_1^2. Note too the asymptotic bias terms of $\widehat{p}$ and $\tilde{p}$ are of the same size (the two biases are asymptotically equivalent if $\nu\delta \to 0$, and have the same magnitude in other cases). Hence, with our procedure the gain in accuracy can be quite substantial, especially if ν is large.

4. Numerical study. We applied the local linear version of our local polynomial estimation procedure (i.e. the one based on (2.5) with $\ell = 1$) on simulated and real examples. This method, which we denote below by DH, is the one we prefer because it works well and it is very easy to implement, and we can easily derive and compute a good data-driven bandwidth for it. The practical advantages of local linear estimators over other local polynomial estimators have been discussed at length in the standard nonparametric regression literature. Of course, other versions of our general local linear smoother procedure can be used, such as a spline approach or more complicated iterative kernel procedures (see Remark 1). Each of the methods gives essentially the same estimator.

In our simulations we compared the DH procedure, calculated by definition from homogeneous groups, with the local linear estimator $\widehat{p}_S$ at (2.1) that we would use if we had access to the original non grouped data. We also compared DH with the local linear version of the method of Delaigle and Meister (2011) which, by definition, is calculated from randomly created groups. We denote these two methods by LL and DM, respectively. We took the kernel, K, equal to the standard normal density. For h, in the DM case we used the plug-in bandwidth of Delaigle and Meister (2011) with their weight ω_0; we used a similar plug-in bandwidth in the LL and DH cases, see section A.2 of the supplemental article (Delaigle and Hall, 2011) for details.

4.1. *Simulation results.* To facilitate the comparison with the DM method, we simulated data according to the four models used by Delaigle and Meister (2011):
(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 generated 200 samples from each model, with X normal or uniform, and with $N = 1,000$, $N = 5,000$ and $N = 10,000$. Then for the DH method we

TABLE 1

Simulation results for models (i) to (iv), when the $X_{i,j}$s are uniform. The numbers show $10^4\times$ MED (IQR) of the ISE calculated from 200 simulated samples.

		$\nu = 1$	$\nu = 5$		$\nu = 10$		$\nu = 20$	
Model	N	LL	DH	DM	DH	DM	DH	DM
(i)	10^3	9.35(7.42)	10.1(8.16)	26.9(24.1)	11.0(8.53)	51.2(49.6)	17.8(484)	122(110)
	5.10^3	2.91(2.01)	2.94(2.38)	7.59(5.34)	3.30(2.06)	14.1(11.4)	4.46(2.94)	29.2(25.2)
	10^4	1.62(1.20)	1.83(1.40)	4.54(3.05)	2.07(1.63)	7.70(6.13)	2.89(1.95)	16.8(13.9)
(ii)	10^3	6.37(8.38)	8.66(9.99)	29.4(28.4)	10.3(11.4)	64.7(69.5)	29.7(1560)	166(169)
	5.10^3	1.48(1.37)	1.66(2.26)	6.37(5.93)	2.41(2.74)	13.8(12.1)	4.47(5.94)	35.8(30.0)
	10^4	.963(.843)	1.02(1.16)	3.39(2.89)	1.35(1.25)	7.04(6.20)	2.35(3.26)	19.1(17.2)
(iii)	10^3	.777(.978)	.860(1.26)	3.44(4.03)	1.02(1.31)	7.26(8.37)	1.90(4.81)	19.9(19.5)
	5.10^3	.176(.220)	.166(.254)	.722(.818)	.214(.298)	1.68(1.67)	.356(.482)	4.48(3.97)
	10^4	.093(.108)	.100(.128)	.355(.344)	.117(.158)	.797(.800)	.200(.212)	2.28(1.79)
(iv)	10^3	2.33(2.11)	2.49(2.32)	7.41(9.81)	2.70(2.55)	17.2(16.3)	5.07(166)	39.7(34.1)
	5.10^3	.590(.510)	.633(.602)	2.01(1.73)	.637(.702)	4.05(3.70)	.964(1.06)	9.62(9.11)
	10^4	.309(.254)	.317(.293)	1.10(.873)	.373(.311)	2.31(1.89)	.570(.539)	5.47(4.80)

TABLE 2

Simulation results for models (i) to (iv), when the $X_{i,j}$s are normal. The numbers show $10^4\times$ MED (IQR) of the ISE calculated from 200 simulated samples.

		$\nu = 1$	$\nu = 5$		$\nu = 10$		$\nu = 20$	
Model	N	LL	DH	DM	DH	DM	DH	DM
(i)	10^3	10.3(6.69)	10.7(7.18)	20.8(19.0)	10.8(8.04)	37.0(35.3)	12.8(9.70)	85.6(72.8)
	5.10^3	4.35(2.80)	4.14(2.71)	9.60(5.49)	4.32(2.95)	12.0(11.1)	4.50(3.44)	17.3(18.8)
	10^4	3.12(1.77)	3.33(2.07)	7.66(4.12)	3.01(2.01)	9.42(5.68)	3.20(2.19)	13.6(11.0)
(ii)	10^3	5.02(5.20)	5.78(6.83)	17.0(23.0)	8.18(10.6)	46.0(57.8)	21.1(64.0)	167(202)
	5.10^3	1.69(1.95)	1.98(2.18)	4.23(5.97)	2.36(3.40)	9.48(12.3)	5.37(6.75)	28.3(36.9)
	10^4	1.02(.925)	1.17(1.21)	2.99(3.12)	1.46(1.64)	5.51(6.81)	3.04(3.22)	15.0(17.7)
(iii)	10^3	.897(1.53)	.885(1.06)	2.95(3.36)	.910(1.27)	5.73(7.10)	1.37(2.14)	23.7(27.3)
	5.10^3	.274(.389)	.263(.325)	.946(.997)	.260(.383)	1.61(2.08)	.448(.692)	4.26(4.93)
	10^4	.204(.270)	.148(.175)	.637(.725)	.182(.219)	1.13(1.10)	.323(.435)	2.42(2.58)
(iv)	10^3	4.13(4.30)	3.60(3.48)	13.2(12.5)	4.32(3.84)	28.1(26.9)	7.60(9.43)	82.3(75.2)
	5.10^3	1.30(1.33)	1.10(1.01)	3.85(3.77)	1.21(1.22)	7.45(6.56)	2.24(2.20)	16.6(18.1)
	10^4	.764(.651)	.566(.474)	2.50(1.86)	.676(.672)	4.63(4.03)	1.01(1.04)	10.1(9.96)

split each sample homogeneously into groups of equal sizes $\nu = 5$, $\nu = 10$ or $\nu = 20$; for the DM method, we created the groups randomly (remember that this estimator is valid only for random groups).

To assess the performance of our DH estimator we calculated, in each case and for each of the 200 generated samples, the integrated squared error ISE $= \int_a^b (\widehat{p} - p)^2$, with a and b denoting the 0.05 and 0.95 quantiles of the distribution of X. We did the same for the DM and LL estimators $\tilde{p}$ and $\widehat{p}_S$. For brevity, figures illustrating the results are provided in section A.4 of the supplemental article (Delaigle and Hall, 2011), and here we show only summary statistics. In the graphs of section A.4, we show the target curve (thin uninterrupted curve) as well as three interrupted curves; these were calculated from the samples that gave the first, second and third quartiles of the 200 ISE values.

In Table 1 we show, for each model with X uniform, the median (MED) and interquartile range (IQR) of the 200 ISE values obtained using the LL estimator based on non grouped data, and, for several values of ν, the DH and the DM approaches based on data pooled in groups of size ν; Table 2 shows the same but for X normal. Note that LL cannot be calculated from grouped data, but we include it to assess the potential loss incurred by pooling the data. The tables show that for $\nu \leq 10$, pooling the data homogeneously hardly affects the quality of the estimator. Sometimes, the results are even slightly better with the DH method than with the LL one. Indeed a careful analysis of the bias and variance of the various estimators shows that for some curves $p(x)$, grouping homogeneously can sometimes be slightly beneficial when ν is small (roughly this is because by grouping a little we lose very little information, but we increase the number of Y_j^* positive, which makes the estimation a little easier for this particular estimator. Theoretical arguments support this conclusion). The situation is much less favourable for the DM random grouping method, whose quality degrades quickly as ν increases. Unsurprisingly, DH beat DM systematically, except when N/ν was small ($N = 1,000$ and $\nu = 20$), where the $J = 50$ grouped observations did not suffice to estimate very well the curves from models (i) and (ii).

4.2. *Real data application.* We also applied our DH method on real data. To make the comparison with the LL estimator possible we used data for which we had access to the entire, non grouped set of observations (X_i, Y_i). Then we grouped the data and compared the DH and LL procedures. We used data from the NHANES study, which are available at `www.cdc.gov/nchs/nhanes/nhanes1999-2000/nhanes99_00.htm`. These data were collected in the US between 1999 and 2000.

As in Delaigle and Meister (2011), our goal was to estimate two conditional probabilities: $p_{\mathrm{HBc}}(x) = E(Y_{\mathrm{HBc}}|X = x)$ and $p_{\mathrm{CL}}(x) = E(Y_{\mathrm{CL}}|X = x)$, where X was the age of a patient, $Y_{\mathrm{HBc}} = 0$ or 1 indicating the absence or presence of antibody to hepatitis B virus core antigen in the patients serum or plasma, and $Y_{\mathrm{CL}} = 0$ or 1 indicating the absence or presence of genital Chlamydia trachomatis infection in the urine of the patient. The sample size was $N = 7,016$ for HBc and $N = 2,042$ for CL. The percentage of Y_i's equal to one was 0.047 in the HBc case and 0.044 in the CL case. See Delaigle and Meister (2011) for more details on these data and the methods employed to collect them.

For brevity here we only present the results obtained using our method by pooling the data homogeneously in groups of equal size $\nu = 2$, 5, 10 and 20.

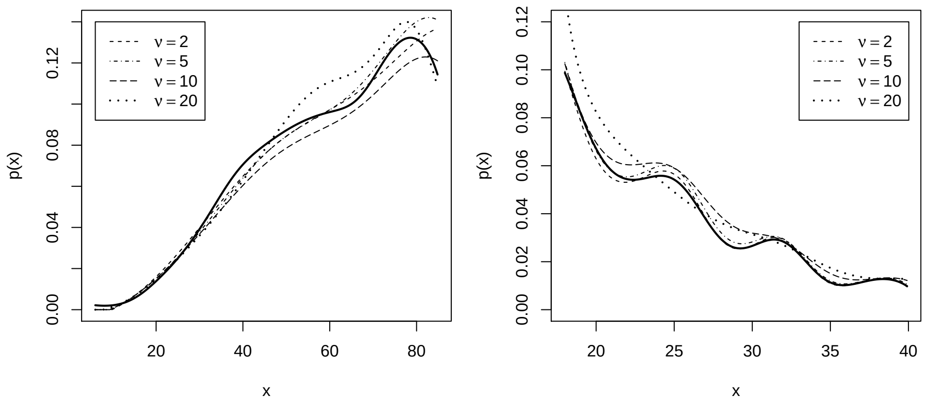

FIG 1. *NHANES study: DH estimator for* $\nu = 2$, 5, 10 *and* 20 *and LL estimator (thick curve) when* $Y = Y_{HBc}$ *(left) or* $Y = Y_{CL}$ *(right).*

As in the simulations, our DH estimator improved considerably on the DM method. An illustration of our procedure with a second covariate is given in section A.3 of the supplemental article (Delaigle and Hall, 2011). In Figure 1 we compare DH with LL. All curves were calculated using our bandwidth procedure described in section A.2 of the supplemental article (Delaigle and Hall, 2011). We see that, in these examples, grouping data in pools of size as large as $\nu = 20$ does not dramatically degrade performance.

5. Generalisations to unequal groups and the multivariate case.

Our procedure for estimating p can be extended to the multivariate setting, where the covariates are random d-vectors, and to unequal group sizes. These extensions can be performed in many different ways, for example by binning on each variable, using bins of potentially different sizes to accommodate different levels of homogeneity. If we group using bins of equal dimension then, to a large extent, the theoretical properties discussed earlier, in the setting of equal-size groups, continue to hold. To briefly indicate this we give, below, details of methodology and results in the case of multivariate histogram binning where, for definiteness, the bin sizes and shapes, but not the group sizes, are equal. Cases where the bin sizes and shapes also vary can be treated in a similar manner, provided the variation is not too great, but since there are so many possibilities we do not treat those cases here. An approach of this type is discussed in section A.3 of the supplemental article (Delaigle and Hall, 2011).

In the analysis below we take $\boldsymbol{X}$ to be a d-vector, and the function p to be d-variate, where $d \geq 1$. We group the data in bins of equal width, specifically width $(\nu/N)^{1/d}$ along each of the d coordinate axes, rather than in groups of equal number. In the theory described below, for notational simplicity we assume that the support of the distribution of $\boldsymbol{X}$ contains the cube $\mathcal{I} = [0,1]^d$, and we estimate p there. We choose ν so that $J = (N/\nu)^{1/d}$ is an integer (on this occasion ν is not necessarily an integer itself), and take the bins to be the cubes $\mathcal{B}(k_1, \ldots, k_d)$ defined by

$$\mathcal{B}(k_1, \ldots, k_d) = \prod_{\ell=1}^{d} \Big(\tfrac{1}{2}\,(2k_\ell + 1)\,(\nu/N)^{1/d} - \tfrac{1}{2}\,(\nu/N)^{1/d}, \\ \tfrac{1}{2}\,(2k_\ell + 1)\,(\nu/N)^{1/d} + \tfrac{1}{2}\,(\nu/N)^{1/d} \Big] ,$$

where $k_\ell = 0, \ldots, J-1$ for $\ell = 1, \ldots, d$. In this setting it is convenient to write the paired data as simply $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_N, Y_N)$, where $\boldsymbol{X}_j$ is a d-vector and each $Y_j = 0$ or 1, and refer to $\boldsymbol{X}_j$ in terms of the bin in which it lies, rather than give it a double subscript (as in the notation $\boldsymbol{X}_{ij}$, where j is the bin index).

Put $b(k_1, \ldots, k_d) = (\frac{1}{2}\,(2k_1 + 1)\,(\nu/N)^{1/d}, \ldots, \frac{1}{2}\,(2k_d + 1)\,(\nu/N)^{1/d})$, representing the centre of the bin $\mathcal{B}(k_1, \ldots, k_d)$, define

$$Z^*(k_1, \ldots, k_d) = 1 - \max_{j\,:\,\boldsymbol{X}_j \in \mathcal{B}(k_1, \ldots, k_d)} Y_j^* ,$$

and compute $\widehat{\mu}$ by applying a d-variate local polynomial smoother to the values of $(b(k_1, \ldots, k_d), Z^*(k_1, \ldots, k_d))$, interpreted as (explanatory variable, response variable) pairs in a conventional d-variate nonparametric regression problem. To derive an estimator of p from $\widehat{\mu}$ we take

$$\widehat{p}(\boldsymbol{x}) = 1 - \widehat{\mu}(\boldsymbol{x})^{1/m(\boldsymbol{x})} , \tag{5.1}$$

where $m(\boldsymbol{x})$ denotes the number of data $\boldsymbol{X}_j$ in the bin containing $\boldsymbol{x} \in \mathcal{I}$.

In developing theoretical properties of this estimator we choose our regularity conditions to simplify exposition. In particular, we replace assumptions (S1)–(S4) and (T5) by the following restriction:

(U) the nonparametric smoother defined by the estimator at (3.2) is a standard d-variate local linear smoother (see e.g. Fan, 1993), where the kernel K, a function of d variables, is a spherically symmetric, compactly supported, Hölder continuous probability density, and, for some $\epsilon > 0$, the bandwidth h satisfies $h + (Nh^d)^{-1} = O(N^{-\epsilon})$ as $N \to \infty$.

Conditions (T1)–(T4) are replaced by (V1)–(V4), below, and (V5) is additional:

Condition V:
(V1) the distribution of $\boldsymbol{X}$ has a continuous density, f, that is bounded away from zero on an open set $\mathcal{J}$ that contains the cube $\mathcal{I} = [0,1]^d$;
(V2) the function $p = \delta\pi$ is bounded below 1 uniformly on $\mathcal{I}$ and in $N \ge 1$;
(V3) the fixed, nonnegative function π has two Hölder-continuous derivatives on $\mathcal{J}$;
(V4) for some $\epsilon > 0$, $h + \nu\delta h + (\nu^2/N^{1-\epsilon} h^d \delta) \to 0$ as $N \to \infty$;
(V5) $C_1\,(\delta N)^4 \le \nu^{d+4} \le C_2\,N^{d+3}/\delta$ for constants $C_1, C_2 > 0$.

THEOREM 5.1. *Assume that conditions U and V hold, and that $\nu\delta = O(1)$. Then, for each $\boldsymbol{x} \in \mathcal{I}$,*

$$\widehat{p}(\boldsymbol{x}) = p(\boldsymbol{x}) + O_p\big\{(\delta/Nh^d)^{1/2} + \delta\,h^2\big\}\,. \tag{5.2}$$

The "O_p" term on the right-hand side of (5.2) has exactly the same size as the dominant remainder term, $A(\boldsymbol{x})\,V(\boldsymbol{x}) + B(\boldsymbol{x})$, on the right-hand side of (3.4) in Theorem 3.1, provided of course that we take $d = 1$ in Theorem 5.1. Refinements given in Theorem 3.1 and in the results in section 3.3 can also be derived in the present setting.

Theorem 5.1 is proved similarly to Theorem 3.1, and so is not derived in detail here. The main difference in the argument comes from incorporating a slightly different definition of $\widehat{p}$, given by (5.1). For example, suppose $\widehat{p}$ is as defined at (5.1), and note that $E(m) = \nu_1 + O\{\nu_1\,(\nu_1/N)^2\}$, where $\nu_1(\boldsymbol{x}) = \nu\,f(\boldsymbol{x})$ and f denotes the density of $\boldsymbol{X}$. Since in addition $m - E(m) = O_p(\nu^{1/2})$, then $m = \nu_1\,(1+\Delta)^{-1}$ where $|\Delta| = O_p\{\nu^{-1/2} + (\nu/N)^2\}$, and, much as in the argument leading to (6.7),

$$\begin{aligned}\widehat{p} &= 1 - \widehat{\mu}^{1/m} = 1 - \big(\widehat{\mu}^{1/\nu_1}\big)^{1+\Delta} \\ &= 1 - \Big[1 - p + O_p\big\{\delta\,h^2 + \big(\delta/Nh^d\big)^{1/2}\big\}\Big]^{1+\Delta} \\ &= 1 - (1-p)\,\Big[1 + O_p\Big\{\delta\,h^2 + \big(\delta/Nh^d\big)^{1/2} + \delta\,|\Delta|\Big\}\Big] \\ &= p + O_p\Big[\delta\,h^2 + \big(\delta/Nh^d\big)^{1/2} + \delta\,\big\{\nu^{-1/2} + (\nu/N)^2\big\}\Big]\,. \end{aligned} \tag{5.3}$$

Now, $\delta\,h^2 + (\delta/Nh^d)^{1/2}$ is minimised by taking $h = (N\delta)^{-1/(d+4)}$, and for this choice of h we have

$$\delta^{-1}\,\big\{\delta\,h^2 + (\delta/Nh^d)^{1/2}\big\} \asymp (\delta N)^{-2/(d+4)}\,.$$

This quantity is not of smaller order than $\nu^{-1/2} + (\nu/N)^2$ if and only if both $\nu^{-1/2} = O\{(\delta N)^{-\rho}\}$ and $(\nu/N)^2 = O\{(\delta N)^{-\rho}\}$, where $\rho = 2/(d+4)$. This is in turn equivalent to

$$C_1\,(\delta N)^{4/(d+4)} \le \nu \le C_2\,\big(N^{d+3}/\delta\big)^{1/(d+4)}\,,$$

for constants $C_1, C_2 > 0$, which is also equivalent to (V5). Therefore if (V5) holds then we can deduce (5.2) from (5.3).

6. Technical arguments.

6.1. *Proof of Theorem 3.1.* Let D_j equal the maximum of $|X_{ij} - \bar{X}_j|$ over $i = 1, \ldots, \nu$. The ratio ν/N equals the order of magnitude of the expected value of the width of the group that contains $x \in \mathcal{I}$, and it can be proved that

for each $\epsilon > 0$, $D_j = O_p(\nu/N^{1-\epsilon})$ uniformly in j such that $|\bar{X}_j - x| \le C\,h$ and $x \in \mathcal{I}$. (6.1)

Note that, by (T4), $\nu/N^{1-\epsilon} \to 0$ for sufficiently small $\epsilon > 0$.

For $k = 1, 2$ let $p^{(k)}$ be the kth derivative of p, and put $p_k = p^{(k)}/\{k!\,(1-p)\}$. Let $\eta > 0$ denote the exponent of Hölder continuity of p'' on $\mathcal{I}$ (see (T3)); that is, $|p''(x_1) - p''(x_2)| = O(|x_1 - x_2|^\eta)$ uniformly in $x_1, x_2 \in \mathcal{I}$. Then, using (6.1) it can be proved that for each $\epsilon > 0$,

$$\begin{aligned}
E(Z_j^* \,|\, \mathcal{X}) &= \prod_{i=1}^{\nu} \{1 - p(X_{ij})\} \\
&= \{1 - p(\bar{X}_j)\}^\nu \prod_{i=1}^{\nu} \big\{1 - p_1(\bar{X}_j)\,(X_{ij} - \bar{X}_j) + O_p\big(\delta\,D_j^2\big)\big\} \\
&= \{1 - p(\bar{X}_j)\}^\nu \prod_{i=1}^{\nu} \exp\big\{-p_1(\bar{X}_j)\,(X_{ij} - \bar{X}_j) + O_p\big(\delta\,D_j^2\big)\big\} \\
&= \{1 - p(\bar{X}_j)\}^\nu \exp\bigg\{-\sum_{i=1}^{\nu} p_1(\bar{X}_j)\,(X_{ij} - \bar{X}_j) + O_p\big(\nu\,\delta\,D_j^2\big)\bigg\} \\
&= \{1 - p(\bar{X}_j)\}^\nu \exp\big\{O_p\big(\nu\,\delta\,D_j^2\big)\big\} \\
&= \{1 - p(\bar{X}_j)\}^\nu \big\{1 + O_p\big(\nu^3\,\delta/N^{2-\epsilon}\big)\big\}\,, \qquad (6.2)
\end{aligned}$$

uniformly in the sense of (6.1) and for each $\epsilon > 0$. (Assumption (T4) implies that $\nu^3\delta/N^{2-\epsilon} \to 0$ for some $\epsilon > 0$.) Observe too that, uniformly in the same sense,

$$\begin{aligned}
\{1 - p(\bar{X}_j)\}^\nu =& \{1 - p(x)\}^\nu \,\big\{1 - p_1(x)\,(\bar{X}_j - x) - p_2(x)\,(\bar{X}_j - x)^2 \\
&+ O_p\big(\delta\,h^{2+\eta}\big)\big\}^\nu
\end{aligned}$$

$$
\begin{aligned}
={}&\{1-p(x)\}^\nu \Big[1-\nu\, p_1(x)\,(\bar X_j-x)+\big\{\tfrac12\,\nu\,(\nu-1)\,p_1(x)^2 \\
&\qquad -\nu\, p_2(x)\big\}\,(\bar X_j-x)^2+O_p\big(\nu\,\delta\, h^{2+\eta}+\nu^3\,\delta^3\,h^3\big)\Big]\,, \quad (6.3)
\end{aligned}
$$

again uniformly in the sense of (6.1). (Note that, by (T4), $\nu\delta h\to0$.) Combining (3.2), (T4), (S1), (S2), (S4), (6.2) and (6.3) we deduce that, for each $\epsilon>0$ and each $x\in\mathcal{I}$,

$$
\begin{aligned}
\tilde\mu(x)\equiv E\{\widehat\mu(x)\,|\,\mathcal{X}\}&=\sum_j w_j(x)\,E(Z_j^*\,|\,\mathcal{X})\Big/\sum_j w_j(x) \\
&=\{1-p(x)\}^\nu\bigg\{1+\Big[\tfrac12\,\nu\,(\nu-1)\,p_1(x)^2-\nu\,p_2(x)\big\}\Big]\,\frac{\sum_j w_j(x)\,(\bar X_j-x)^2}{\sum_j w_j(x)} \\
&\qquad +O_p\big(\nu\,\delta\,h^{2+\eta}+\nu^3\,\delta^3\,h^3+\nu^3\,\delta\,N^{\epsilon-2}\big)\bigg\} \\
&=\{1-p(x)\}^\nu\Big[1+h^2\,\big\{\tfrac12\,\nu\,(\nu-1)\,p_1(x)^2-\nu\,p_2(x)\big\}\,b(x) \\
&\qquad +o_p\big(\nu\,\delta\,h^2+\nu^3\,\delta\,N^{\epsilon-2}\big)\Big]\,,
\end{aligned}
$$

whence, for all $\epsilon>0$,

$$
\tilde\mu(x)^{1/\nu}=\{1-p(x)\}\big[1-h^2\big\{p_2(x)-\tfrac12(\nu-1)p_1(x)^2\big\}\,b(x)+o_p\big(\delta\,h^2+\nu^2\delta\,N^{\epsilon-2}\big)\big]\,, \quad (6.4)
$$

uniformly in $x\in\mathcal{I}$. Hence, defining

$$
\Delta(x)=\widehat\mu(x)-\tilde\mu(x)=\sum_j w_j(x)\,\{Z_j^*-E(Z_j^*\,|\,\mathcal{X})\}\Big/\sum_j w_j(x)\,, \quad (6.5)
$$

noting that $1-p$ is bounded away from zero (see (T2)), and taking the argument of the functions below to equal the specific point x referred to in (3.4), we deduce that:

$$
\begin{aligned}
\widehat p ={}&1-\widehat\mu^{1/\nu}=1-(\tilde\mu+\Delta)^{1/\nu} \\
={}&1-\big(\tilde\mu^{1/\nu}+\nu^{-1}\,\tilde\mu^{-(\nu-1)/\nu}\,\Delta\big)+O_p\big(\nu^{-1}\,\tilde\mu^{-(2\nu-1)/\nu}\,\Delta^2\big) \\
={}&1-(1-p)\,\Big[1-h^2\,\big\{p_2-\tfrac12\,(\nu-1)\,p_1^2\big\}\,b+o_p\big(\delta\,h^2+\nu^2\,\delta\,N^{\epsilon-2}\big)\Big] \\
&\qquad -\nu^{-1}\,\tilde\mu^{-(\nu-1)/\nu}\,\Delta+O_p\big(\nu^{-1}\,\tilde\mu^{-(2\nu-1)/\nu}\,\Delta^2\big) \\
={}&p+(1-p)\,\Big[h^2\,\big\{p_2-\tfrac12\,(\nu-1)\,p_1^2\big\}\,b+o_p\big(\delta\,h^2+\nu^2\,\delta\,N^{\epsilon-2}\big)\Big] \\
&\qquad -\{1+o_p(1)\}\,\nu^{-1}\,(1-p)^{-(\nu-1)}\,\Delta+O_p\big\{\nu^{-1}\,(1-p)^{-(2\nu-1)}\Delta^2\big\} \quad (6.6) \\
={}&p+(1-p)\,\Big[h^2\,\big\{p_2-\tfrac12\,(\nu-1)\,p_1^2\big\}\,b+o_p\big(\delta\,h^2\big)\Big]
\end{aligned}
$$

$$- \{1 + o_p(1)\} \, \nu^{-1} \, (1-p)^{-(\nu-1)} \, \Delta \, , \tag{6.7}$$

where (6.6) holds without the assumption $\nu\delta = O(1)$ (it holds under either that condition or (3.11)), but (6.7) requires $\nu\delta = O(1)$. Note that, by (T4), $\nu/N^{1-\epsilon}h \to 0$ for some $\epsilon > 0$, and so $\nu^2\delta N^{2\epsilon-2}/(\delta h^2) = (\nu/N^{1-\epsilon}h)^2 \to 0$. Additionally, it will follow from (6.9) below that, when $\delta = O(1)$, $\Delta = O_p\{(\nu^2\delta/Nh)^{1/2}\}$, and by (T4), $\delta/Nh \to 0$, so $\Delta = o_p(1)$. The identity leading from (6.6) to (6.7) follows from this property.

Observe that, by (6.2) and (6.3), $E(Z_j^* \,|\, \mathcal{X}) = \{1 + o_p(1)\}\, \{1 - p(x)\}^\nu$ and

$$1 - E(Z_j^* \,|\, \mathcal{X}) = 1 - \{1 - p(x)\}^\nu + O_p\Big[\{1 - p(x)\}^\nu \left(\nu \, \delta \, h + \nu^3 \, \delta \, N^{\epsilon-2}\right)\Big] \, ,$$

uniformly in j such that $|\bar{X}_j - x| \le C\,h$, where C is as in (T5), and moreover,

$$\mathrm{var}(\Delta \,|\, \mathcal{X}) = \frac{\sum_j w_j^2 \, \mathrm{var}(Z_j^* \,|\, \mathcal{X})}{(\sum_j w_j)^2} = \frac{\sum_j w_j^2 \, E(Z_j^* \,|\, \mathcal{X}) \, \{1 - E(Z_j^* \,|\, \mathcal{X})\}}{(\sum_j w_j)^2} \, .$$

(Here and in (6.8)–(6.10) the argument of the functions is the point x in (3.4).) Therefore, by (S3),

$$\begin{aligned} \mathrm{var}(\Delta \,|\, \mathcal{X}) = \{1 + o_p(1)\} \, (\nu/Nh) \, (1-p)^\nu \left\{1 - (1-p)^\nu\right\} v \\ + O_p\big[(\nu/Nh) \left(\nu \, \delta \, h + \nu^3 \, \delta \, N^{\epsilon-2}\right)\big] \, . \end{aligned} \tag{6.8}$$

Properties (T4) and (6.8), and Lyapounov's central limit theorem (see the next paragraph for details), imply that when $\nu\delta = O(1)$ and $\pi(x) > 0$ (the latter is assumed here and below; the proof when $\pi(x) = 0$ is simpler), we can write

$$\begin{aligned} \Delta &= \Big((\nu/Nh)(1-p)^\nu\{1 - (1-p)^\nu\} \, v + O_p\big[(\nu/Nh)\big(\nu\delta h + \nu^3\delta N^{\epsilon-2}\big)\big]\Big)^{1/2} V_4 \\ &= \{1 + o_p(1)\} \, \Big[(\nu/Nh) \, (1-p)^\nu \left\{1 - (1-p)^\nu\right\} v\Big]^{1/2} V_4 \, , \end{aligned} \tag{6.9}$$

where the second identity follows from the fact that $h + \nu^2 N^{\epsilon-2} \to 0$ for some $\epsilon > 0$ (see (T4)), and V_4 denotes a random variable that is asymptotically distributed as normal $\mathrm{N}(0,1)$. This result and (6.7) imply that

$$\begin{aligned} \widehat{p} = p + (1-p) \, \Big[h^2 \left\{p_2 - \tfrac{1}{2} \, (\nu - 1) \, p_1^2\right\} b + o_p\big(\delta \, h^2\big)\Big] \\ - \{1 + o_p(1)\} \, \Big[(\nu Nh)^{-1} \, (1-p)^{2-\nu} \left\{1 - (1-p)^\nu\right\} v\Big]^{1/2} V_4 \, . \end{aligned} \tag{6.10}$$

Result (3.4) follows from (6.10).

When applying a generalised from of Lyapounov's theorem to establish a central limit theorem for Δ, conditional on $\mathcal{X}$, we should, in view of (S4), prove that for some integer $k > 2$, $\big[(\nu/Nh)\,(1-p)^\nu\,\big\{1-(1-p)^\nu\big\}\,v\big]^{-k/2}(\nu/Nh)^{k-1} \to 0$. When $\nu\delta = O(1)$ this is equivalent to $(\delta/Nh)^{-k/2}$ $(\nu/Nh)^{k-1} \to 0$, and hence to $(Nh/\nu)^2\,(\nu^2/Nh\delta)^k \to 0$; call this result (R). Now, (T4) ensures that for some $\epsilon > 0$, $\nu^2/N^{1-\epsilon}h\delta \to 0$. Therefore (R) holds for all sufficiently large k.

Next we outline the derivation of (3.7). It can be proved from (3.3) that if $C_1 > 0$ is given, if $C_2 = C_2(C_1) > 0$ is chosen sufficiently large, if $\mathcal{I}_N$ is a regular grid of n^{C_2} points in $\mathcal{I}$, and if, for each $x \in \mathcal{I}$, we define x_N to be the point in $\mathcal{I}_N$ nearest to x, then

$$P\Big\{\sup_{x\in\mathcal{I}} |\Delta(x) - \Delta(x_N)| \le N^{-C_1}\Big\} \to 1\,. \tag{6.11}$$

Note that, by (T4), Applying (S3), (S4), Rosenthal's and Markov's inequalities, we can prove that, for each $C, \epsilon > 0$, $\sup_{x\in\mathcal{I}} P\big\{|\Delta(x)| > N^\epsilon\,(\nu^2\delta/Nh)^{1/2}$ $\mid \mathcal{X}\big\} = O_p\big(N^{-C}\big)$. It follows that, for all $C, \epsilon > 0$,

$$P\Big\{\sup_{x\in\mathcal{I}_N} |\Delta(x)| > N^\epsilon\,(\nu^2\delta/Nh)^{1/2}\,\Big|\,\mathcal{X}\Big\} = O\big(N^{-C}\big)\,. \tag{6.12}$$

Together (6.11) and (6.12) imply that, for each $C, \epsilon > 0$,

$$P\Big\{\sup_{x\in\mathcal{I}} |\Delta(x)| > N^\epsilon\,(\nu^2\delta/Nh)^{1/2}\Big\} \to 0\,. \tag{6.13}$$

Results (6.4) (which holds uniformly in $x \in \mathcal{I}$) and (6.13) imply that (6.7) holds uniformly in $x \in \mathcal{I}$. Hence,

$$\begin{aligned}\int_{\mathcal{I}} (\widehat{p}-p)^2 = \int_{\mathcal{I}} B^2 + \int_{\mathcal{I}} \big\{\nu^{-1}\,(1-p)^{-(\nu-1)}\,\Delta\big\}^2 \\ - 2\int_{\mathcal{I}} B\,\big\{\nu^{-1}\,(1-p)^{-(\nu-1)}\,\Delta\big\} + o_p\Big\{(\delta\,h^2)^2 + \int_{\mathcal{I}} (\Delta/\nu)^2\Big\}\,.\end{aligned} \tag{6.14}$$

Conditional on $\mathcal{X}$ the random variable Δ, at (6.5), equals a sum of independent random variables with zero means, and using that property, condition S (which, for this part of the theorem, holds uniformly in $x \in \mathcal{I}$) and (3.4) it can be proved that

$$E\bigg[\int_{\mathcal{I}} \big\{\nu^{-1}\,(1-p)^{-(\nu-1)}\,\Delta\big\}^2\,\bigg|\,\mathcal{X}\bigg] = \int_{\mathcal{I}} A^2 + o_p(\delta/Nh)\,, \tag{6.15}$$

$$\operatorname{var}\biggl[\int_{\mathcal{I}} \bigl\{\nu^{-1}\,(1-p)^{-(\nu-1)}\,\Delta\bigr\}^2 \biggm| \mathcal{X}\biggr] = o_p\bigl\{(\delta/Nh)^2\bigr\}\,, \tag{6.16}$$

$$\operatorname{var}\biggl[\int_{\mathcal{I}} B\,\bigl\{\nu^{-1}\,(1-p)^{-(\nu-1)}\,\Delta\bigr\} \biggm| \mathcal{X}\biggr] = o_p\bigl\{(\delta/Nh)^2 + (\delta h^2)^4\bigr\}\,. \tag{6.17}$$

Result (6.15) follows from (6.8). To derive (6.16), note that by (6.4) we have, uniformly in $x_1, x_2 \in \mathcal{I}$,

$$E\bigl\{\Delta(x_1)^2\,\Delta(x_2)^2 \mid \mathcal{X}\bigr\} = E\bigl\{\Delta(x_1)^2 \mid \mathcal{X}\bigr\}\,E\bigl\{\Delta(x_2)^2 \mid \mathcal{X}\bigr\} + O_p\{t_1(x_1,x_2)\}\,, \tag{6.18}$$

where

$$\begin{aligned} t_1(x_1,x_2) &= \frac{\sum_j w_j(x_1)^2\,w_j(x_2)^2\,E\bigl[\{Z_j^* - E(Z_j^*\,|\,\mathcal{X})\}^4 \mid \mathcal{X}\bigr]}{\{\sum_j w_j(x_1)\}^2\,\{\sum_j w_j(x_2)\}^2} = O_p\{t_2(x_1,x_2)\}\,,\\ t_2(x_1,x_2) &= \frac{\sum_j w_j(x_1)^2\,w_j(x_2)^2\operatorname{var}(Z_j^*\,|\,\mathcal{X})}{\{\sum_j w_j(x_1)\}^2\,\{\sum_j w_j(x_2)\}^2}\\ &= O_p\biggl[\frac{\nu\delta\,\sum_j w_j(x_1)^4}{\{\sum_j w_j(x_1)\}^4}\biggr] = O_p\biggl\{\Bigl(\frac{\nu\delta}{Nh}\Bigr)^2\Bigl(\frac{\nu^2}{Nh\delta}\Bigr)\biggr\} = o_p\biggl\{\Bigl(\frac{\nu\delta}{Nh}\Bigr)^2\biggr\}\,, \end{aligned}$$

again uniformly in $x_1, x_2 \in \mathcal{I}$. (The last and second-last identities here follow from (T4) and (S4) respectively.) Noting these bounds, defining $\xi_1 \equiv \{\nu^{-1}\,(1-p)^{-(\nu-1)}\}^2$, and integrating (6.18) over $x_1, x_2 \in \mathcal{I}$, we deduce that

$$E\biggl\{\int \xi_1(x)\,\Delta(x)^2\,dx \biggm| \mathcal{X}\biggr\}^2 = \biggl[\int \xi_1(x)\,E\bigl\{\Delta(x)^2\,|\,\mathcal{X}\bigr\}\,dx\biggr]^2 + o_p\bigl\{(\nu\delta/Nh)^2\bigr\}\,,$$

which implies (6.16).

To derive (6.17), define $\xi_2 = B\,\xi_1$ and $e_j = E[\{Z_j^* - E(Z_j^*\,|\,\mathcal{X})\}^2\,|\,\mathcal{X}]$, write M for the left-hand side of (6.17), and note that

$$M = \int_{\mathcal{I}}\!\int_{\mathcal{I}} \xi_2(x_1)\,\xi_2(x_2)\;\frac{\sum_j w_j(x_1)\,w_j(x_2)\,e_j}{\{\sum_j w_j(x_1)\}\,\{\sum_j w_j(x_2)\}}\,dx_1\,dx_2\,.$$

In view of (T5), $w_j(x) = 0$ if $|\bar{X}_j - x| > C\,h$, and so the series in the numerator inside the integrand can be confined to indices j for which both $|\bar{X}_j - x_1| \le C\,h$ and $|\bar{X}_j - x_2| \le C\,h$. Therefore the integrand equals zero unless $|x_1 - x_2| \le 2\,C\,h$. Hence, defining $J(x_1,x_2) = 1$ if $|x_1 - x_2| \le 2\,C\,h$, and $J(x_1,x_2) = 0$ otherwise; using the Cauchy-Schwarz inequality to derive both the inequalities below; and writing $\|\mathcal{I}\|$ for the length of the interval $\mathcal{I}$; we have:

$$M \le \int_{\mathcal{I}}\!\int_{\mathcal{I}} J(x_1,x_2)\xi_2(x_1)\xi_2(x_2)\biggl[\prod_{k=1}^{2}\sum_j w_j(x_k)^2 e_j/\{\sum_j w_j(x_k)\}^2\biggr]^{1/2} dx_1\,dx_2$$

$$
\begin{aligned}
&= \int_{\mathcal{I}}\int_{\mathcal{I}} J(x_1,x_2)\,\xi_2(x_1)\,\xi_2(x_2)\left[\prod_{k=1}^{2}\operatorname{var}\{\Delta(x_k)\,|\,\mathcal{X}\}\right]^{1/2} dx_1\,dx_2 \\
&\le \|\mathcal{I}\|\left(\int_{\mathcal{I}}\int_{\mathcal{I}} J(x_1,x_2)\left[\prod_{k=1}^{2}\xi_2(x_k)^2\operatorname{var}\{\Delta(x_k)\,|\,\mathcal{X}\}\right] dx_1\,dx_2\right)^{1/2}. \quad (6.19)
\end{aligned}
$$

Using (6.8) show that $\operatorname{var}\{\Delta(x_k)\,|\,\mathcal{X}\} = O(\nu^2\delta/Nh)$, uniformly in $x_k \in \mathcal{I}$; noting that $B = O(\delta\,h^2)$ uniformly in $x \in \mathcal{I}$ (the bound at (3.7) holds uniformly in the argument of B); and observing that $\xi_1(x) = O(\nu^{-2})$ uniformly in $x \in \mathcal{I}$, whence it follows from the bound for B that $\xi_2(x) = O(\delta\,h^2\,\nu^{-2})$ uniformly in $x \in \mathcal{I}$; we deduce from (6.19) that

$$
\begin{aligned}
M &= O_p\Big[\big(\nu^2\delta/Nh\big)\,\big\{\big(\delta\,h^2\,\nu^{-2}\big)\big\}^2\Big]\left(\int_{\mathcal{I}}\int_{\mathcal{I}} J(x_1,x_2)\,dx_1\,dx_2\right)^{1/2} \\
&= O_p\big(\delta^2\,h^{7/2}/N\big) = o_p\big\{(\delta/Nh)^2 + (\delta h^2)^4\big\}\,. \quad (6.20)
\end{aligned}
$$

Result (6.17) follows directly from (6.20).

6.2. *Proof of Theorem 3.2.* The proof is similar to that of the first part of Theorem 3.1, the main difference occurring at the point at which the remainder term, $O_p(R)$ where $R = \nu^{-1}\,(1-p)^{-(2\nu-1)}\,\Delta^2$, in (6.6), is shown to be negligible relative to the term $\nu^{-1}\,(1-p)^{-(\nu-1)}\,\Delta$ there. It suffices to prove that $(1-p)^{-\nu}\,\Delta \to 0$ in probability, or equivalently, in view of (6.9), that $(\nu/Nh)\,(1-p)^{-\nu} \to 0$. However, the latter result is ensured by (3.11).

6.3. *Proof of Theorem 3.3.* Without loss of generality the point x in (3.16) is $x = 0$. Recall that $p^0 \equiv \delta$, and take $p^1(u) = \delta\,\{1 + h^2\,\psi(u/h)\}$, where ψ is bounded and has two bounded derivatives on the real line, is supported on $[-\frac{1}{2}, \frac{1}{2}]$ and satisfies $\psi(0) \neq 0$. The respective functions $\pi^0 \equiv 1$ and $\pi^1(u) = 1 + h^2\,\psi(u/h)$ satisfy (3.14). (The quantity $h = h(N) > 0$ here is not a bandwidth, but converges to 0 as $N \to \infty$.) Therefore, $p^0(u) = p^1(u)$ except when $u \in (-\frac{1}{2}\,h, \frac{1}{2}\,h)$. We assume that $\nu\delta \to \infty$ as $N \to \infty$, and consider the problem of discriminating between p^0 and p^1 using the data pairs $(\mathcal{X}_j, Y_j^*)$.

Without loss of generality we confine attention to those pairs $(\mathcal{X}_j, Y_j^*)$ for which $\mathcal{X}_j$ is wholly contained in $[-\frac{1}{2}\,h, \frac{1}{2}\,h]$. Pairs for which $\mathcal{X}_j$ has no intersection with $[-\frac{1}{2}\,h, \frac{1}{2}\,h]$ convey no information for discriminating between p^0 and p^1, and it is readily proved that including pairs for which $\mathcal{X}_j$ overlaps the boundary does not affect the results we derive below. In a slight abuse of notation we shall take the integers j for which $\mathcal{X}_j \subseteq [-\frac{1}{2}\,h, \frac{1}{2}\,h]$ to be $1, \ldots, m$, where $m = hN/\nu + o_P(1)$ and is assumed to be an integer.

The likelihood of the data pairs $(\mathcal{X}_j, Y_j^*)$ for $1 \leq j \leq m$, conditional on $\mathcal{X} = \{X_1, \ldots, X_N\}$, is $\prod_{j=1}^m P_j^{Y_j^*}\,(1-P_j)^{1-Y_j^*}$ where $P_j = P(Y_j^* = 1 \,|\, \mathcal{X}) = 1 - \prod_{i=1}^{\nu} \{1 - p(X_{ij})\}$. Let P_j^0 and P_j^1 denote the versions of P_j when $p = p^0$ and $p = p^1$, respectively. Also, let $\Theta_j^+ = P_j^1/P_j^0$ and $\Theta_j^- = (1-P_j^1)/(1-P_j^0)$. In this notation the log-likelihood ratio statistic is given by

$$L = \sum_{j=1}^m \big\{Y_j^* \, \log(\Theta_j^+) + (1 - Y_j^*)\, \log(\Theta_j^-)\big\}$$

$$= \sum_{j=1}^m (1 - Y_j^*)\, \log(\Theta_j^-/\Theta_j^+) + \sum_{j=1}^m \log(\Theta_j^+)\,, \tag{6.21}$$

and therefore, $E(L \,|\, \mathcal{X}) = \sum_{j=1}^m (1 - P_j)\, \log(\Theta_j^-/\Theta_j^+) + \sum_{j=1}^m \log(\Theta_j^+)$, $\mathrm{var}(L \,|\, \mathcal{X}) = \sum_{j=1}^m P_j\,(1-P_j)\,\big\}\log(\Theta_j^-/\Theta_j^+)\big\}^2$. Writing E^0 and var^0 to denote expectation and variance when $p = p^0$, we deduce that

$$E^0(L \,|\, \mathcal{X}) = (1-\delta)^{\nu} \sum_{j=1}^m \log(\Theta_j^-/\Theta_j^+) + \sum_{j=1}^m \log(\Theta_j^+)\,, \tag{6.22}$$

$$\mathrm{var}^0(L \,|\, \mathcal{X}) = (1-\delta)^{\nu} \big\{1 - (1-\delta)^{\nu}\big\} \sum_{j=1}^m \big\{\log(\Theta_j^-/\Theta_j^+)\big\}^2 . \tag{6.23}$$

Assume for the time being that

$$\nu \delta h^2 \to 0 \tag{6.24}$$

as $N \to \infty$, and observe that, since $1 - P_j^0 = (1-\delta)^{\nu}$, then

$$\Theta_j^- = \big(1 - P_j^0\big)^{-1} \prod_{i=1}^{\nu} \big[1 - \delta\,\big\{1 + h^2\,\psi(X_{ij}/h)\big\}\big]$$

$$= \prod_{i=1}^{\nu} \bigg\{1 - \frac{\delta}{1-\delta}\, h^2\, \psi(X_{ij}/h)\bigg\} = 1 - \rho\, h^2\, S_j + R_j\,, \tag{6.25}$$

where $\rho = \delta/(1-\delta)$, $S_j = \sum_i \psi(X_{ij}/h)$ and $R_j = O_p(\nu\rho^2 h^4)$ uniformly in $1 \leq j \leq m$. (We used (6.24) to derive the last identity in (6.25). To obtain uniformity in the bound for R_j, and in later bounds, we used the fact that ψ is bounded.) Hence,

$$\log(\Theta_j^-) = -\Big\{\rho\, h^2\, S_j - R_j + \tfrac{1}{2}\,\big(\rho\, h^2\, S_j - R_j\big)^2 + \tfrac{1}{3}\,\big(\rho\, h^2\, S_j - R_j\big)^3 - \ldots\Big\}\,.$$

Similarly, since

$$\begin{aligned} P_j^1 &= 1 - \big(1 - P_j^1\big) = 1 - \big(1 - P_j^0\big) \prod_{i=1}^{\nu} \bigg\{1 - \frac{\delta}{1-\delta}\, h^2\, \psi(X_{ij}/h)\bigg\} \\ &= 1 - \big(1 - P_j^0\big)\, \big(1 - \rho\, h^2\, S_j + R_j\big) = P_j^0 + \big(1 - P_j^0\big)\, \big(\rho\, h^2\, S_j - R_j\big)\,, \end{aligned}$$

then

$$\begin{aligned} \log(\Theta_j^+) &= \log\big\{1 + \big(\rho\, h^2\, S_j - R_j\big)(1 - P_j^0)/P_j^0\big\} \\ &= \frac{(1-\delta)^\nu}{1-(1-\delta)^\nu}\, \big(\rho\, h^2\, S_j - R_j\big) - \tfrac{1}{2}\, (1-\delta)^{2\nu}\, \big(\rho\, h^2\, S_j - R_j\big)^2 \\ &\quad + O_p\Big\{(1-\delta)^{3\nu}\, \big(\nu\, \rho\, h^2\big)^2\Big\}\,, \end{aligned}$$

uniformly in $1 \le j \le m$. It follows that

$$\begin{aligned} \log(\Theta_j^-/\Theta_j^+) &= -\,\Big\{\big(\rho\, h^2\, S_j - R_j\big) + \tfrac{1}{2}\, \big(\rho\, h^2\, S_j - R_j\big)^2 + \tfrac{1}{3}\, \big(\rho\, h^2\, S_j - R_j\big)^3 \\ &\qquad + \ldots\Big\} - \frac{(1-\delta)^\nu}{1-(1-\delta)^\nu}\, \big(\rho\, h^2\, S_j - R_j\big) + \tfrac{1}{2}\, (1-\delta)^{2\nu}\, \big(\rho\, h^2\, S_j \\ &\qquad - R_j\big)^2 + O_p\Big\{(1-\delta)^{3\nu}\, \big(\nu\, \rho\, h^2\big)^2\Big\} \\ &= -\, \rho\, h^2\, S_j + O_p\big\{\nu\, \rho^2\, h^4 + (1-\delta)^\nu\, \nu\, \rho\, h^2\big\}\,, \end{aligned} \tag{6.26}$$

$$\begin{aligned} &(1-\delta)^\nu\, \log(\Theta_j^-/\Theta_j^+) + \log(\Theta_j^+) \\ &\quad = -(1-\delta)^\nu\, \Big\{\big(\rho\, h^2\, S_j - R_j\big) + \tfrac{1}{2}\, \big(\rho\, h^2\, S_j - R_j\big)^2 + \tfrac{1}{3}\, \big(\rho\, h^2\, S_j - R_j\big)^3 + \ldots\Big\} \\ &\qquad - \frac{(1-\delta)^{2\nu}}{1-(1-\delta)^\nu}\, \big(\rho\, h^2\, S_j - R_j\big) \\ &\qquad + \frac{(1-\delta)^\nu}{1-(1-\delta)^\nu}\, \big(\rho\, h^2\, S_j - R_j\big) - \tfrac{1}{2}\, (1-\delta)^{2\nu}\, \big(\rho\, h^2\, S_j - R_j\big)^2 \\ &\qquad + O_p\Big\{(1-\delta)^{3\nu}\, \big(\nu\, \rho\, h^2\big)^2\Big\} \\ &\quad = -(1-\delta)^\nu\, \Big\{\tfrac{1}{2}\, \big(\rho\, h^2\, S_j - R_j\big)^2 + \tfrac{1}{3}\, \big(\rho\, h^2\, S_j - R_j\big)^3 + \ldots\Big\} \\ &\qquad - \tfrac{1}{2}\, (1-\delta)^{2\nu}\, \big(\rho\, h^2\, S_j - R_j\big)^2 + O_p\Big\{(1-\delta)^{3\nu}\, \big(\nu\, \rho\, h^2\big)^2\Big\} \\ &\quad = -\tfrac{1}{2}\, (1-\delta)^\nu\, \big(\rho\, h^2\, S_j\big)^2 + o_p\Big\{(1-\delta)^\nu\, \big(\nu\, \rho\, h^2\big)^2\Big\}\,, \end{aligned} \tag{6.27}$$

uniformly in $1 \le j \le m$. Using (6.22), (6.23), (6.26) and (6.27) we deduce that:

$$E^0(L \,|\, \mathcal{X}) = -\tfrac{1}{2}\,(1-\delta)^\nu\,\big(\rho\,h^2\big)^2 \sum_{j=1}^m S_j^2 + o_p\Big\{m\,(1-\delta)^\nu\,\big(\nu\,\rho\,h^2\big)^2\Big\}\,,$$

$$\mathrm{var}^0(L \,|\, \mathcal{X}) = \{1+o_p(1)\}\,(1-\delta)^\nu\,\big(\rho\,h^2\big)^2 \sum_{j=1}^m S_j^2 + o_p\Big\{m\,(1-\delta)^\nu\,\big(\nu\,\rho\,h^2\big)^2\Big\}\,.$$

Choose h so that

$$\text{the squared mean, and the variance, are of the same order}\,. \tag{6.28}$$

In particular, take $\big\{m\,(1-\delta)^\nu\,\big(\nu\,\rho\,h^2\big)^2\big\}^2 = C_1\,m\,(1-\delta)^\nu\,\big(\nu\,\rho\,h^2\big)^2$, and hence

$$m\,(1-\delta)^\nu\,\big(\nu\,\rho\,h^2\big)^2 = C_2 + o_P(1)\,, \tag{6.29}$$

or equivalently, using the fact that $m = Nh/\nu + o_P(1)$,

$$h = C_3\,\big\{(N\,\nu\,\rho^2)^{-1}\,(1-\delta)^{-\nu}\big\}^{1/5}\,, \tag{6.30}$$

where C_1, C_2, C_3 are positive constants; C_3 can be chosen arbitrarily. It follows that

$$\rho h^2 = C_3^2\,\big(\rho/N^2\big)^{1/5}\,\nu^{-2/5}\,\lambda_N^2\,, \tag{6.31}$$

where $\lambda_N^5 = (1-\delta)^{-\nu}$. If h is given by (6.31) then $\nu\rho h^2 = C_3^2\,(\nu^3\rho/N^2)^{1/5}\,\lambda_N^2$ and therefore (6.24) follows from (3.15).

It can be shown that, conditional on the explanatory variables, the log-likelihood ratio L, centred at the conditional mean and variance, is asymptotically normally distributed with zero mean and unit variance. (We shall give a proof below.) Therefore by taking C_1, and hence C_3, sufficiently small, we can ensure that: (i) The probability of discriminating between p^0 and p^1, when $p = p^0$, is bounded below 1 as $N \to \infty$. (This follows from (6.28).) Similarly it can be proved that: (ii) The probability of discriminating between p^0 and p^1, when $p = p^1$, is bounded below 1. Consider the assertion: (iii) $\check{p}(0) - p(0)$ converges in probability to 0, along a subsequence, at a strictly faster rate than h^2. If (iii) is true then the error rate of the classifier which asserts that $p = p^0$ if $\check{p}(0)$ is closer to $p(0)$ than to $p^1(0)$, and $p = p^1$ otherwise, and converges to 0 as $N \to \infty$. However, properties (i) and (ii) show that even the optimal classifier, based on the likelihood ratio rule, does not enjoy this degree of accuracy, and so (iii) must be false. This proves (3.16).

Finally we derive the asymptotic normality of L claimed in the previous paragraph. We do this using Lindeberg's central limit theorem, as follows. In view of the definition of L at (6.21) it is enough to prove that for each $\eta > 0$,

$$S_{N1}(\eta) \equiv \sigma(\mathcal{X})^{-2} \sum_{j=1}^{m} E^0\Big[\big|Y_j^* - E(Y_j^* \,|\, \mathcal{X})\big|^2 \sigma_j(\mathcal{X})^2 \\ \times I\Big\{\big|Y_j^* - E(Y_j^* \,|\, \mathcal{X})\big|\, \sigma_j(\mathcal{X}) > \eta\, \sigma(\mathcal{X})\Big\} \;\Big|\; \mathcal{X}\Big] \to 0 \tag{6.32}$$

in probability, where we define

$$\sigma_j(\mathcal{X})^2 = \big\{\log(\Theta_j^-/\Theta_j^+)\big\}^2 = \big(\rho h^2 S_j\big)^2 + o_p\big\{(\nu\delta h^2)^2\big\}, \tag{6.33}$$

$$\sigma(\mathcal{X})^2 = \sum_{j=1}^{m} \operatorname{var}^0(Y_j^* \,|\, \mathcal{X})\, \sigma_j(\mathcal{X})^2 = \{1 + o_p(1)\}\, C_4\, m\, \big(\nu\, \rho\, h^2\big)^2 (1-\delta)^\nu\,, \tag{6.34}$$

with $C_4 > 0$ and (6.33) holding uniformly in j. (We used (6.26) to obtain the second identities in each of (6.33) and (6.34).) Since $m = hN/\nu + o_P(1)$ then, by (6.30) and (6.34), $\sigma(\mathcal{X})^2 \to C_5$ in probability, where $C_5 > 0$. Hence, by (6.32), with probability converging to 1 as $N \to \infty$,

$$C_6\, S_{N1}(\eta) \le S_{N2}(\eta) \equiv \sum_{j=1}^{m} E^0\Big[\big|Y_j^* - E(Y_j^* \,|\, \mathcal{X})\big|^2 \sigma_j(\mathcal{X})^2 \\ \times I\Big\{\big|Y_j^* - E(Y_j^* \,|\, \mathcal{X})\big|\, \sigma_j(\mathcal{X}) > C_7\Big\} \;\Big|\; \mathcal{X}\Big],$$

where $C_6, C_7 > 0$ are constants and C_7 depends on η.

Note too that, using (6.30) to obtain the second relation below, and (3.15) to get the last relation, we have: $\big(\nu\, \rho\, h^2\big)^5 \asymp \big(\nu\delta h^2\big)^5 \asymp \big\{\big(\nu^3\, \delta\big)^{1/2} N^{-1} (1 - \delta)^{-\nu}\big\}^2 \to 0$. Therefore, (6.24) holds. Since $|Y_j^* - E(Y_j^* \,|\, \mathcal{X})| \le 1$ then, if $\sigma_j(\mathcal{X}) \le C_7$, we have $I\{|Y_j^* - E(Y_j^* \,|\, \mathcal{X})|\, \sigma_j(\mathcal{X}) > C_7\} = 0$. Hence, using (6.26) and (6.24),

$$\begin{aligned} S_{N2}(\eta) &\le \sum_{j=1}^{m} E^0\Big\{\big|Y_j^* - E(Y_j^* \,|\, \mathcal{X})\big|^2 \;\Big|\; \mathcal{X}\Big\}\, \sigma_j(\mathcal{X})^2\, I\{\sigma_j(\mathcal{X}) > C_7\} \\ &= (1-\delta)^\nu \left\{1 - (1-\delta)^\nu\right\} \sum_{j=1}^{m} \sigma_j(\mathcal{X})^2\, I\{\sigma_j(\mathcal{X}) > C_7\} \\ &\le (1-\delta)^\nu\, C_7^{-2} \sum_{j=1}^{m} \sigma_j(\mathcal{X})^4 = O_p\big\{(1-\delta)^\nu\, m\, (\nu\, \delta\, h^2)^4\big\} \end{aligned}$$

$$= o_p\Big\{m\,(1-\delta)^\nu\,\big(\nu\,\delta\,h^2\big)^2\Big\} = o_p(1)\,,$$

since $m\,(1-\delta)^\nu\,(\nu\delta h^2)^2 = C_2$; see (6.29). This completes the proof of (6.32).

Acknowledgements. We thank three referees and an associate editor for their helpful comments which led to an improved version of the manuscript

SUPPLEMENTARY MATERIAL

Supplement A: additional material
(Provided in a separate file). The supplementary article contains a description of Delaigle and Meister's method, details for bandwidth choice, an alternative procedure for multivariate setting and unequal groups, and additional numerical results.

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.

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

Delaigle, A. and Hall, P. (2011). Supplement to "Nonparametric regression with homogeneous group testing data".

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. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196–216.

Fan, J., and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman & Hall, London.

Fan, J., Heckman, N. and Wand, M. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90**, 141–150.

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.

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.

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., Wand, M. P. and Carroll, R.J. (2003). *Semiparametric Regression*, Cambridge University Press.

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.

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

Department of Mathematics and Statistics, University of Melbourne, Parkville, VIC, 3010, Australia.
E-mail: A.Delaigle@ms.unimelb.edu.au
halpstat@ms.unimelb.edu.au