Biometrika (2014), **xx**, x, *pp.* 1–19

Printed in Great Britain

New approaches to non- and semi-parametric regression for univariate and multivariate group testing data

By A. Delaigle, P. Hall

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

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

J.R. Wishart

Department of Mathematics and Statistics, University of New South Wales, Kensington, New South Wales 2052, Australia.

j.wishart@unsw.edu.au

Summary

We consider non- and semi-parametric estimation of a conditional probability curve in the case of group testing data, where the individuals are pooled randomly into groups, and only the pooled data are available. We derive a nonparametric weighted estimator that has optimality properties accounting for group sizes, and show how to extend it to multivariate settings, including the partially linear model. In the group testing context, it is natural to assume that the probability curve depends on the covariates only through a linear combination of them. Motivated by this, we develop a nonparametric estimator based on the single-index model. We study theoretical properties of the suggested estimators, and derive data-driven procedures. Practical properties of the methods are demonstrated via real and simulated examples and shown to have smaller median integrated square error than existing competitors.

Some key words: Bandwidth; local polynomial regression; multivariate kernel estimator; partially linear model, single-index model, weighted estimator.

1. Introduction

Group testing (Dorfman, 1943) is a method employed when collecting data on a Bernoulli variable Y, where, instead of observing the value of Y for each individual in a sample, the individuals are pooled in J groups of sizes $n_1, \ldots, n_J$; and only the maximum of the Y-values of the individuals within each group is observed. More specifically, let Y_{ij} denote the value of Y for the ith individual in the jth group. In the group testing setting, instead of observing Y_{ij} $(i = 1, \ldots, n_j;\ j = 1, \ldots, J)$, we observe

$$Y_j^* = \max_{i=1,\ldots,n_j} Y_{ij} \quad (j = 1, \ldots, J)\,. \tag{1}$$

This technique was originally introduced in infectious disease studies, to reduce the cost and increase the speed of data collection. Often, Y is the result of a blood or urine test, typically a test for an infectious disease, and $Y = I(\text{test is positive})$, where $I(\mathcal{E})$

denotes the indicator function of an event $\mathcal{E}$; that is, $I(\mathcal{E}) = 1$ if $\mathcal{E}$ is true, and $I(\mathcal{E}) = 0$ otherwise. There, $Y_j^* = 1$ if one or more individuals within the jth group test positive. Of course, individual here should be interpreted in a broad sense. In particular, it does not necessarily have to refer to a person, and could for example be an animal, or a water or milk sample tested for pollution (Nagi & Raggi, 1972; Wahed et al., 2006; Lennon, 2007; Fahey et al., 2006). As described by Hepworth (2005), the topic of group testing includes applied studies of plant disease (Fletcher et al., 1999), fisheries (Worlund & Taylor, 1983), and spread of disease by insects (Swallow, 1985).

In those studies it is common also to observe one or several covariates X, in which case it is of interest to estimate the probability of being contaminated, for example by an infectious disease, given X, that is $p(x) = \mathrm{pr}(Y = 1 \mid X = x)$. Unlike Y, X is often observed for each individual. In the parametric context, this problem has been studied by, for example, Vansteelandt et al. (2000) and Xie (2001). Delaigle & Meister (2011) suggested a consistent nonparametric estimator of p, which does not exploit fully properties of unequal group sizes. See Delaigle & Hall (2012) for the particular context where the data are grouped homogeneously, and also Gastwirth & Hammick (1989); Chen & Swallow (1990); Farrington (1992); Gastwirth & Johnson (1994); Hardwick et al. (1998); Hung & Swallow (2000); Bilder & Tebbs (2009); Chen et al. (2009); Huang (2009); Huang & Tebbs (2009); Li & Xie (2012) and Wang et al. (2013) for related work.

While the univariate, nonparametric estimator of Delaigle & Meister (2011) performs well when all groups have the same size, it does not account fully for the fact that groups of unequal size do not contain the same amount of information about p. As a result, when groups are of unequal size, the corresponding estimators suffer from excessive variance. We suggest a new, nonparametric estimator which addresses this difficulty through adaptive weights, and allows for discrete covariates determined by a partially linear model.

We generalise our method to the multivariate context and also derive a single-index version of our estimator. For the latter, we use ideas proposed in the standard regression context by Härdle et al. (1993), although the grouped nature of our data makes the adaptation to our setting highly nontrivial, in both the development of the method and the derivation of its theoretical aspects. We establish asymptotic properties of our estimators, and propose automatic, data-driven procedures for choosing the smoothing parameters in practice. We extend our methodology to cases where the tests are imperfect.

2. Methodology when a single covariate is modelled nonparametrically

2·1. *Local polynomial estimator and Delaigle & Meister's (2011) estimator*

In the univariate case considered by Delaigle & Meister (2011), we observe a sample (X_{ij}, Y_j^*) $(i = 1, \ldots, n_j;\ j = 1, \ldots, J)$, with Y_j^* as in (1). Here $Y_{ij} \mid X_{ij} = x \sim \mathrm{Be}\{p(x)\}$; that is, $p(x) = \mathrm{pr}(Y_{ij} = 1 \mid X_{ij} = x) = 1 - \mathrm{pr}(Y_{ij} = 0 \mid X_{ij} = x)$, the Y_{ij}s are independent, and the X_{ij}s are independent and identically distributed random variables. The goal is to construct a consistent nonparametric estimator of p based on the grouped testing data (X_{ij}, Y_j^*) $(i = 1, \ldots, n_j;\ j = 1, \ldots, J)$.

Before discussing estimators for such data, it is useful to address the simpler problem of estimating p nonparametrically when the Y_{ij}s are available. Since $p(x) = E(Y_{ij} \mid X_{ij} = x)$, p is a regression curve and can be estimated by the standard ℓth order local polynomial estimator, constructed as follows. First, approximate p locally around x by an ℓth degree

polynomial $p_\ell(z) = \sum_{0\le k\le \ell} \alpha_{k,x}\,(x-z)^k$; and then, fit the local coefficients $\alpha_{k,x}$ by minimising the locally weighted least squares sum, $\sum_{j\le J}\sum_{i\le n_j}\{Y_{ij} - p_\ell(X_{ij})\}^2\,K_h(X_{ij} - x)$, where K is a kernel function, $h>0$ is a bandwidth, and $K_h(x) = h^{-1}\,K(x/h)$. For $k = 0,\ldots,\ell$, let $\hat\alpha_{k,x}$ be the resulting estimator of $\alpha_{k,x}$. The ℓth order local polynomial estimator of $p(x)$ is defined by $\hat p(x) = \hat\alpha_{0,x}$. See Fan & Gijbels (1996, p. 19).

In the group testing data context introduced above, Delaigle & Meister (2011) suggested the first nonparametric estimator of p. Their approach consists in constructing the standard local polynomial estimator based on the group testing data, and converting this naive estimator into a consistent estimator of p through a correction factor. Although their estimator has good properties when the group sizes n_j are equal, it assigns the same weight to each observation Y_j^*, regardless of the size of the group it comes from. In many applications, the groups cannot be taken of equal size. See, e.g., Hepworth (2005) for a study involving plant viruses, where n_j varies from 1 to 25. In such cases the quality of the information contained in Y_j^* depends heavily on the group size n_j. By not taking this into account, the estimator of Delaigle & Meister (2011) estimator can suffer from a large variance.

To overcome this difficulty we suggest a new nonparametric estimator of p, which we construct so as to optimise asymptotic theoretical properties, whether the n_js are equal or not. We use this estimator to construct a rescaled estimator which satisfies a centralised bias property. Instead of treating only this simple univariate context, in Section 2·2 we introduce our method in a more general partially linear model, which enables us to include discrete explanatory variables. The case of multiple continuous covariates will be treated in Section 3. See also Section 7 for the case of imperfect tests.

2·2. *Partially linear model*

Suppose we observe a continuous variable and one or more discrete covariates, such as gender. Specifically, we observe pairs (X_{ij}, Y_j^*), with Y_j^* as in (1), and where $X_{ij} = (U_{ij}, V_{ij}^{\mathrm T})^{\mathrm T}$ is a d-dimensional vector, with $d\ge 1$. Here, U_{ij} is a continuous variable and, when $d\ge 2$, $V_{ij}\in\mathbb{R}^{d-1}$ is a discrete variable or vector. When $d=1$, we observe only a continuous variable $X = U$, as in Delaigle & Meister (2011).

We model the continuous and discrete parts simultaneously through a partially linear model. There it is assumed that

$$p(X_{ij}) = g(U_{ij}) + \gamma^{\mathrm T} V_{ij} \quad (j=1,\ldots,J;\, i=1,\ldots,n_j)\,, \tag{2}$$

where g is an unknown function and $\gamma\in\mathbb{R}^{d-1}$ is an unknown parameter. We let $N = \sum_{j\le J} n_j$, $Z_j^* = 1 - Y_j^*$, $m = 1-g$ and $q_0 = 1 - E\{p(X_{ij})\}$.

In the standard non-grouped data case, methods have been developed in the literature for estimating g and γ. See, for example, Speckman (1988) and Härdle et al. (2000). In the group testing context,

$$E(Z_j^* \mid X_{1j},\ldots,X_{n_j j}) = \prod_{i=1}^{n_j}\left\{m(U_{ij}) - \gamma^{\mathrm T} V_{ij}\right\}, \tag{3}$$

$$E(q_0^{1-n_j} Z_j^* \mid X_{ij}) = m(U_{ij}) - \gamma^{\mathrm T} V_{ij}\,. \tag{4}$$

As in the standard case, below we suggest estimating m and γ in two steps.

To estimate m, assume temporarily that γ and q_0 are known, and let $T_{ij}^* = Z_j^* + q_0^{n_j-1}\,\gamma^{\mathrm T} V_{ij}$. It follows from (4) that $m(u) = E(q_0^{1-n_j}\,T_{ij}^* \mid U_{ij} = u)$. Borrowing techniques

from the local polynomial estimator introduced in Section 2·1, this equation can be used to construct an ℓth order local polynomial estimator of m, with $\ell \geq 0$, as follows. First, approximate $m(z)$, for z in a neighbourhood of u, by the ℓth order polynomial $m_\ell(z) = \sum_{0\leq k\leq \ell} \alpha_{k,u}\,(z-u)^k$. Then, at each u, estimate the coefficients $\alpha_{k,u}$ by minimising a locally weighted least squares sum. Using the standard approach discussed in Section 2·1, the ith individual from the jth group would be assigned a weight $K_h(U_{ij}-u)$. We use different weights in our case, since groups of unequal size do not contain information of the same quality. Motivated by the relation $m(u) = E(q_0^{1-n_j}\,T_{ij}^* \mid U_{ij} = u)$, we suggest estimating $\alpha_u = (\alpha_{0,u},\ldots,\alpha_{\ell,u})^{\mathrm{T}}$ by

$$(\hat\alpha_{0,u},\ldots,\hat\alpha_{\ell,u})^{\mathrm{T}} = \operatorname*{argmin}_{\alpha_u} \sum_{j=1}^{J}\sum_{i=1}^{n_j} \left\{q_0^{1-n_j}\,T_{ij}^* - m_\ell(U_{ij})\right\}^2 K_h(U_{ij}-u)\,q_0^{n_j-1}\,\psi_j(q_0)\,,$$

where $\psi_1,\ldots,\psi_J$ are smooth, positive weight functions defined on $[0,1]$. We shall show later how to choose the ψ_js to optimise properties of our estimator; see Section 5·1.

As in the standard case, discussed in Section 2·1, for γ and q_0 known we define the ℓth order local polynomial estimator of $m(u)$ by $\hat m^0(u) = \hat\alpha_{0,u}$. This estimator can also be written as $\hat m^0(u) = e_1^{\mathrm{T}}\,\tilde{\mathrm{S}}_N^{-1}\,\tilde{\mathrm{T}}_N$, where $e_1 = (1,0,\ldots,0)^{\mathrm{T}}$, and where $\tilde{\mathrm{S}}_N = (\tilde S_{N,k,k'})_{0\leq k,k'\leq\ell}$ and $\tilde{\mathrm{T}}_N = (\tilde T_{N,0},\ldots,\tilde T_{N,\ell})^{\mathrm{T}}$, with

$$\tilde S_{N,k,k'} = \frac{1}{Nh^{k+k'}} \sum_{j=1}^{J} \psi_j(q_0)\,q_0^{n_j-1} \sum_{i=1}^{n_j} K_h(U_{ij}-u)\,(U_{ij}-u)^{k+k'}\,, \tag{5}$$

$$\tilde T_{N,k} = \frac{1}{Nh^{k}} \sum_{j=1}^{J} \psi_j(q_0) \sum_{i=1}^{n_j} T_{ij}^*\,K_h(U_{ij}-u)\,(U_{ij}-u)^{k}\,.$$

In practice q_0 is unknown, but it is estimated root-N consistently by the maximum likelihood estimator $\hat q$ derived by Delaigle & Meister (2011), which is recalled in the supplementary material. To estimate γ, let $\epsilon_{ij} = q_0^{1-n_j}\,Z_j^* - m(U_{ij}) + \gamma^{\mathrm{T}}V_{ij}$. It follows from (4) that $E(\epsilon_{ij} \mid X_{ij}) = 0$. Therefore,

$$q_0^{1-n_j}\,Z_j^* - E\big(q_0^{1-n_j}\,Z_j^* \,\big|\, U_{ij}\big) = -\gamma^{\mathrm{T}}\{V_{ij} - E(V_{ij} \mid U_{ij})\} + \epsilon_{ij}\,.$$

We can replace q_0 by $\hat q$, and the functions $g_{ZU}(u) = E(q_0^{1-n_j}\,Z_j^* \mid U_{ij} = u)$ and $g_{VU}(u) = E(V_{ij} \mid U_{ij} = u)$ can be estimated using standard nonparametric regression, for example local-linear estimators with a cross-validation bandwidth. Let the estimators be $\hat g_{ZU}$ and $\hat g_{VU}$, computed from the data $(U_{ij}, \hat q^{1-n_j}\,Z_j^*)$ $(i=1,\ldots,n_j;\ j=1,\ldots,J)$ and (U_{ij},V_{ij}) $(i=1,\ldots,n_j;\ j=1,\ldots,J)$, respectively. This suggests estimating γ by

$$\hat\gamma = \operatorname*{argmin}_{\gamma} \sum_{j=1}^{J}\sum_{i=1}^{n_j} \Big[\hat q^{1-n_j}\,Z_j^* - \hat g_{ZU}(U_{ij}) + \gamma^{\mathrm{T}}\big\{V_{ij} - \hat g_{VU}(U_{ij})\big\}\Big]^2\,. \tag{6}$$

Replacing q_0 by $\hat q$ and γ by $\hat\gamma$ in the definition of $\hat m^0(u)$, we deduce the estimator

$$\hat m(u) \;=\; e_1^{\mathrm{T}}\,\hat{\mathrm{S}}_N^{-1}\,\hat{\mathrm{T}}_N\,, \tag{7}$$

where $\hat{\mathrm{S}}_N$ and $\hat{\mathrm{T}}_N$ denote the versions of $\tilde{\mathrm{S}}_N$ and $\tilde{\mathrm{T}}_N$ with every occurrence of q_0 and γ replaced by $\hat q$ and $\hat\gamma$, respectively. If $d > 1$, let $x = (u,v)$, where $u \in \mathbb{R}$ and $v \in \mathbb{R}^{d-1}$.

Since $p(x) = 1 - m(u) + \gamma^{\mathrm{T}} v$, we deduce that $p(x)$ can be estimated by

$$\hat{p}(x) = 1 - \hat{m}(u) + \hat{\gamma}^{\mathrm{T}} v\,. \tag{8}$$

When $X = U$, the estimator of p is found by letting $x = u$ and $\gamma = \hat{\gamma} = 0$ in (8). Moreover, in that case, the property $q_0 = E\{m(X)\}$ suggests a second estimator, $\hat{m}_{\mathrm{CB}}$, of m, found by standardising $\hat{m}$ so as to satisfy the identity $\hat{q} = N^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \hat{m}_{\mathrm{CB}}(X_{ij})$. Motivated by this, we define $\hat{m}_{\mathrm{CB}}$ by

$$\hat{m}_{\mathrm{CB}}(x) = \hat{q}\,\hat{m}(x) \Big/ \left\{ \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \hat{m}(X_{ij}) \right\}, \tag{9}$$

and we put $\hat{p}_{\mathrm{CB}}(x) = 1 - \hat{m}_{\mathrm{CB}}(x)$. The index CB stands for centralised bias. Indeed, we shall see in Section 4·1 that the asymptotic bias term of the estimator $\hat{m}_{\mathrm{CB}}$ is centralised; these quantities will be defined more precisely in Section 4·1. This estimator could be generalised to the case $d > 1$, but there, $q_0 = E\{m(U)\} - \gamma^{\mathrm{T}} V$, and thus instead of (9), the rescaling depends on $\hat{\gamma}$, which is much less attractive.

3. Multivariate case

3·1. *General multivariate estimator*

Our ideas can be extended to the multivariate case, where $X_{ij} = (X_{ij,1}, \ldots, X_{ij,d})^{\mathrm{T}}$ is a d-dimensional continuous vector. There, instead of using a partially linear model, we could estimate the function p completely nonparametrically. The extension can be made along the lines of standard multivariate local polynomial regression. For example, in the local-constant case we can estimate p at $x \in \mathbb{R}^d$ by $\hat{p}(x) = 1 - \hat{m}(x)$, where

$$\hat{m}(x) \,=\, \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} \psi_j(\hat{q})\, Z_j^* K_H(X_{ij} - x)}{\sum_{j=1}^{J} \sum_{i=1}^{n_j} \psi_j(\hat{q})\, \hat{q}^{n_j - 1}\, K_H(X_{ij} - x)}\,, \tag{10}$$

with K denoting a d-variate kernel, H a bandwidth matrix, and $K_H(x) = |H|^{-1/2} K(H^{-1/2}x)$. See Delaigle & Meister (2011) for an alternative local-constant estimator of m, which does not take the unequal sample sizes fully into account.

In the supplementary material we derive a local-linear version of this estimator; see equation (**??**). As we shall discuss in our numerical section, the local-linear estimator can work very well for d small, but it can also suffer from too much variability. In general we recommend using the local-constant estimator. See Section 6·1.

The purely multivariate estimators at (10) and (**??**) are consistent, but suffer from the usual curse of dimensionality. As in the standard non-grouped case, it can be proved that their convergence rate is order $N^{-2/(d+4)}$. To overcome this difficulty it is common to reduce dimension, for example using additive models or single-index models; see Fan & Gijbels (1996, pp. 274–276). We take up this issue in the next section.

3·2. *Single-index model*

In the parametric context with group testing data, p is often assumed to follow a parametric generalised linear model, where p depends on x only through $\beta_0^{\mathrm{T}} x$, with β_0 denoting a d-dimensional parameter. See, for example, Vansteelandt et al. (2000). Therefore, in the nonparametric case it seems natural to reduce dimension through single-index models, where the nonparametric form for p depends on x only through $\beta_0^{\mathrm{T}} x$.

Motivated by this, and recalling equation (4), we consider the single-index model, where we assume that the data (X_{ij}, Z_j^*) $(j = 1, \ldots, J;\, i = 1, \ldots, n_j)$, are generated from

$$Z_j^* = q_0^{n_j - 1}\, g\big(\beta_0^{\mathrm{T}} X_{ij}\big) + \epsilon_{ij}\,, \tag{11}$$

in which

$$q_0 = E\big\{g\big(\beta_0^{\mathrm{T}} X_{ij}\big)\big\}\,, \quad g\big(\beta_0^{\mathrm{T}} X_{ij}\big) = q_0^{1-n_j}\, E(Z_j^* \mid X_{ij})\,, \tag{12}$$

$Z_1^*, \ldots, Z_{n_j}^*$ are independent and identically distributed random variables, the X_{ij}s are independent and identically distributed random d-vectors, the sets $(Z_j^*, X_{1j}, \ldots, X_{n_j j})$ $(j = 1, \ldots, J)$, are independent of one another, β_0 is a fixed d-vector of unit length, g is a smooth, nonnegative function, $0 < q_0 < 1$, and ϵ_{ij} is defined by (11). In view of (12), $E(\epsilon_{ij} \mid X_{ij}) = 0$ for each i and j, and Z_j^* has the Bernoulli distribution with $\mathrm{pr}(Z_j^* = 1) = q_0^{n_j}$ and $\mathrm{pr}(Z_j^* = 1 \mid X_{ij}) = q_0^{n_j - 1}\, g(\beta_0^{\mathrm{T}} X_{ij})$. We wish to estimate q_0, β_0 and g, the latter nonparametrically. From there we can deduce an estimator of $p(x) = 1 - g(\beta_0^{\mathrm{T}} x)$.

3·3. *Estimation in the single-index model*

It is awkward to estimate q_0 and g together, since g can be estimated only at nonparametric rates, whereas q_0 can be approximated root-N consistently. We estimate q_0 by the maximum likelihood estimator, $\hat{q}$, of Delaigle & Meister (2011). Let ψ_j, for $j = 1, \ldots, J$, denote smooth, positive functions defined on $[0, 1]$, let f_β denote the density of $\beta^{\mathrm{T}} X_{ij}$, and let $g_\beta(t) = E(q_0^{1-n_j}\, Z_j^* \mid \beta^{\mathrm{T}} X_{ij} = t)$. Motivated by the estimator $\hat{m}$ introduced in Section 2·2, we define an estimator of g, when $\beta_0 = \beta$, by

$$\hat{g}_\beta(t \mid h) = \hat{a}_\beta(t \mid h) \big/ \hat{b}_\beta(t \mid h)\,, \tag{13}$$

where β is in the set B_0 of all unit d-vectors,

$$\hat{a}_\beta(t \mid h) = \frac{1}{N} \sum_{j=1}^{J} \psi_j(\hat{q})\, Z_j^* \sum_{i=1}^{n_j} K_h\big(t - \beta^{\mathrm{T}} X_{ij}\big)\,, \tag{14}$$

$$\hat{b}_\beta(t \mid h) = \frac{1}{N} \sum_{j=1}^{J} \psi_j(\hat{q})\, \hat{q}^{n_j - 1} \sum_{i=1}^{n_j} K_h\big(t - \beta^{\mathrm{T}} X_{ij}\big)\,. \tag{15}$$

The quantities $\hat{a}_\beta(t)$ and $\hat{b}_\beta(t)$ can be viewed as estimators of, respectively,

$$a_\beta(t) = f_\beta(t)\, g_\beta(t)\, \times M/N\,, \quad b_\beta(t) = f_\beta(t)\, \times M/N\,, \tag{16}$$

where $M = \sum_{j \le J} n_j\, \psi_j(q_0)\, q_0^{n_j - 1}$. See the proof of Theorem 2 for details.

As (14) and (15) indicate, we describe here in detail the single-index model using local-constant methods, but it is also possible to use more general local polynomial methods; see the supplementary material. As will be discussed in Section 6, in practice, while the local-linear estimator can sometimes improve on the local-constant one for d small, local-constant fitting provides substantial robustness against the variance problems that can afflict higher order techniques when the sample size, N, is not sufficiently large. These difficulties reflect the fact that a local-constant estimator never takes the form of a nonzero number divided by zero, whereas a local-linear estimator can have that form.

We shall use the notation $\hat{\beta}$ to denote an estimator of β_0, and $\hat{a}_{\hat{\beta}}(t \mid h) / \hat{b}_{\hat{\beta}}(t \mid h)$ will denote an estimator of $a_{\beta_0}(t)/b_{\beta_0}(t) = g(t)$. Next, to reflect the first part of (12), and

using arguments similar to those for $\hat{m}_{\mathrm{CB}}$ in Section 2·1, we define a second estimator by

$$\hat{g}_{\beta,\mathrm{CB}}(t \mid h) = \hat{q}\,\hat{g}_\beta(t \mid h) \Big/ \Big\{ \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \hat{g}_\beta\big(\beta_0^{\mathrm{T}} X_{ij} \mid h\big) \Big\}. \tag{17}$$

To choose h and define $\hat{\beta}$ we shall use a cross-validation approach inspired by work of Härdle et al. (1993). However, two difficulties prevent us from applying their method directly to our context: the data are not identically distributed, in fact the Z_j^*s do not even have the same mean; and all the X_{ij}s from group j share the same Z_j^*. For the same reason, theoretical properties of our estimator are much more difficult to derive than those in Härdle et al. (1993). To overcome these challenges, we proceed as follows. Take A to be a compact subset of $\mathbb{R}^d$, and define

$$\hat{g}_\beta^{(-j)} = \hat{a}_\beta^{(-j)} \big/ \hat{f}_\beta^{(-j)} \tag{18}$$

to be the version of $\hat{g}_\beta$ computed on omitting all pairs (Z_j^*, X_{ij}) $(i = 1, \ldots, n_j)$; the estimators $\hat{a}_\beta^{(-j)}$ and $\hat{f}_\beta^{(-j)}$ are also constructed in this leave-out manner. We introduce the squared-error cross-validation criterion,

$$S_1(h, \beta) = \sum_{X_{ij} \in A} \sum \Big\{ Z_j^* - \hat{q}^{n_j - 1}\, \hat{g}_\beta^{(-j)}\big(\beta^{\mathrm{T}} X_{ij} \mid h\big) \Big\}^2 \phi_j(\hat{q}), \tag{19}$$

where $\sum \sum_{X_{ij} \in A}$ denotes summation over (i, j) such that $1 \le i \le n_j$, $1 \le j \le J$ and $X_{ij} \in A$, and $\phi_1, \ldots, \phi_J$ are smooth, positive weight functions. Omitting pairs (Z_j^*, X_{ij}) $(i = 1, \ldots, n_j)$ involves reducing sample size from N to $N - n_j$. Observe too that, in (19), it is unnecessary to omit data from the jth group when computing $\hat{q}$.

We choose $(h, \beta) = (\hat{h}, \hat{\beta})$ to minimise $S_1(h, \beta)$. In this notation our estimator of $g(t)$, where g is the function in the model at (11), is either $\hat{g}_{\hat{\beta}}(t \,|\, \hat{h})$ or $\hat{g}_{\hat{\beta},\mathrm{CB}}(t \,|\, \hat{h})$, where $\hat{g}_\beta(t \mid h)$ and $\hat{g}_{\beta,\mathrm{CB}}(t \mid h)$ are given by (13) and (17), respectively.

4. Asymptotic properties

4·1. *Asymptotic normality of the nonparametric estimator in Section 2·2*

Recall that, in the general case, $X = (U, V^{\mathrm{T}})^{\mathrm{T}}$, where U is a continuous variable and $V \in \mathbb{R}^{d-1}$ is discrete. Let f_U denote the density of U. We make the following assumptions:

(H1) K is real and symmetric, $\|K\|_\infty < \infty$, $\int K(x)\,dx = 1$, $\int |x|^{2\ell+3} |K(x)|\,dx < \infty$, and $\int \big(|x|^{3\ell+1} + x^{4\ell}\big) K(x)^2\,dx < \infty$;

(H2) $h \to 0$ and $Nh \to \infty$ as $N \to \infty$;

(H3) $f_U(u) > 0$ and f_U is twice differentiable and satisfies $\|f_U^{(j)}\|_\infty < \infty$ for $j = 0, 1, 2$;

(H4) m is $\ell + 2$ times differentiable, and $\|m^{(j)}\|_\infty < \infty$ for $j = 0, \ldots, \ell + 2$;

(H5) $0 < \inf_j \psi_j(q_0) < \sup_j \psi_j(q_0) < \infty$;

(H6) $\sup_j n_j < \infty$ and $0 < q_0 < 1$.

We introduce the following notation: $\mu_j = \int x^j\, K(x)\,dx$, $\nu_j = \int x^j\, K(x)^2\,dx$, $\mu = (\mu_{\ell+1}, \ldots, \mu_{2\ell+1})^{\mathrm{T}}$, $\widetilde{\mu} = (\mu_{\ell+2}, \ldots, \mu_{2\ell+2})^{\mathrm{T}}$, $\mathrm{S} = \big(S_{k,k'}\big)_{0 \le k,k' \le \ell}$, $\mathrm{S}^* = \big(S^*_{k,k'}\big)_{0 \le k,k' \le \ell}$

where $S_{k,k'} = \mu_{k+k'}$, $S^*_{k,k'} = \nu_{k+k'}$, $M = \sum_{j\le J} n_j\,\psi_j(q_0)\,q_0^{n_j-1}$ and

$$\tau_j(u)^2 = m(u)\,q_0^{n_j-1}\left\{1 - q_0^{n_j-1}\,m(u)\right\} - q_0^{n_j-1}\,E\left\{\left(\gamma^{\mathrm{T}}V_{ij}\right) \mid U_{ij} = u\right\}\left\{1 - 2\,q_0^{n_j-1}\,m(u)\right\} - q_0^{2n_j-2}\,E\left\{\left(\gamma^{\mathrm{T}}V_{ij}\right)^2 \mid U_{ij} = u\right\}.$$

The next theorem establishes asymptotic normality of $\hat m^0$, defined above equation (5). Its proof is similar to that of Theorem 3.1 of Delaigle et al. (2009); see the supplementary material.

THEOREM 1. *Under Conditions (H1)–(H6), we have*

$$\hat m^0(u) = m(u) + B(u) + V(u)^{1/2}\,\mathcal{N}_N + o_p\{B(u)\} + o_p\{V(u)^{1/2}\}, \tag{20}$$

where the random variable $\mathcal{N}_N$ is asymptotically normal $\mathrm{N}(0,1)$,

$$V(u) = e_1^{\mathrm{T}}\,\mathrm{S}^{-1}\,\mathrm{S}^*\,\mathrm{S}^{-1}\,e_1\,\frac{1}{M^2\,h\,f_U(u)}\sum_{j\le J} n_j\,\psi_j(q_0)^2\,\tau_j(u)^2,$$

$$B(u) = \begin{cases} e_1^{\mathrm{T}}\,\mathrm{S}^{-1}\mu\,\dfrac{1}{(\ell+1)!}\,m^{(\ell+1)}(u)\,h^{\ell+1} & \ell \text{ odd;} \\ e_1^{\mathrm{T}}\,\mathrm{S}^{-1}\,\widetilde{\mu}\,\dfrac{1}{(\ell+2)!}\left\{(\ell+2)\,m^{(\ell+1)}(u)\,\dfrac{f_U'(u)}{f_U(u)} + m^{(\ell+2)}(u)\right\}h^{\ell+2} & \ell \text{ even.} \end{cases}$$

Standard arguments for partially linear models can be used to prove that $\hat\gamma = \gamma + O_p(N^{-1/2})$. Similarly, it follows from Delaigle & Meister (2011) that $\hat q = q_0 + O_p(N^{-1/2})$. This rate of convergence is so fast that, if γ and q_0 are replaced by $\hat\gamma$ and $\hat q$ in the formula for $\hat m^0$, then the error that is introduced is negligible, to first order, relative to the error in $\hat m^0$ as an approximation to m. Consequently, (20) also holds if $\hat m^0(u)$, on the left-hand side, is replaced by $\hat m(u)$. The methods used are conventional, and ask of the ψ_js that they have uniformly bounded first derivatives.

Thus, for both $\hat m$ and $\hat m^0$, the best convergence rate is obtained using h such that $B(u)^2 \asymp V(u)$, where $a \asymp b$ means that $a = O(b)$ and $b = O(a)$. Then, since (H5)–(H6) imply that $M \asymp N$ and $V(u) \asymp 1/(Nh)$, we have $\hat m(u) - m(u) = O_p(N^{-(\ell+1)/(2\ell+3)})$ if ℓ is odd and $O_p(N^{-(\ell+2)/(2\ell+5)})$ if ℓ is even.

Recall the estimator $\hat m_{\mathrm{CB}}$, defined at (9). Using arguments similar to those employed in Step 7 of the proof of Theorem 2, it can be shown that with h chosen as in the previous paragraph, $\hat m_{\mathrm{CB}}(x) = \hat m(x) - q_0^{-1}\,m(x)\,E\{B(X)\} + o_p\{B(x)\}$ and

$$\hat m_{\mathrm{CB}}(x) = m(x) + B_{\mathrm{CB}}(x) + V(x)^{1/2}\,\mathcal{N}_N + o_p\{B(x)\},$$

where $B_{\mathrm{CB}}(x) = q_0^{-1}\,[\mathrm{E}\{m(X)\}\,B(x) - m(x)\,E\{B(X)\}]$. Here we used the fact that $q_0 = E\{m(X)\}$. In particular, $E\{B_{\mathrm{CB}}(X)\} = 0$, whence the name centralised bias estimator.

A more general consistent estimator of m can be defined by replacing the ψ_js in (5) by $\tilde\psi_j$s potentially different from the ψ_js, and satisfying $\sum_{j\le J} n_j\,\tilde\psi_j = \sum_{j\le J} n_j\,\psi_j\,q_0^{n_j-1}$. It can be proved that this estimator has the same asymptotic bias, B, as ours, and that its asymptotic variance, V, is minimised by taking $\tilde\psi_j = \psi_j$. The estimator of Delaigle & Meister (2011) can be expressed in this general form, taking $\tilde\psi_j = 1$ and $\psi_j = N/(\sum_{k\le J} n_k\,q_0^{n_k-1})$. Our analysis shows that our estimator has more attractive asymptotic properties than theirs.

4·2. *Theoretical properties of the single-index model*

Let $A \subset \mathbb{R}^d$ be the closure of a union of a finite number of bounded, open convex sets, let X have the distribution of a generic X_{ij}, assume that the distribution of X is continuous, write f for the density of X, and recall that f_β is the density of $\beta^{\mathrm{T}} X$. Let B_{nhd} represent an open neighbourhood of $\beta_0 \in B_0$.

To establish properties of our estimator, assume that conditions (I1)–(I8) in Section A·1 hold. Recall the definition of $\hat{g}_\beta(t \mid h)$ at (13). Let $\hat{g}_\beta^{[0\,(-j)]}$ be the version of $\hat{g}_\beta$ that arises if we replace $\hat{q}$ by q_0, and if we leave out all pairs $(Z_j^*, X_{1j}), \ldots, (Z_j^*, X_{n_j j})$. That is, $\hat{g}_\beta^{[0\,(-j)]}$ is the version of $\hat{g}_\beta^{(-j)}$, at (18), that is obtained on replacing $\hat{q}$ by q_0. Put $S_2(\beta) = \sum \sum_{X_{ij} \in A} \{Z_j^* - q_0^{n_j - 1}\, g_\beta(\beta^{\mathrm{T}} X_{ij})\}^2\, \phi_j(q_0)$, $S_3(h) = \sum \sum_{X_{ij} \in A} \{\hat{g}_{\beta_0}^{[0\,(-j)]}(\beta_0^{\mathrm{T}} X_{ij} \mid h) - g(\beta_0^{\mathrm{T}} X_{ij})\}^2\, q_0^{2(n_j-1)}\, \phi_j(q_0)$, with g_β as in Section 3·3. Thus, $S_2(\beta)$ is the sum-of-squares criterion we would use to compute $\hat{\beta}$ if we knew q_0 and g_β, and $S_3(h)$ is the cross-validation criterion we would employ to compute the weighted least-squares bandwidth for estimating g if we knew q_0 and β_0.

Let $(\hat{h}, \hat{\beta})$ be the minimiser of $S_1(h, \beta)$, at (19), over $(h, \beta) \in H_N \times B_N$. Let X have the distribution of an X_{ij} but be independent of all the data Z_j^* and X_{ij}, and let $c_0 > 0$ denote the constant such that the bandwidth $h_0 = h_0(N)$ that minimises $E[\{\hat{g}_{\beta_0}(\beta_0^{\mathrm{T}} X \mid h) - g(\beta_0^{\mathrm{T}} X)\}^2\, I(X \in A)]$ satisfies $h_0 \sim c_0\, N^{-1/5}$ as $N \to \infty$. Finally, let

$$\hat{g}^0(t) = \frac{\sum_j \psi_j(q_0)\, Z_j^* \sum_i K_{h_0}(t - \beta_0^{\mathrm{T}} X_{ij})}{\sum_j \psi_j(q_0)\, q_0^{n_j - 1} \sum_i K_{h_0}(t - \beta_0^{\mathrm{T}} X_{ij})} \tag{21}$$

denote the estimator of g that we would use if we knew h_0, q_0 and β_0. If we substitute these values for h, $\hat{q}$ and β in the definition of $\hat{g}_\beta(t \mid h)$, at (13), we obtain $\hat{g}^0(t \mid h_0)$.

It follows from Theorem 1 that

$$\hat{g}^0(t) = g(t) + (N h_0)^{-1/2}\, g_2(t)\, \mathcal{N}_N + h_0^2\, g_3(t) + o_p\big\{(N h_0)^{-1/2} + h_0^2\big\}\,, \tag{22}$$

where the random variable $\mathcal{N}_N$ is asymptotically normal $\mathrm{N}(0, 1)$, $g_3(t) = \mu_2\, \{g'(t)\, f'_{\beta_0}(t)/f_{\beta_0}(t) + g''(t)/2\}$, and, with $R(K) = \int K^2$, we define

$$g_2(t)^2 = R(K)\, N\, \big\{M^2\, h_0\, f_{\beta_0}(t)\big\}^{-1} \sum_{j=1}^{J} n_j\, \psi_j(q_0)^2\, q_0^{n_j - 1}\, g(t)\, \big\{1 - q^{n_j - 1}\, g(t)\big\}\,.$$

In particular, $g_2 \geq 0$ and g_3 are continuous functions. Result (22) holds for all $t \in \mathcal{T}$, where $\mathcal{T}$ is the set of all $\beta^{\mathrm{T}} x$ with x constrained to lie in any given open set in the interior of the support of the distribution of X, and β lies in a sufficiently small open neighbourhood of β_0, depending on the aforementioned open set.

Finally we are ready to establish properties of our estimator, in the next theorem. See the supplementary material for a proof.

THEOREM 2. *Take v_0, W_1, Σ_0, Σ_1 and Σ_2 as in the supplementary material. If conditions (I1)–(I8) in the supplementary material hold, then (i)*

$$\begin{aligned} S_1(h, \beta) =& S_2(\beta) + S_3(h) + 2N\, (\hat{q} - q_0)\, (\beta - \beta_0)^{\mathrm{T}}\, v_0 + V_1 \\ &+ o_p\Big[N\, \big\{(Nh)^{-1} + h^4\big\} + N^{1/2}\, \|\beta - \beta_0\| + N\, \|\beta - \beta_0\|^2 + 1\Big] \qquad (23) \\ =& N\, \Big[(\beta - \beta_0)^{\mathrm{T}}\, \Sigma_0\, (\beta - \beta_0) - 2\, (\beta - \beta_0)^{\mathrm{T}}\, \{W_1 - (\hat{q} - q_0)\, v_0\}\Big] + S_3(h) + V_2 \end{aligned}$$

$$+ o_p\Big[N\left\{(Nh)^{-1} + h^4\right\} + N^{1/2}\,\|\beta - \beta_0\| + N\,\|\beta - \beta_0\|^2 + 1\Big]\,, \tag{24}$$

uniformly in $(h, \beta) \in H_N \times B_N$*, and the random variables* V_1 *and* V_2 *do not depend on* β *or* h*; (ii)* $N^{1/5}\,\hat{h} \to c_0$ *in probability as* $N \to \infty$*; (iii)* $N^{1/2}\,(\hat{\beta} - \beta_0, \hat{q} - q_0)$ *is asymptotically normally distributed with zero mean and covariance matrix* Σ_2*; (iv)* $\hat{g}_{\hat{\beta}}(t \,|\, \hat{h}) = \hat{g}^0(t \mid h_0) + o_p(N^{-2/5})$*, uniformly in* $t \in \mathcal{T}$*, where* $\hat{g}^0$ *is as at* (21) *and* $\mathcal{T}$ *is as defined below* (22)*; and (v)* $\hat{g}_{\hat{\beta},\mathrm{CB}}(t \,|\, \hat{h}) = \hat{g}^0(t \mid h_0) - q_0^{-1}\,g(t)\,h_0^2\,E\{g_3(\beta_0^{\mathrm{T}} X)\} + o_p(N^{-2/5})$*, uniformly in* $t \in \mathcal{T}$*.*

It follows from part (v) of the theorem that the estimator $\hat{g}_{\hat{\beta},\mathrm{CB}}$, at (17), generally differs from $\hat{g}^0$ by a term of size h_0^2, resulting from a bias contribution associated with the denominator on the right-hand side of (17). In particular, using arguments similar to those in Section 4·1, the asymptotic bias term of $\hat{g}_{\hat{\beta},\mathrm{CB}}(x)$ is centred, since $\hat{g}_{\hat{\beta},\mathrm{CB}}(x) = g(x) + (Nh_0)^{-1/2}\,g_2(t)\,\mathcal{N}_N + h_0^2\,g_{3,\mathrm{CB}}(x) + o_p(N^{-2/5})$, where $E\{g_{3,\mathrm{CB}}(X)\} = 0$.

5. Computing the estimators in practice

5·1. *Computing the weights* ψ_j

Abusing terminology a little, in the sequel we shall refer to B and V as, respectively, the asymptotic bias and variance of $\hat{m}$. Theoretically, we can define locally optimal weights $\psi_j^*(q_0; x)$ to be the ψ_js which minimise $B(x)^2 + V(x)$. See the supplementary material for an explicit formula for $\psi_j^*(q_0; x)$. Instead of using local weights, we can use global weights, which are simpler to calculate. We define the global weights $\psi_j^*(q_0)$ to be the ψ_js which minimise the asymptotic weighted integrated mean squared error, $\mathrm{AMISE}_w = \int \{B(x)^2 + V(x)\}\,w(x)\,dx$, where $w = f_X\,\omega$, with ω denoting a nonnegative function. It is common, in local polynomial regression, to use a weighted criterion of this type; see Fan & Gijbels (1996, p. 67). Specific choice of ω will be discussed in Section 6.

Since the ψ_js influence only the variance part, the ψ_j^*s are found by minimising $\int V\omega$ with respect to the ψ_js, which gives:

$$\psi_j^*(q_0) = \left\{ \int m(x)\,\omega(x)\,dx - q_0^{n_j - 1} \int m(x)^2\,\omega(x)\,dx \right\}^{-1}.$$

The ψ_j^*s depend on m and q_0, which are unknown and have to be estimated from the data. In the case $X = U$ we suggest estimating these weights by

$$\hat{\psi}_j^*(\hat{q}) = \left\{ \int \hat{m}_{\mathrm{PILOT}}(x)\,\omega(x)\,dx - \hat{q}^{n_j - 1} \int \hat{m}_{\mathrm{PILOT}}(x)^2\,\omega(x)\,dx \right\}^{-1}, \tag{25}$$

where $\hat{q}$ denotes the maximum likelihood estimator of q_0, discussed in the supplementary material, and $\hat{m}_{\mathrm{PILOT}}$ denotes the estimator of Delaigle & Meister (2011) computed using their plug-in bandwidth.

These arguments can be extended, and used to compute optimal weights for the partially linear model, where $X = (U, V)$, but the weights in this case are difficult to compute in practice. We suggest approximating them by the weights at (25), taking $\hat{m}_{\mathrm{PILOT}}$ to be the local-constant estimator $\hat{m}$ at (7) computed with weights $\psi_j(\hat{q}) = (1 - \hat{q}^{n_j})^{-1}$ and the cross-validation bandwidth of the supplementary material. These pilot weights $\psi_j(\hat{q})$ result from replacing m by $\hat{q}$ in the definition of $\psi_j^*(q_0; x)$.

For the single-index model estimator of Section 3·3 we use the formula at (25), replacing there $\hat{m}_{\text{PILOT}}(x)$ by $\hat{m}_{\text{PILOT}}(\hat{\beta}_{\text{PILOT}}^{\text{T}} x)$, where $\hat{m}_{\text{PILOT}}$ denotes the estimator obtained when computing $\hat{m}$ with weights $\psi_j(\hat{q}) = (1 - \hat{q}^{n_j})^{-1}$ and with a bandwidth h chosen by cross-validation, as at (26), and $\hat{\beta}_{\text{PILOT}}$ is a pilot estimator of β, which we obtain by fitting a linear regression to the data, using the global polynomial procedure described in Section 4.1.2 of Delaigle & Meister (2011).

5·2. *Bandwidth for the estimator in Section 2·2*

In the case where $X = U$, we suggest a plug-in bandwidth procedure for the local-linear estimator of m, where $\ell = 1$, which is probably the most popular form of the local polynomial estimator. In univariate nonparametric regression, it is well known that plug-in bandwidths usually outperform cross-validation bandwidths. We define our plug-in bandwidth as the bandwidth that minimises an estimator of AMISE_w, which was introduced in Section 5·1, and which, using Theorem 1, can be written as:

$$\text{AMISE}_w = \frac{h^4}{4}\,\mu_2^2\,\theta_2 + \frac{R(K)}{h}\,\frac{\sum_{j=1}^J n_j\, q^{n_j-1}\psi_j^2}{(\sum_{k=1}^J n_k\,\psi_k\,q^{n_k-1})^2}\int \left\{1 - q^{n_j-1}\,m(x)\right\} m(x)\,\omega(x)\,dx\,,$$

where $\theta_2 = \int \{m''(x)\}^2\, f_X(x)\,\omega(x)\,dx$ and ω is as in Section 5·1. As in Delaigle & Meister (2011), to estimate AMISE_w we use ideas employed by Ruppert et al. (1995). However, our procedure differs from that of Delaigle & Meister (2011) in that, unlike them, our approach does not require us to introduce extra weights.

The second term in the formula for AMISE_w is the easiest to estimate, since it requires only a pilot estimator of m. For this we use the local-constant estimator of m computed with the cross-validation bandwidth obtained by minimising

$$\text{CV}(h) = \sum_{j=1}^J \sum_{i=1}^{n_j} \left\{Z_j^* - \hat{q}^{n_j-1}\,\hat{m}^{(-j)}(X_{ij})\right\}^2 1_{[a,b]}(X_{ij})\,, \tag{26}$$

where $\hat{m}^{(-j)}$ denotes the local-constant estimator of m computed without using observations from the jth group, and a and b are a lower and an upper empirical quantile of the X_{ij}s, for example the 10th and 90th percentiles.

The quantity θ_2 is the most difficult to estimate. To estimate it, our local polynomial technique is extended to construct consistent estimators of derivatives of m. For $\nu \le \ell$, define the ℓth order local polynomial estimator of $m^{(\nu)}(x)$ by $\hat{m}^{(\nu)}(x) = \nu!\, h^{-\nu} e_{\nu+1}^{\text{T}}\, \hat{\text{S}}_N^{-1}\, \hat{\text{T}}_N$, in which $e_{\nu+1} = (0,\ldots,0,1,0,\ldots,0)^{\text{T}}$ where the 1 is at the $(\nu+1)$th position. Consistency of this estimator can be established along the lines of Theorem 3.1 of Delaigle et al. (2009). Motivated by this, and following ideas of Ruppert et al. (1995), we take

$$\hat{\theta}_2 = \frac{1}{N}\sum_{j=1}^J \sum_{i=1}^{n_j} \{\hat{m}''_{(-j)}(X_{ij})\}^2\,\omega(X_{ij})\,,$$

where $\hat{m}''_{(-j)}$ denotes the local polynomial estimator of m'' of order $\ell = 3$, computed without using the data from the jth group. To compute $\hat{\theta}_2$ we need a bandwidth, h_2 say, which is necessarily different from h. However, to choose it we cannot use the approach of Ruppert et al. (1995), which is valid only for standard regression. In the supplementary material we show how to extend their method to our group testing setting.

Table 1. *Simulation results for models (i) to (iv) using the estimator* $\hat{g} = 1 - \hat{m}$ *with* $\hat{m}$ *at* (7). *The numbers show* $10^4 \times$ *median integrated squared error, and interquartile range in square brackets, calculated from 200 simulated samples.*

Model	$N = 1{,}000$	$N = 5{,}000$	$N = 10{,}000$	Model	$N = 1{,}000$	$N = 5{,}000$	$N = 10{,}000$
(i) A	28·4[31·7]	10·5[7·04]	6·54[4·82]	(iii) A	4·62[8·08]	1·16[1·37]	0·76[0·87]
(i) B	38·7[40·1]	12·7[9·35]	8·09[5·28]	(iii) B	6·84[8·79]	1·64[2·06]	1·07[1·20]
(ii) A	17·5[24·3]	4·86[5·79]	2·82[2·92]	(iv) A	25·5[32·8]	7·43[6·98]	4·56[4·07]
(ii) B	21·5[32·3]	6·16[7·35]	3·56[4·78]	(iv) B	32·8[46·7]	11·1[10·8]	5·32[6·09]

Table 2. *Simulation results for models (i) to (iv) using the estimator* $\hat{\gamma}$ *at* (6). *The numbers show* $10^3 \times$ *median squared error, and interquartile range in square brackets, calculated from 200 simulated samples.*

Model	$N = 1{,}000$	$N = 5{,}000$	$N = 10{,}000$	Model	$N = 1{,}000$	$N = 5{,}000$	$N = 10{,}000$
(i) A	6·67[20·2]	1·23[4·17]	0·74[1·59]	(iii) A	12·8[28·0]	1·44[4·67]	1·10[3·43]
(i) B	11·2[30·8]	2·49[6·01]	1·01[2·56]	(iii) B	15·5[42·7]	3·03[9·46]	1·37[4·06]
(ii) A	6·50[23·4]	1·41[3·98]	0·67[1·78]	(iv) A	11·5[27·7]	1·78[5·55]	1·08[2·85]
(ii) B	11·1[38·0]	3·21[7·19]	1·44[4·36]	(iv) B	12·8[37·6]	3·32[11·0]	1·76[4·44]

Our arguments can be extended for computing a plug-in bandwidth in the case of the partially linear model, where $X = (U, V)$. However, the resulting formula involves many unknowns; recall the definition of τ_j^2 above Theorem 1. Consequently the bandwidth is too variable to work well in practice. We experimented with this approach and found that better results could be obtained by using instead the plug-in bandwidth described above, pretending that $X = U$. This is the bandwidth we recommend using in practice.

6. NUMERICAL ILLUSTRATIONS

6·1. *Simulations*

We ran simulations for the univariate, partially linear, and multivariate procedures. Our goal was threefold: (a) in the univariate case, demonstrate the superiority of our approach over that of Delaigle & Meister (2011); (b) in the partially linear case, illustrate the performance of our method; (c) in the multivariate case, compare the purely multivariate nonparametric estimator with the single-index estimator.

For (a) we used the same four models as Delaigle & Meister (2011). Due to space considerations we provide the results of the comparison of our univariate estimator $\hat{p}_{\mathrm{CB}}$ with the method of Delaigle & Meister (2011) only in the supplementary material. Those results show that our procedure can improve significantly on the method of Delaigle & Meister (2011). For (b) we generalised the univariate models by incorporating a discrete variable V. Specifically, for g, U, and γ in model (2) we took:

(i) $g(u) = \{\sin(\pi u/2) + 1 \cdot 2\}/[20 + 40u^2\{\mathrm{sign}(u) + 1\}]$, $U \sim \mathrm{N}(0, 1 \cdot 5^2)$ and $\gamma = 0 \cdot 1$;
(ii) $g(u) = \exp(-4 + 2u)/\{8 + 8\exp(-4 + 2u)\}$, $U \sim \mathrm{N}(2, 1 \cdot 5^2)$ and $\gamma = 0 \cdot 05$;
(iii) $g(u) = u^2/8$, $U \sim \mathrm{N}(0 \cdot 5, 0 \cdot 5^2)$ and $\gamma = 0 \cdot 1$;
(iv) $g(u) = u^2/8$, $U \sim \mathrm{N}(0, 0 \cdot 75^2)$ and $\gamma = 0 \cdot 1$.

We took V to be a Bernoulli variable independent of U, with $P(V = 0) = 0 \cdot 75$. We also considered a version where V was dependent of U; see the supplementary material for details.

In each case we grouped the data in two different ways, as follows. A: $[N/4]$ groups of size 2 and $[N/12]$ groups of size 6 and B: $[N/4]$ groups of size 2 and $[N/20]$ groups of size 10. Since group testing is most often employed to save money in large studies, the total sample size, N, is typically rather large. Reflecting this, here we consider three values of

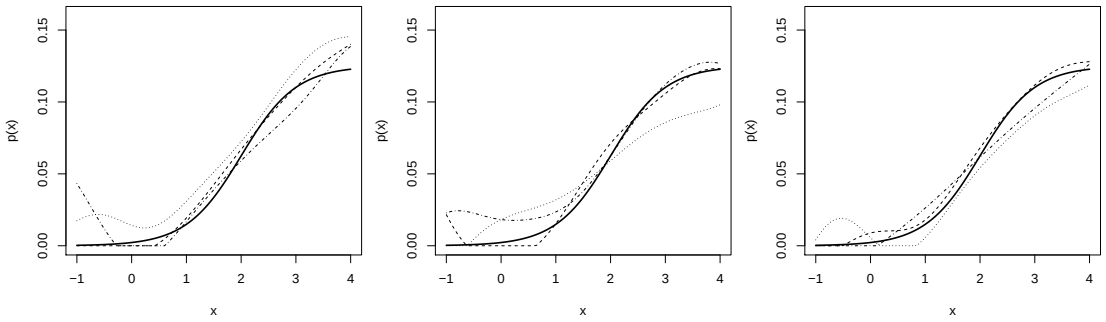

Fig. 1. Curves $\hat{g}$ corresponding to the 20th, dashed line; 100th, dot-dashed line; and 180th, dotted line; values of $\int_{[a,b]} |\hat{g} - g|$ for model (ii) with grouping A when $N = 5{,}000$ and U and V are dependent for the left column, $N = 5{,}000$ and U and V are independent for the centre column, and $N = 10{,}000$ and U and V are independent for the right column. The true curve is depicted by the solid line.

N: 1,000, 5,000 and 10,000. We generated 200 samples from each model, and applied the estimators $\hat{\gamma}$ at (6) and $\hat{g} = 1 - \hat{m}$ with $\hat{m}$ at (7) to each sample. As in Delaigle & Meister (2011), all estimators of g were truncated to the interval $[0, 1]$. To assess the performance of our estimators, we computed, for the 200 samples, the squared error $(\hat{\gamma} - \gamma)^2$ and the integrated squared error, $\int_{[a,b]} (\hat{g} - g)^2$, on the same interval $[a, b]$ as Delaigle & Meister (2011).

We computed the local-linear estimator $\hat{m}$ at (7) using the plug-in bandwidth of Section 5·2 and the weights at (25). For the ω used to compute the weights in (25) and the plug-in bandwidth, we took $\omega(x) = 1_{[q_{0\cdot1}, q_{0\cdot9}]}(x)$, where q_α denotes the empirical α quantile of the distribution of the U_{ij}s.

In Table 1 we show, for each of models (i) to (iv), combined with each of the groupings A and B, the median and interquartile range of the 200 integrated squared error values for $\hat{g}$, obtained for three values of N. In Table 2 we show the median and interquartile range of the squared errors of $\hat{\gamma}$. In Fig. 1, for model (ii) with grouping A, we show three estimated curves $\hat{g}$ for $N = 5{,}000$ and 10,000 when U and V are independent, and for $N = 5,000$ in the case treated in the supplementary material where U and V are dependent. In each case, the three curves correspond to the samples that resulted in the 20th, 100th and 180th smallest values of $\int_{[a,b]} |\hat{g} - g|$.

In the multivariate case we compared the method of Section 3·1 for the single-index model with the purely multivariate estimator of Section 3·2. We simulated from the following models:

(v) $p(x_1, x_2) = \exp(-4 - 4x_1 + 4x_2)/\{8 + 8\exp(-4 - 4x_1 + 4x_2)\}$ and $X^{\mathrm{T}} = (V_1 + V_2, V_1 + V_3)$, where $V_1 \sim \mathrm{N}(0, 0 \cdot 2^2)$, $V_2 \sim \mathrm{N}(0, 1 \cdot 5^2)$ and $V_3 \sim \mathrm{N}(0, 1 \cdot 5^2)$;

(vi) $p(x_1, x_2) = (-x_1 + 2x_2 - 0 \cdot 5)^2/8$ and $X^{\mathrm{T}} = (V_1 + V_2, V_1 + V_3)$, where $V_1 \sim \mathrm{N}(0, 0 \cdot 1^2)$, $V_2 \sim \mathrm{N}(0, 0 \cdot 5^2)$ and $V_3 \sim \mathrm{N}(0, 0 \cdot 5^2)$;

(vii) $p(x_1, x_2, x_3) = \exp(-4 - 10x_1 + 6x_2 + 10x_3)/\{8 + 8\exp(-4 - 10x_1 + 6x_2 + 10x_3)\}$ and $X^{\mathrm{T}} = (X_1, X_2, X_3)$, where $X_3 \sim \mathrm{N}(0, 1 \cdot 5^2)$, $X_1 \sim U[-2, 2]$ and $X_2 \sim U[-2, 2]$;

(viii) $p(x_1, x_2, x_3) = \{(1/2) - (-5x_1 + 3x_2 + 5x_3)^2/8\}\, \phi(-10x_1 + 6x_2 + 10x_3)$ and X is as in (vii). Here ϕ denotes the standard normal density.

In each case we grouped the data in two ways, as follows. A: $[N/4]$ groups of size 2, and $[N/12]$ groups of size 6; B: $[N/10]$ groups of size 5, and $[N/20]$ groups of size 10.

For each h on a grid, estimators of β_0 were computed by minimising the cross-validation criterion at (19), where minimisation was undertaken numerically and where we took

$\phi_j = 1$. Owing to local minima, the success of the procedure depends on the quality of the starting point of the numerical minimisation procedure. We took two starting points: the vector β found by fitting a linear model to the data and the estimator $\tilde{\beta}$ obtained using a gradient approach. See the supplementary material for details. We found this to work reasonably well for the examples we considered. An alternative could be to use the logistic fit from Vansteelandt et al. (2000); see also Bilder et al. (2010).

Let $\hat{\beta}_h$ denote the solution for a given h. We estimated β_0 by $\hat{\beta}_{\hat{h}} = \text{argmin}_{h,\hat{\beta}_h} S_1(h, \hat{\beta}_h)$, where $h \in H_n$ with H_n as in the supplementary material. We compared our single-index estimator from Section 3·3 with the purely multivariate estimator from Section 3·1, which we computed by taking $K(x) = \phi(\|x\|)/\int \phi(\|y\|)\, dy$ with ϕ as above, and H to be diagonal, with diagonal elements equal to $h_1^2, \ldots, h_d^2$. Here $h_j = h\hat{\sigma}_j$, with $\hat{\sigma}_j^2$ denoting the empirical variance of X_j and with h chosen by cross-validation. For the multivariate estimator we took $\psi_j = 1$; the optimal weights in this context are complex to estimate, and empirical versions might even introduce too much variability. We compared the two methods through their integrated squared errors computed on a domain that contained most of the data.

The results are shown in Table 3, where we report the median and the first and third quartiles of 200 integrated squared errors computed from 200 samples. We implemented the local-constant and local-linear versions of both methods, to which we refer below as estimators of order $\ell = 0$ and $\ell = 1$, respectively. It is well known that in practice, local-linear estimators are more variable than local-constant estimators, and the problem increases with dimension. This was reflected by our simulation results. As can be seen from the table, in the bivariate case the local-linear estimators sometimes brought significant improvement over the local-constant estimators. However, in our trivariate examples, the variability of local-linear estimators was too great for them to compete with local-constant estimators, and we report the results only for $\ell = 0$. Note too that computing the estimators for $\ell = 1$ is much more time consuming than for $\ell = 0$. For all these reasons, in general we recommend using $\ell = 0$. Table 3 also shows that overall, the single-index-based estimator performed significantly better than the purely multivariate estimator. We found this to be particularly true if the distributions of the components X_j differed strongly, as in our three dimensional setting; see also the real data example in the next section. In Table **??** in the supplementary material, we illustrate the effect of grouping by comparing our results with those obtained for standard estimators applied to the non grouped data.

6·2. *Real data example*

We conclude our numerical illustrations with a reanalysis of the bivariate example considered by Delaigle & Hall (2012), which was also used by Delaigle & Meister (2011) in the univariate case. The data come from the National Health and Nutrition Examination Survey and were collected in the US between 1999 and 2000. They are available at `www.cdc.gov/nchs/nhanes/nhanes1999-2000/nhanes99_00.htm`. These data are not grouped, and are therefore ideal for illustrating the effect of grouping on estimators. Since we are using them merely as an illustration, we follow the precedent in other papers of ignoring issues of sample weights originating in survey design.

As in the supplementary file of Delaigle & Hall (2012), let $X = (X_1, X_2)$, where X_1 is the age of a patient and X_2 is the total cholesterol measured in 100 mg per dL; let Y be the indicator, 0 or 1, of the presence of an antibody to hepatitis B virus core antigen in a patient serum or plasma. As in Delaigle & Hall (2012), our goal is to estimate

Table 3. *Simulation results for models (v) to (viii). The numbers show* $10^4 \times$ *median integrated squared error, and within square brackets the interquartile range, calculated from 200 simulated samples, where the data were grouped according to grouping A or B. We show results for the multivariate method and the single-index approach, and indicate the order of the local polynomial by* ℓ.

Model	ℓ	Grouping	$N = 5{,}000$				$N = 10{,}000$			
			Multivariate		Single-index		Multivariate		Single-index	
(v)	0	A	30	[21,46]	22	[12,36]	21	[14,37]	12	[7,19]
		B	52	[33,73]	35	[19,56]	33	[23,49]	16	[10,30]
	1	A	28	[21,38]	17	[8,31]	18	[13,25]	8	[5,12]
		B	41	[30,54]	21	[11,36]	27	[19,35]	10	[6,16]
(vi)	0	A	14	[10,17]	6	[4,10]	9	[6,12]	4	[2,5]
		B	19	[13,28]	12	[6,17]	14	[11,19]	7	[4,10]
	1	A	8	[6,11]	3	[2,5]	5	[4,8]	2	[1,3]
		B	11	[8,15]	6	[4,11]	8	[6,10]	3	[2,4]
(vii)	0	A	47	[36,61]	30	[21,53]	36	[26,44]	19	[13,28]
		B	67	[51,85]	55	[34,82]	50	[39,62]	30	[20,46]
(viii)	0	A	1127	[991,1270]	229	[159,324]	767	[691,851]	141	[96,211]
		B	1611	[1349,1840]	405	[263,610]	1206	[1099,1338]	246	[173,361]

$p(x) = E(Y \mid X = x)$ from these $N = 6{,}960$ observations, but where these authors use homogeneous pools, we grouped the data randomly, in groups of size 10 and 5.

To assess the quality of our estimator we repeated this 200 times. For each of the 200 randomly grouped testing samples created in this way, we computed the multivariate estimator derived in Section 3·1, and the single-index version in Section 3·2. In this example, the local-linear estimator suffered from too great variability and we used the local-constant estimator. As in the simulation Section we computed the integrated squared error. Here, the true curve p is unknown. As in Delaigle & Hall (2012), we approximated it by the standard multivariate nonparametric estimator computed from non-grouped data. In this example the distributions of X_1 and X_2 are quite different, and the single-index-based estimator performed considerably better than the purely multivariate estimator. In both cases considered, that is groups of size 10 and groups of size 5, the median, respectively the interquartile range, of the 200 integrated squared error values for the single-index estimator was about 50, respectively more than 10, times smaller than that for the multivariate estimator.

7. Imperfect tests

When the tests are imperfect, the test result $\tilde{Y}_j^* = 0$ or 1 potentially does not reflect the true status, Y_j^*. Consequently, the estimators of m, γ and q_0 introduced in the previous sections, with the unobserved Y_j^*s replaced by the $\tilde{Y}_j^*$s, are not consistent for m, γ and q_0. We follow Vansteelandt et al. (2000) and assume that the test accuracy does not depend on the n_js, and that the test result depends only on the true status. Let $1 - p_1 = \text{pr}(\tilde{Y}_j^* = 0 \mid Y_j^* = 0)$ and $1 - p_2 = \text{pr}(\tilde{Y}_j^* = 1 \mid Y_j^* = 1)$ be respectively the known test specificity and sensitivity, where p_1 and p_2 are less than $0 \cdot 5$.

We can estimate q_0 by the estimator $\check{q}$ defined in Section 5 of Delaigle & Meister (2011). Let $\tilde{Z}_j^* = 1 - \tilde{Y}_j^*$. As indicated by our calculations in the supplementary material, to estimate γ consistently we can take

$$\hat{\gamma} = \text{argmin}_\gamma \sum_{j=1}^{J} \sum_{i=1}^{n_j} \Big[\check{q}^{1-n_j} \, \tilde{Z}_j^* - \hat{g}_{\tilde{Z}U}(U_{ij}) + (1 - p_1 - p_2)\, \gamma^{\mathrm{T}} \big\{ V_{ij} - \hat{g}_{VU}(U_{ij}) \big\} \Big]^2,$$

where $\hat{g}_{VU}$ is as defined in Section 2·2 and $\hat{g}_{\tilde{Z}U}$ is a standard nonparametric regression estimator of $g_{\tilde{Z}U}(u) = E(q_0^{1-n_j} \tilde{Z}_j^* \mid U_{ij} = u)$ computed using the data $(U_{ij}, \check{q}^{1-n_j} \tilde{Z}_j^*)$ $(i = 1, \ldots, n_j;\ j = 1, \ldots, J)$.

Let $\hat{q}$ denote the maximum likelihood estimator from the supplementary material computed with the $\tilde{Z}_j^*$s instead of the Z_j^*s. To estimate m consistently, we can take

$$\hat{m}_{\mathrm{C}}(u) = C_0^{-1} \left\{ \hat{m}(u)\, \hat{M}_E - (\hat{M}_E - C_0)\, \hat{\gamma}^{\mathrm{T}} \hat{g}_{VU}(u) - p_2 \sum_{j=1}^{J} n_j\, \hat{\psi}_j \right\},$$

where $\hat{M}_E = \sum_{j \le J} n_j\, \hat{\psi}_j\, \hat{q}^{n_j - 1}$ and $C_0 = (1 - p_1 - p_2) \sum_{j \le J} n_j\, \hat{\psi}_j\, \check{q}^{n_j - 1}$. Here we take the $\hat{\psi}_j$s equal to $\hat{\psi}(\hat{q})$ in Section 5·1, replacing there the Y_j^*s by the $\tilde{Y}_j^*$s. Some adjustment is also needed for computing the centralised biased estimator at (9), where we should replace $\hat{m}$ by $\hat{m}_{\mathrm{C}}$, and $\hat{q}$ by $\check{q}$.

See the supplementary material for a practical illustration of the method.

Acknowledgement

Research supported by the Australian Research Council.

Supplementary material

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

A. Technical arguments for the single-index model

A·1. *Assumptions for Theorem* 2

Theorem 2 is derived under the following assumptions:

(I1) f is compactly supported, bounded away from 0 on A, and has two bounded, Hölder-continuous derivatives;

(I2) for $k = 0, 1, 2$, $f_\beta^{(k)}(\beta^{\mathrm{T}} x)$ is bounded uniformly in x and in $\beta \in B_{\mathrm{nhd}}$, $f_\beta(\beta^{\mathrm{T}} x)$ is bounded away from zero uniformly in $\beta \in B_{\mathrm{nhd}}$ and $x \in A$, and for each $\eta_1 > 0$ there exists $\eta_2 > 0$ such that $\sup_{\beta \in B_{\mathrm{nhd}}} |f''_\beta(\beta^{\mathrm{T}} x_1) - f''_\beta(\beta^{\mathrm{T}} x_2)| \le \eta_1$ for all $x_1, x_2 \in A$ with $\|x_1 - x_2\| \le \eta_2$;

(I3) the function g, at (11), is bounded and has two bounded, Hölder-continuous derivatives;

(I4) K is a symmetric, compactly supported, probability density with a Hölder continuous derivative;

(I5) the n_js are uniformly bounded and satisfy $J^{-1} \sum_{j \le J} I(n_j = k) \to \rho_k$, for each k, as $J \to \infty$;

(I6) the functions ψ_j and ϕ_j depend on j only through n_j, and in particular can be written as $\psi_j = \psi(\cdot \mid n_j)$ and $\phi_j = \phi(\cdot \mid n_j)$, respectively;

(I7) the functions ψ_j and ϕ_j, appearing in (16) and (19) and denoted below collectively by χ_j, are uniformly bounded and satisfy $\inf_j \inf_{u \in [\eta_1, 1-\eta_1]} \chi_j(u) > \eta_2$ for each $\eta_1 \in (0, 1/2)$, where $\eta_2 > 0$ depends on η_1, and they have two bounded derivatives and, for each $\eta_3 \in (0, 1/2)$ and some $\eta_4 > 0$, they satisfy $\sup_j |\psi_j(u) - \psi_j(v) - (u - v)\, \psi'_j(v) - 1/2\, (u - v)^2\, \psi''_j(v)| \le C_4(\eta_3)\, |u - v|^{2+\eta_4}$ and $\sup_j |\phi_j(u) - \phi_j(v) - (u - v)\, \phi'_j(v)| \le C_4(\eta_3)\, (u - v)^2$ whenever $u, v \in [\eta_3, 1 - \eta_3]$, where $C_4(\eta_3) < \infty$ depends on neither j nor N;

(I8) $h \in H_N \equiv [N^{\eta_5-(1/4)}, N^{-(1/6)-\eta_5}]$ for some $\eta_5 \in (0, 1/24]$, and $\beta \in B_N$ where B_N is any nonempty set of unit d-vectors β such that $\sup_{\beta\in B_N} \|\beta - \beta_0\| = O(N^{-(1/4)-\eta_6})$ for a value $\eta_6 > 0$.

We view N as the asymptotic parameter, and consider J and $n_1, \ldots, n_J$ to be functions of N. Thus, (I5) asserts that $\max_{j\le J} n_j(N) \le C$, where $C > 0$ is fixed, and implies that the long-run proportion of values of j such that $n_j = k$ equals ρ_k. In particular, the n_js can take no more than a finite number of fixed values, although within that range they can depend on N, and $N/J \to \sum_{k\ge1} \rho_k\, k$ as $N \to \infty$. Conditions (I1) and (I2) are conventional; they confer second-order smoothness properties on f and f_β, and permit us to avoid cases where the denominators in definitions of estimators of g are effectively estimators of zero. Assumption (I3) asks that g enjoy the same level of smoothness as the density of X; (I4) is a standard assumption on the kernel function, K; (I7) implies that the weight functions ψ_j and ϕ_j, which may depend on N, are uniformly bounded and smooth; and (I8) defines the regions around 0 and β_0, for h and β respectively, where we search for the minimum of the criterion $S_1(h, \beta)$ at (19).

A·2. *Notations used in statement of Theorem* 2

We can consider $g_\beta(t) = E\{g(\beta_0^{\mathrm{T}} X) \mid \beta^{\mathrm{T}} X = t\}$, as a functional of β. This fact is justified to be consistent with the definition in Section 3·3 in Section A·3. Under conditions (I2) and (I3) the functional has a continuous derivative in β:

$$g_\beta(\beta^{\mathrm{T}} x) = g(\beta_0^{\mathrm{T}} x) + (\beta - \beta_0)^{\mathrm{T}}\, g_1(x) + o(\|\beta - \beta_0\|)\,, \tag{A1}$$

where $g = g_{\beta_0}$ is as in (11), g_1 is a d-vector of functions, the components of $g_1(x)$ are bounded uniformly in $x \in A$, and the remainder in (A1) is of the stated order uniformly in such values of x.

Recall that X is distributed as a generic X_{ij}. Noting the definitions of ρ_k and $\phi(q \mid k)$ in (I5) and (I6), and letting $S_\rho = \sum_{k\ge1} \rho_k\, k$, define the d-vectors v_0 and W_1 by

$$v_0 = E\{I(X \in A)\, g(\beta_0^{\mathrm{T}} X)\, g_1(X)\} \sum_{k\ge1} \rho_k\, \phi(q_0 \mid k)\, q_0^{2(k-1)} / S_\rho\,, \tag{A2}$$

$$W_1 = \frac{1}{N} \sum\sum_{X_{ij}\in A} q_0^{n_j - 1}\, \phi_j(q_0)\, g_1(X_{ij})\, \epsilon_{ij}\,, \tag{A3}$$

where W_1 is asymptotically normal $\mathrm{N}(0, N^{-1}\,\Sigma_1)$, with Σ_1 defined at (A6).

Define the d-vector v_1 and the both positive semidefinite $d \times d$ matrices Σ_0 and Σ_1 by

$$v_1 = S_\rho^{-1} \sum_{k\ge1} \rho_k\, \phi(q_0 \mid k)\, \frac{k^2\, q_0^{2(k-1)}}{1 - q_0^k} E\Big[I(X \in A)\, g(\beta_0^{\mathrm{T}} X)\, g_1(X)\, \{1 - q_0^{k-1}\, g(\beta_0^{\mathrm{T}} X)\}\Big]\,, \tag{A4}$$

$$\Sigma_0 = S_\rho^{-1}\, E\Big\{g_1(X)\, g_1(X)^{\mathrm{T}}\, I(X \in A)\Big\} \sum_{k\ge1} k\, \rho_k\, \phi(q_0 \mid k)\, q_0^{2(k-1)}\,, \tag{A5}$$

$$\begin{aligned}\Sigma_1 = S_\rho^{-1} \Bigg\{ \sum_{k\ge1} \rho_k\, q_0^{2(k-1)}\, \phi(q_0 \mid k)^2 \Bigg(& k\, E\Big[I(X \in A)\, g_1(X)^2\, q_0^{k-1}\, g(\beta_0^{\mathrm{T}} X)\, \{1 - q_0^{k-1}\, g(\beta_0^{\mathrm{T}} X)\}\Big] \\ & + k\,(k-1)\, E\Big[I(X_1 \in A)\, I(X_2 \in A)\, g_1(X_1)\, g_1(X_2)^{\mathrm{T}} \times \Big\{q_0^{k-2}\, g(\beta_0^{\mathrm{T}} X_1)\, g(\beta_0^{\mathrm{T}} X_2) \\ & - 2q_0^{2k-3}\, g^2(\beta_0^{\mathrm{T}} X_1)\, g(\beta_0^{\mathrm{T}} X_2) + q_0^{2(k-1)}\, g(\beta_0^{\mathrm{T}} X_1)\, g(\beta_0^{\mathrm{T}} X_2)\Big\}\Big]\Bigg)\Bigg\}\,.\end{aligned} \tag{A6}$$

We assume that Σ_0 is nonsingular. Let c and c_1 be positive scalars defined by

$$c = S_\rho \Big/ \sum_{k\ge1} \rho_k\, k^2\, q_0^{k-1}\,(1-q_0^k)^{-1}\,, \quad c_1 = c^2\, S_\rho^{-1} \sum_{k\ge1} \rho_k\, k^2\, q_0^k\,(1-q_0^k)^{-1}\,. \tag{A7}$$

Recalling the definitions at (A2) and (A4), let Σ_2 be the $(d+1)\times(d+1)$ covariance matrix with

$$\Sigma_0^{-1/2}\left\{\Sigma_1 - c\left(v_0\, v_1^{\mathrm{T}} + v_1\, v_0^{\mathrm{T}}\right) + c_1\, v_0\, v_0^{\mathrm{T}}\right\}\Sigma_0^{-1/2}$$

in the upper $d\times d$ diagonal block, c_1 as the lowest diagonal element, that is the element in row $d+1$ and column $d+1$, and where $\Sigma_0^{-1/2}\,(c\,v_1 - c_1\,v_0)$ is the off-diagonal column.

A·3. *Technical arguments for Sections* 3·3 *and* 4·2

The definition of g_β in Section 3·3 requires justification, since we are asserting that for all β, not just for $\beta=\beta_0$, the left-hand side does not depend on j. To appreciate that our claim is correct, let $\mathcal{F}_1$ and $\mathcal{F}_2$ denote the sigma-fields generated by $\beta^{\mathrm{T}}X_{ij}$ and X_{ij}, respectively. Then $\mathcal{F}_1\subseteq\mathcal{F}_2$, and so, for any random variable V for which $E|V|<\infty$, it holds true that $E(V\,|\,\mathcal{F}_1) = E\{E(V\,|\,\mathcal{F}_2)\,|\,\mathcal{F}_1\}$. Taking $V=\epsilon_{ij}$ we deduce that

$$E\big(\epsilon_{ij}\;\big|\;\beta^{\mathrm{T}}X_{ij}\big) = E\big\{E(\epsilon_{ij}\mid X_{ij})\;\big|\;\beta^{\mathrm{T}}X_{ij}\big\} = E\big(\,0\;\big|\;\beta^{\mathrm{T}}X_{ij}\big) = 0\,,$$

where we have used the fact that $E(\epsilon_{ij}\mid X_{ij})=0$. Therefore,

$$\begin{aligned} E\big(Z_j^*\;\big|\;\beta^{\mathrm{T}}X_{ij}\big) &= E\big\{E(Z_j^*\mid X_{ij})\;\big|\;\beta^{\mathrm{T}}X_{ij}\big\} = E\big\{q_0^{n_j-1}\,g\big(\beta_0^{\mathrm{T}}X_{ij}\big)\;\big|\;\beta^{\mathrm{T}}X_{ij}\big\} \\ &= q_0^{n_j-1}\,E\big\{g\big(\beta_0^{\mathrm{T}}X_{ij}\big)\;\big|\;\beta^{\mathrm{T}}X_{ij}\big\} = q_0^{n_j-1}\,g_\beta\big(\beta^{\mathrm{T}}X_{ij}\big)\,, \end{aligned} \tag{A8}$$

In particular, the argument at (A8) shows that the definition of g_β at the beginning of Section A·2 is equivalent to that given in Section 3·3, and does not depend on j.

REFERENCES

BILDER, C. R. & TEBBS, J. M. (2009). Bias, efficiency, and agreement for group-testing regression models. *J. Stat. Comput. Simul.* **79**, 67–80.

BILDER, C. R., ZHANG, B., , SCHAARSCHMIDT, F. & TEBBS, J. M. (2010). binGroup: A Package for Group Testing. *The R Journal* **2**, 56–60.

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

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

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

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

FAHEY, J. W., OURISSON, P. J. & DEGNAN, F. H. (2006). Pathogen detection, testing, and control in fresh broccoli sprouts. *Nutrition Journal* **5**.

FAN, J. & GIJBELS, I. (1996). *Local polynomial modelling and its applications*, vol. 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

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

FLETCHER, J., RUSSELL, A. & BUTLER, R. (1999). Seed-borne cucumber mosaic virus in New Zealand lentil crops: yield effects and disease incidence. *New. Zeal. J. Crop. Hort.* **27**, 197–204.

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

GASTWIRTH, J. L. & JOHNSON, W. O. (1994). Screening with cost-effective quality control: Potential applications to hiv and drug testing. *J. Amer. Statist. Assoc.* **89**, 972–981.

Härdle, W., Hall, P. & Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21**, 157–178.

Härdle, W., Liang, H. & Gao, J. (2000). *Partially linear models.* Contributions to Statistics. Physica-Verlag, Heidelberg.

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

Hepworth, G. (2005). Confidence intervals for proportions estimated by group testing with groups of unequal size. *J. Agric. Biol. Envir. S.* **10**, 478–497.

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

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

Hung, M.-C. & 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. *Appl. Environ. Microbiol.* **73**, 2799–2805.

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

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

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

Speckman, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B* **50**, 413–436.

Swallow, W. H. (1985). Group testing for estimating infection rates and probabilities of disease transmission. *Phytopathology* **75**, 882–889.

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

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

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

Worlund, D. D. & Taylor, G. (1983). Estimation of disease incidence in fish populations. *Can. J. Fish. Aquat. Sci.* **40**, 2194–2197.

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