Estimation of conditional prevalence from group testing data with missing covariates

Aurore Delaigle[1], Wei Huang[1] and Shaoke Lei[2]
[1]School of Mathematics and Statistics and Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), University of Melbourne, Australia.
[2]Health Services, Murdoch Children's Research Institute and Health Services Research Unit, The Royal Children's Hospital, Australia.

Abstract: We consider estimating the conditional prevalence of a disease from data pooled according to the group testing mechanism. Consistent estimators have been proposed in the literature, but they rely on the data being available for all individuals. In infectious disease studies where group testing is frequently applied, the covariate is often missing for some individuals. There, unless the missing mechanism occurs completely at random, applying the existing techniques to the complete cases without adjusting for missingness does not generally provide consistent estimators, and finding appropriate modifications is challenging. We develop a consistent spline estimator, derive its theoretical properties, and show how to adapt local polynomial and likelihood estimators to the missing data problem. We illustrate the numerical performance of our methods on simulated and real examples.

Keywords: bandwidth, kernel, likelihood, local polynomial, pooled data, spline.

1 Introduction

Group testing is a technique that was originally applied to test for syphilis during the Second World War (Dorfman, 1943). It consists in pooling randomly individuals into groups, and instead of observing the disease status of each individual, $Y = 1$ (infected) or 0 (not infected), we observe only the infection status Y^* of each group. Typically, Y^* is obtained through a blood, serum, urine or other fluid test performed on the pooled fluid of the individuals from the group, which is less time and cost consuming than testing the fluid of each individual separately. The group testing technique is also applied in other contexts such as pollution detection in water, milk, etc (e.g. Nagi and Raggi, 1972, Wahed et al., 2006, Fahey et al., 2006 and Lennon, 2007) and plant disease or transgene (e.g. Fletcher et al., 1999, Montesinos-López et al., 2012 and Montesinos-López et al., 2013).

As pointed by Lindan et al. (2005), in contexts where resources are limited, for example in developing countries, testing of infectious diseases like chlamydia or gonorrhea is often not performed routinely due to the prohibitive cost of individual testing. One of the keys to wider use of group testing is the availability of methods for analysing data that have been grouped. Therefore, in the group testing literature, a lot of effort has been dedicated to developing methods for analysing group testing data, demonstrating that they give reasonable results, and assessing the loss incurred when using such methods compared to methods for non grouped data; see for example Verstraeten et al. (1998), Vansteelandt et al. (2000), Sarov et al. (2007), Lewis et al. (2012) and Zhang et al. (2013).

In group testing studies, an important quantity is the prevalence p of the disease conditional on an explanatory variable X. Many techniques have been suggested for estimating p, using either parametric (e.g. Vansteelandt et al., 2000, Xie, 2001, Chen et al., 2009 and Huang and Tebbs, 2009), semiparametric (e.g. Li and Xie, 2012, Wang et al., 2013, Wang et al., 2014 and Delaigle et al., 2014), or nonparametric approaches (e.g. Delaigle and Meister, 2011, Delaigle and Hall, 2012, Delaigle et al., 2014, Delaigle and Zhou, 2015 and Delaigle and Hall, 2015).

The existing techniques for estimating p all rely on the fact that X is observed for each individual. However, in practice, for reasons such as mechanism breakdowns or individuals refusing to answer some questions, X is often subject to missingness. This is particularly the case for sensitive studies such as those involving infectious disease, which are yet the ones where group testing techniques are the most often used. In the missing data literature, it is well known that in general, when data are missing, we cannot naively discard the incompletely observed individuals and apply existing techniques to the complete cases. In general, this approach, often referred to as the complete-case analysis, only provides satisfactory estimators when the data are missing completely at random (MCAR), that is, when the probability of being missing is unrelated to the variables of interest (e.g. Little and Rubin, 2014, chap. 3).

In practice, the data are often missing not completely at random, and applying the existing techniques to the complete cases produces biased estimators. Following Rubin (1976) and Little and Rubin (2014), this missing mechanism can be classified into two categories: 1) missing at random (MAR), where given the value of the completely observable variables, the probability that a variable is missing is independent of the value of the missing variable. In our notation, if we let $\Delta = 1$ if X is observed and $\Delta = 0$ if X is missing and if Y is fully observable, this means $\mathbb{P}(\Delta = 1|Y, X) = \mathbb{P}(\Delta = 1|Y)$; 2) missing not at random (MNAR), where, given the completely observable variables, the probability of the missingness still depends on the missing variable.

In the missing data literature (e.g. Little and Rubin, 2014), it is well known that both mechanisms (MAR and MNAR) can lead to consistent estimators. However this requires appropriate correction for missingness, which, under the MNAR mechanism, is often only possible if we have a parametric model for the way the data are missing. In practice, it is often not possible to determine whether the data are MAR or MNAR and we rarely know the true parametric model for the missing mechanism. Moreover, even when the missing mechanism is MNAR, estimators constructed under a MAR assumption can sometimes give more accurate results than those constructed under a MNAR assumption based on a wrong parametric model (see e.g. Rubin et al., 1995). Therefore, it is common in practice to use the MAR assumption. Estimators obtained under a MAR assumption are often considered to be a more useful starting point to analyse data with missing values than a MCAR assumption; see e.g. Little and Rubin (2014), page 19 and Molenberghs et al. (2014), page 281.

To our knowledge, despite the fact that there are often missing data in infectious studies where group testing techniques have the biggest potential to be used, statistical inference with missing data in the group testing context has not been conducted before. Consistent methods exist for missing data that are not grouped. In the parametric case, these are often based on a likelihood function (see Chapters 5–12 of Little and Rubin, 2014). In the nonparametric context, procedures are typically based on minimising an adjusted squared error function of the type $\sum Q\{Y_i, p(X_i)\}\Delta_i / \mathbb{P}(\Delta_i = 1|Y_i, X_i)\}$, where Q is a loss function and $\mathbb{P}(\Delta_i = 1|Y_i, X_i)$ is estimated from the data, which can be done easily if the non-grouped (X_i, Δ_i, Y_i)'s are observed and the MAR mechanism or a parametric model for $\mathbb{P}(\Delta_i = 1|Y_i, X_i)$ is assumed (e.g. Wang et al., 1998, Liang et al., 2007, Kim and Yu, 2012, Jiang et al., 2016). Imputation-based procedures also exist, where the missing data are replaced by values computed from observed data, and standard regression methods are directly applied to the "reconstructed data" (e.g. Chapter 4.5 of Little and Rubin, 2014, Oh and Scheuren, 1983 and Little, 1986), but they often produce biased estimators (Dempster and Rubin, 1983).

These methods cannot be applied in our context where the individual Y_i's are not observed. In this paper, under a particular MNAR assumption for our observed grouped data which derives from a standard MAR assumption on the individual non grouped (X_i, Y_i)'s, we develop consistent estimators of p in the case where the Y_i's are observed only in a pooled form. In section 3, we propose a nonparametric spline estimator. Unlike other methods, theoretical properties of spline estimators have never been studied in the group testing literature, and we rigorously establish consistency of our estimator in section 4.1. In section 5, we show how to adapt to our context two methods often used in the group testing case without missing data: a nonparametric local polynomial estimator (section 5.1) and a parametric likelihood estimator (section 5.2). We also show how to extend our ideas to the multivariate context (section 5.3). We apply our methods to real and simulated data in section 6, where we show that they work well in practice. Proofs and technical details are provided in the online appendix.

2 Model and data

We are interested in estimating the conditional prevalence $p(x) = \mathbb{E}(Y|X = x)$ of a disease given a covariate X, where Y denotes the disease status of a patient ($Y = 1$ if infected and $Y = 0$ if not infected). Instead of observing the disease status of each of n patients in a study, the individuals are pooled into J groups of sizes $n_1, \dots, n_J$, with $\sum_{j=1}^{J} n_j = n$, and only the disease status of each group is observed. The covariate X is observed at an individual level, but is missing for some of the individuals.

More formally, for $i = 1, \dots, n_j$, $j = 1, \dots, J$, let $Y_{i,j}$ and $X_{i,j}$ denote, respectively, the disease status (0 or 1) and the covariate of the ith individual in the jth group, and let $\Delta_{i,j} = 1$ if $X_{i,j}$ is observed and 0 otherwise. We assume that the $(\Delta_{i,j}, X_{i,j}, Y_{i,j})$'s are all independent and identically distributed (i.i.d.), and $\mathbb{P}(Y_{i,j} = 1|X_{i,j} = x) = \mathbb{E}(Y_{i,j}|X_{i,j} = x) = p(x)$. Instead of observing these individual data, for $i = 1, \dots, n_j$ and $j = 1, \dots, J$, we observe only $(\Delta_{i,j}, X_{i,j}, Y_j^*)$, where

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

When $\Delta_{i,j} = 0$, the notation $(\Delta_{i,j}, X_{i,j}, Y_j^*)$ means that Y_j^* is observed but $X_{i,j}$ is not.

We assume that

$$\mathbb{P}(\Delta_{i,j} = 1|Y_{i,j}, X_{i,j}) = \mathbb{P}(\Delta_{i,j} = 1|Y_{i,j}) \,, \tag{2.2}$$

which, if the data were not grouped (i.e., if the $Y_{i,j}$'s were observed), would correspond to a MAR assumption on the the $X_{i,j}$'s. In our case, the $Y_{i,j}$'s are not observed, but this assumption implies a MNAR one on the observed $(X_{i,j}, Y_j^*)$'s of a special type, which makes nonparametric estimation possible. An illustration of (2.2) is the situation where healthy individuals think they are safe because they feel fit, and are therefore less likely to report personal information, while individuals with the disease feel sick, thereby being more willing to provide information. Finally we let f_X and $\pi(x)$ denote, respectively, the density function of the $X_{i,j}$'s and $\mathbb{P}(\Delta_{i,j} = 1|X_{i,j} = x)$. Our goal is to construct consistent nonparametric and parametric estimators of p from the data $(X_{i,j}, \Delta_{i,j}, Y_j^*)$, $i = 1, \dots, n_j$, and $j = 1, \dots, J$.

3 Nonparametric spline estimator

In this section, we construct a consistent nonparametric spline estimator of $p(x) = \mathbb{P}(Y_{i,j} = 1|X_{i,j} = x)$ from the data $(X_{i,j}, \Delta_{i,j}, Y_j^*)$, $i = 1, \dots, n_j$, and $j = 1, \dots, J$, defined in section 2. We denote by $[a, b]$ a finite interval on which we estimate $p(x)$, where a and b are some fixed constants such that $-\infty < a < b < \infty$.

3.1 Main ideas

We start by describing the main ideas leading to our estimator. Let $Z_j^* = 1 - Y_j^* = \prod_{i=1}^{n_j}(1 - Y_{i,j})$, and $q_0 = \mathbb{E}\{1 - p(X_{i,j})\}$, and assume temporarily that q_0 is known (it can be easily estimated from the Z_j^*'s by the $\sqrt{n}$-consistent estimator $\widehat{q}_0$ of Delaigle and Meister, 2011; see Appendix A of the supplemental file). We know from Delaigle and Meister (2011) that $\mathbb{E}(Z_j^*|X_{1,j}, X_{2,j}, \dots, X_{n_j,j}) = \prod_{i=1}^{n_j}\{1 - p(X_{i,j})\}$ and

$$\mathbb{P}(Z_j^* = 1) = \mathbb{E}(Z_j^*) = q_0^{n_j}\ ,\ \mathbb{P}(Z_j^* = 1|X_{i,j}) = \mathbb{E}(Z_j^*|X_{i,j}) = q_0^{n_j - 1}\{1 - p(X_{i,j})\}\,, \tag{3.1}$$

so that

$$q_0^{1-n_j}\mathbb{P}(Z_j^* = 1|X_{i,j} = x) = \mathbb{E}(q_0^{1-n_j} Z_j^*|X_{i,j} = x) = 1 - p(x)\,. \tag{3.2}$$

Therefore, $1-p$ is a regression curve, and if the $(X_{i,j}, Z_j^*)$'s were all observed, as in Delaigle et al. (2014) we could estimate p using a standard nonparametric regression estimator constructed from the $(X_{i,j}, q_0^{1-n_j} Z_j^*)$'s.

However we cannot compute this estimator in our case where some of the $X_{i,j}$'s are missing; as noted in the introduction, applying it to the complete cases only would lead to a non consistent estimator. On the other hand, a regression curve that can be consistently estimated from our data using a standard nonparametric estimator is

$$g(x) \equiv \mathbb{E}(q_0^{1-n_j} Z_j^* | X_{i,j} = x, \Delta_{i,j} = 1) = q_0^{1-n_j} \mathbb{P}(Z_j^* = 1 | X_{i,j} = x, \Delta_{i,j} = 1) \,, \qquad (3.3)$$

since it needs only the data for which $\Delta_{i,j} = 1$, i.e. for which $X_{i,j}$ has been observed. We will see how to estimate g by spline and kernel methods in sections 3.2 and 5.1, but for this to be useful, it remains to see if we can express p in terms of g.

In order to do this, applying the result $\mathbb{P}(A, B|C) = \mathbb{P}(A|B, C)\mathbb{P}(B|C)$ to the events $A = \{Z_j^* = 1\}$, $B = \{\Delta_{i,j} = 1\}$ and $C = \{X_{i,j} = x\}$, we write

$$\mathbb{P}(Z_j^* = 1 | X_{i,j} = x, \Delta_{i,j} = 1)\mathbb{P}(\Delta_{i,j} = 1 | X_{i,j} = x) = \mathbb{P}(Z_j^* = 1, \Delta_{i,j} = 1 | X_{i,j} = x) \,.$$

Then applying this result again, but swapping the role of A and B, we get

$$\mathbb{P}(\Delta_{i,j} = 1, Z_j^* = 1 | X_{i,j} = x) = \mathbb{P}(\Delta_{i,j} = 1 | X_{i,j} = x, Z_j^* = 1)\mathbb{P}(Z_j^* = 1 | X_{i,j} = x) \,.$$

Combining those two equalities, using (3.2) and (3.3), and recalling the notation $\pi(x) = \mathbb{P}(\Delta_{i,j} = 1 | X_{i,j} = x)$ from page 6, we deduce that

$$\{1 - p(x)\}\mathbb{P}(\Delta_{i,j} = 1 | X_{i,j} = x, Z_j^* = 1) = g(x)\pi(x) \,. \qquad (3.4)$$

While we have expressed p in terms of g, (3.4) involves two unknown functions, $\mathbb{P}(\Delta_{i,j} = 1 | X_{i,j} = x, Z_j^* = 1)$ and $\pi(x)$. Since they depend on the $X_{i,j}$'s, some of which are missing, in order to be able to estimate these functions, we first need to express them in a way that depends only on the non missing data.

For the first function, using repeatedly the assumption at (2.2) and the fact that the $(X_{i,j}, \Delta_{i,j}, Y_{i,j})$'s are i.i.d., we prove in Appendix A.2 of the supplemental file that

$$\mathbb{P}(\Delta_{i,j} = 1 | X_{i,j} = x, Z_j^* = 1) = \mathbb{P}(\Delta_{i,j} = 1 | Z_j^* = 1) \equiv p_0 \,. \qquad (3.5)$$

Thus, $\mathbb{P}(\Delta_{i,j} = 1|X_{i,j} = x, Z_j^* = 1) = p_0$ is constant in x and depends only on the $\Delta_{i,j}$'s and the Z_j^*'s, which are observed. The function π is more problematic. Using the result $\mathbb{P}(A, B|C) = \mathbb{P}(A|B, C)\mathbb{P}(B|C)$ and (2.2), we show in Appendix A.2 that

$$\pi(x) = \mathbb{P}(\Delta_{i,j} = 1|X_{i,j} = x) = p_0\{1 - p(x)\} + p_1 p(x)\,, \tag{3.6}$$

where $p_1 = \mathbb{P}(\Delta_{i,j} = 1|Y_{i,j} = 1)$. We will see below how to estimate p_0 and p_1, but a problem with the expression for π at (3.6) is that we are estimating π because we need it in order to estimate p using (3.4), and (3.6) depends on p, which we cannot estimate since it depends on π. However, plugging (3.6) into (3.4), we get $\{1-p(x)\}p_0 = g(x)\big[p_0\{1-p(x)\}+p_1 p(x)\big]$, so that

$$p(x) = \{1 - g(x)\}/\{1 + (p_1 p_0^{-1} - 1)g(x)\}\,. \tag{3.7}$$

(Under Assumption (A4) from section 4.1, the denominator of (3.7) does not vanish.)

In (3.7), p is expressed in terms of g, p_0 and p_1 only. In particular, estimating p no longer requires estimating π, and we have resolved the circular argument. To estimate p_0, using (3.5) we write $p_0 = \mathbb{E}(\Delta_{i,j} Z_j^*)/\mathbb{E}(Z_j^*)$. Thus, using (3.1), we can estimate it by $\widehat{p}_0 = \max(\tilde{p}_0, c_0)$, where $\tilde{p}_0 = \sum_{j=1}^J \sum_{i=1}^{n_j} \Delta_{i,j} Z_j^* / \sum_{j=1}^J \sum_{i=1}^{n_j} Z_j^*$ and $c_0 > 0$ is a small number, taking the convention that $0/0 = 0$. Here we use the constant c_0 to ensure that $\widehat{p}_0$, which is a denominator of our estimator, is always strictly larger than zero. We will discuss how to choose c_0 in section 6.

To estimate p_1, which depends on the unobserved $Y_{i,j}$'s, we show in Appendix A.2 that

$$p_1 = \mathbb{P}(\Delta_{i,j} = 1, Y_{i,j} = 1)/\mathbb{P}(Y_{i,j} = 1) = (\mu_\Delta - p_0 q_0)/(1 - q_0)\,, \tag{3.8}$$

where $\mu_\Delta \equiv \mathbb{E}(\Delta_{i,j}) = \mathbb{P}(\Delta_{i,j} = 1)$ can be estimated by the empirical mean $\bar{\Delta} = n^{-1} \sum_{j=1}^J \sum_{i=1}^{n_j} \Delta_{i,j}$. Thus we can estimate p_1 by $\widehat{p}_1 = \max(\tilde{p}_1, c_0)$ where $\tilde{p}_1 = (\bar{\Delta} - \widehat{p}_0 \widehat{q}_0)/(1 - \widehat{q}_0)$, where again, we use c_0 to ensure that our estimator of p_1 is always strictly greater than zero.

Finally, recalling (3.7), we can estimate $p(x)$ by

$$\widehat{p}(x) = \{1 - \widehat{g}(x)\}/\{1 + (\widehat{p}_1\widehat{p}_0^{-1} - 1)\widehat{g}(x)\}\,, \tag{3.9}$$

where $\widehat{g}$ is an estimator of g. We derive two estimators of g in sections 3.2 and 5.1.

3.2 Penalised spline estimator

In this section we construct a penalised spline estimator of g at (3.3); using (3.9) we deduce a spline estimator of p. A spline function is a piecewise polynomial of a degree $d \geq 1$, where the pieces are such that the spline is $d-1$ times continuously differentiable. The pieces are defined on intervals whose extremities are called knots. More formally, for any integer $d \geq 1$, given an interval $[a, b]$ on the real line and a sequence $\boldsymbol{t} = \{t_{-d}, t_{-d+1}, \ldots, t_{K+d}, t_{K+d+1}\}$ of knots such that $t_{-d} = \ldots = t_0 = a < t_1 < t_2 < \ldots < t_K < b = t_{K+1} = \ldots = t_{K+d+1}$, with K a positive integer, a spline of degree d (equivalently, of order $d+1$) with knots $\boldsymbol{t}$ is a function s that is $(d-1)$ times continuously differentiable on $[a, b]$, i.e. $s \in C^{d-1}[a, b]$, and is such that on each $[t_i, t_{i+1}]$, $i = 0, \ldots, K$, s is a polynomial of degree d. The extreme knots are repeated for technical convenience; see Appendix E.1 of the supplemental file where we summarise some useful facts about splines. We denote the class of such spline functions by $S_d(\boldsymbol{t})$.

If the $(X_{i,j}, Y_{i,j})$'s were observed, then since $p(x) = \mathbb{E}(Y_{i,j}|X_{i,j} = x)$ is a regression curve, we could estimate it by the standard penalised spline estimator of degree d:

$$\widehat{s}(x) = \underset{s(x)\in S_d(\boldsymbol{t})}{\operatorname{argmin}} \left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}\{Y_{i,j} - s(X_{i,j})\}^2 + \lambda\int_a^b \{s^{(\ell)}(x)\}^2\,dx\right], \tag{3.10}$$

where the first term is a residual sum of squares, the second is a smoothness penalty term, $s^{(\ell)}$ denotes the ℓth order derivative of s with $0 < \ell < d$, and $\lambda > 0$ is a smoothing parameter controlling the strength of the penalty; see e.g. Ruppert et al. (2003) and Claeskens et al. (2009).

Of course, in our case we cannot compute (3.10) since we do not observe the $Y_{i,j}$'s and some of the $X_{i,j}$'s are missing. As suggested in section 3.1 (in particular, see (3.9)), instead we focus on the curve $g(x) = \mathbb{E}(q_0^{1-n_j} Z_j^* | X_{i,j} = x, \Delta_{i,j} = 1)$, which depends only on our observed data. Mimicking (3.10), replacing there $Y_{i,j}$ by $q_0^{1-n_j} Z_j^*$ and using only the data for which $\Delta_{i,j} = 1$, we propose the following penalised spline estimator of $g(x)$:

$$\widehat{g}(x) = \underset{s(x)\in S_d(\boldsymbol{t})}{\operatorname{argmin}} \left[\sum_{j=1}^{J} \sum_{i=1}^{n_j} \{\widehat{q}_0^{1-n_j} Z_j^* - s(X_{i,j})\}^2 \Delta_{i,j} + \lambda \int_a^b \{s^{(\ell)}(x)\}^2 \, dx \right], \tag{3.11}$$

where $\widehat{q}_0$ is the estimator of q_0 defined in Appendix A.1 of the supplemental file.

A refined version of $\widehat{g}(x)$ can be defined in the case where the group sizes n_j are unequal. As pointed by Delaigle et al. (2014), in that case data from different groups are not identically distributed, and should not contribute to the estimator equally. Instead, each group should be given a weight that depends on its size. Motivated by this, we introduce the following weighted version of our estimator at (3.11):

$$\widehat{g}_s(x) = \underset{s(x)\in S_d(\boldsymbol{t})}{\operatorname{argmin}} \left[\phi(\widehat{q}_0) \sum_{j=1}^{J} \sum_{i=1}^{n_j} \{\widehat{q}_0^{1-n_j} Z_j^* - s(X_{i,j})\}^2 \varphi_j(\widehat{q}_0) \Delta_{i,j} + \lambda \int_a^b \{s^{(\ell)}(x)\}^2 \, dx \right], \tag{3.12}$$

where φ_j, $j = 1, \ldots, J$, are smooth positive functions similar to the weights of Delaigle et al. (2014) and which can be chosen using the method described in section 4.2, and where $\phi = n/(\sum_{j=1}^{J} n_j \varphi_j)$ is a normalising factor. Expressed in this form, the estimator $\widehat{g}_s(x)$ at (3.12) may seem difficult to compute, but in the next two paragraphs we derive a closed form expression for $\widehat{g}_s(x)$, which is given at (3.14) below.

To find an analytic expression for $\widehat{g}_s$, we use the fact that splines can be written explicitly as a linear combination of spline basis functions. There are many theoretically equivalent different ways to choose a spline basis, but some have more attractive properties. We use the B-spline basis, which is one of the most popular ones and for which lots of theoretical results exist in the literature (see e.g. Schumaker, 1981 and De Boor, 2001). We denote the B-spline basis functions of the spline space $S_d(\boldsymbol{t})$ by $N_{i,d+1}(x)$, for $i = -d, \dots, K$ (we recall their definition in Appendix E.1). Any function $s \in S_d(\boldsymbol{t})$ can be expressed as $s(x) = \sum_{i=-d}^{K} \beta_i N_{i,d+1}(x)$, where the coefficients $\beta_i \in \mathbb{R}$.

In this notation, solving (3.12) is equivalent to minimising, w.r.t. the β_i's $\in \mathbb{R}$,

$$\phi(\widehat{q}_0)\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left\{\widehat{q}_0^{\,1-n_j} Z_j^* - \sum_{k=-d}^{K} \beta_k N_{k,d+1}(X_{i,j})\right\}^2 \varphi_j(\widehat{q}_0)\Delta_{i,j} + \lambda \int_a^b \left\{\sum_{k=-d}^{K} \beta_k N^{(\ell)}_{k,d+1}(x)\right\}^2 dx\,. \tag{3.13}$$

Let $\boldsymbol{\Delta} = \mathrm{diag}(\Delta_{1,1}, \Delta_{2,1}, \dots, \Delta_{n_J,J})$, $\widehat{\boldsymbol{Q}} = \mathrm{diag}(\widehat{q}_0^{\,1-n_1} \overset{\times n_1}{\cdots}, \cdots, \widehat{q}_0^{\,1-n_J} \overset{\times n_J}{\cdots})$ and $\boldsymbol{\Phi}(\widehat{q}_0) = \mathrm{diag}\big(\varphi_1(\widehat{q}_0) \overset{\times n_1}{\cdots}, \cdots, \varphi_J(\widehat{q}_0) \overset{\times n_J}{\cdots}\big)$ be diagonal matrices, where $\overset{\times n_j}{\cdots}$ means that the quantity is repeated n_j times. Finally, let $\boldsymbol{Z}^* = (Z_1^* \overset{\times n_1}{\cdots}, \cdots, Z_J^* \overset{\times n_J}{\cdots})^\top$. Then the solution $\widehat{g}_s(x)$ to (3.12), which is found by minimising (3.13), can be expressed as

$$\widehat{g}_s(x) = \phi(\widehat{q}_0)\boldsymbol{N}^\top(x)\{\phi(\widehat{q}_0)\boldsymbol{\mathcal{N}}^\top \boldsymbol{\Phi}(\widehat{q}_0)\boldsymbol{\Delta}\boldsymbol{\mathcal{N}} + \lambda \boldsymbol{D}_\ell\}^{-1}\boldsymbol{\mathcal{N}}^\top \boldsymbol{\Phi}(\widehat{q}_0)\boldsymbol{\Delta}\widehat{\boldsymbol{Q}}\boldsymbol{Z}^*\,, \tag{3.14}$$

where $\boldsymbol{N}^\top(x) = \{N_{-d,d+1}(x), \dots, N_{K,d+1}(x)\}$, $\boldsymbol{\mathcal{N}} = \{\boldsymbol{N}(X_{1,1}), \dots, \boldsymbol{N}(X_{n_J,J})\}^\top$ is an $n \times (K+d+1)$ matrix and $\boldsymbol{D}_\ell$ is a matrix with elements $(\boldsymbol{D}_\ell)_{ij} = \int_a^b N^{(\ell)}_{i,d+1}(x) N^{(\ell)}_{j,d+1}(x)\, dx$, for $i, j = -d, \dots, K$.

The expression for $\widehat{g}_s(x)$ at (3.14) can be computed easily since $\boldsymbol{N}, \boldsymbol{\mathcal{N}}, \boldsymbol{D}_\ell, \boldsymbol{\Delta}$, as well as $\boldsymbol{Z}^*$ defined above (3.1), depend only on the B-spline basis and on the observed data, both of which are known. Recall too that $\widehat{q}_0$ is defined in Appendix A, and that ϕ and $\boldsymbol{\Phi}$ both depend on the φ_j's, which are computed in section 4.2. Finally, the number of knots K and the smoothing parameter λ can be chosen by cross-validation as in section 6.1. Using (3.7), we deduce an estimator of $p(x)$ by taking

$$\widehat{p}_s(x) = \{1 - \widehat{g}_s(x)\}\big/\big\{1 + (\widehat{p}_1\widehat{p}_0^{-1} - 1)\widehat{g}_s(x)\big\}\,. \tag{3.15}$$

Remark 1. *Throughout the paper, the indices of vectors and matrices do not necessarily start at 1. Each time they do not start at 1 we shall indicate their range.*

4 Asymptotic properties of $\widehat{p}_s$ and optimal φ_j's

4.1 Asymptotic properties of $\widehat{p}_s$

We derive asymptotic properties of $\widehat{p}_s(x)$ under the following assumptions, which we discuss in Appendix A.3 of the supplemental file:

Assumption A

(A1) p is $d+1$ times continuously differentiable on $[a,b]$.

(A2) $\sup_{x\in[a,b]} |p^{(j)}(x)| < \infty$ for $j = 1, \ldots, d+1$.

(A3) d and ℓ are positive integers such that $0 < \ell < d < \infty$.

(A4) There exists a constant $c_p > 0$ such that $\min(p_0, p_1) > c_p$.

(A5) $0 < \pi_{\min} \equiv \inf_{x\in[a,b]} \pi(x) \leq \sup_{x\in[a,b]} \pi(x) \equiv \pi_{\max} < 1$.

(A6) $K \to \infty$ as $n \to \infty$ and $K(\log K)^2/n \to 0$ as $n \to \infty$.

(A7) There exists a constant $1 \leq M < \infty$ such that $\delta / \min_{0\leq i\leq K} \delta_i \leq M$, where $\delta_i = t_{i+1} - t_i$ for $i = -d, \ldots, K+d$, and $\delta = \max_{0\leq i\leq K} \delta_i$. Furthermore, $\max_{0\leq i\leq K} |\delta_{i+1} - \delta_i| = o(\delta)$.

(A8) f_X is twice continuously differentiable on $[a,b]$, $\sup_{x\in[a,b]} |f_X^{(k)}(x)| < \infty$, for $k = 1, 2$, and there exist constants $f_{\min}$ and $f_{\max}$ such that $0 < f_{\min} \leq \inf_{x\in[a,b]} f_X(x) \leq \sup_{x\in[a,b]} f_X(x) \leq f_{\max} < \infty$.

(A9) $\sup_{j=1,\ldots,J} n_j < \infty$ and $0 < q_0 < 1$.

(A10) There exist constants ϕ_1 and ϕ_2 such that $0 < \phi_1 \leq \inf_{j=1,\ldots,J} \varphi_j(q_0) < \sup_{j=1,\ldots,J} \varphi_j(q_0) \leq \phi_2 < \infty$.

(A11) $0 \leq \inf_{n\in\mathbb{N}} \lambda K^{2\ell}/n \leq \sup_{n\in\mathbb{N}} \lambda K^{2\ell}/n < \infty$ and $\lambda = o(K)$.

(A12) For $j = 1, \ldots, J$, φ_j depends on j only through n_j, is uniformly bounded and has a bounded derivative on $(0,1)$.

The next theorem establishes asymptotic properties of the estimator $\widehat{p}_s$ at (3.15).

Theorem 1. *Under Assumption* **A**, *we have, for any* $x \in [a, b]$,

$$\widehat{p}_s(x) - p(x) = B_\delta(x)\xi(x) + V_{n,\delta}^{1/2}(x)\xi(x)\Psi_{n,\delta}(x) + o_p\big(\delta^{d+1}\big) + o_p\big\{(n\delta)^{-1/2}\big\}\,, \tag{4.1}$$

where $\Psi_{n,\delta}(x) \xrightarrow{D} N(0,1)$ *as* $n \to \infty$, $\xi(x) = \pi^2(x)/(p_0 p_1)$,

$$B_\delta(x) = \frac{g^{(d+1)}(x)}{(d+1)!}\sum_{i=0}^{K} \mathbb{1}_{[t_i, t_{i+1})}(x)\delta_i^{d+1} B_{d+1}\left(\frac{x - t_i}{t_{i+1} - t_i}\right), \tag{4.2}$$

$$V_{n,\delta}(x) = \frac{\sum_{j=1}^{J} n_j \varphi_j^2(q_0)}{\delta\{\sum_{j=1}^{J} n_j \varphi_j(q_0)\}^2}\boldsymbol{N}^\top(x)\boldsymbol{H}_{n,\lambda}^{-1}\boldsymbol{G}_{g,j}\boldsymbol{H}_{n,\lambda}^{-1}\boldsymbol{N}(x) \asymp (n\delta)^{-1}\,, \tag{4.3}$$

and where

$$B_{d+1}(x) = \sum_{i=0}^{d+1}\frac{1}{i+1}\sum_{j=0}^{i}(-1)^j\left(\begin{array}{c} i \\ j \end{array}\right)(x+j)^{d+1} \tag{4.4}$$

is the Bernoulli polynomial of degree $d + 1$,

$$\boldsymbol{G}_{g,j} = \delta^{-1}\int_a^b \pi(y)[g(y)\{q_0^{1-n_j} - g(y)\}]\boldsymbol{N}(y)\boldsymbol{N}^\top(y) f_X(y)\,dy\,, \tag{4.5}$$

$$\boldsymbol{H}_{n,\lambda} = \lambda(n\delta)^{-1}\boldsymbol{D}_\ell + \delta^{-1}\int_a^b \pi(y)\boldsymbol{N}(y)\boldsymbol{N}^\top(y) f_X(y)\,dy\,. \tag{4.6}$$

Remark 2. *Abusing terminology and following the literature, throughout we will often refer to* $B_\delta\xi$ *and* $V_{n,\delta}\xi^2$ *as asymptotic bias and variance. While they do play the role of bias and variance, it is clear from the above expressions that they are in fact asymptotic expressions coming from our convergence in distribution results.*

The asymptotic bias of our estimator is equal to $B_\delta(x)\xi(x)$, and the term $\xi(x)$ is the effect that missing data have on bias (grouping has no effect); it reduces to 1 when there are no missing data. Our asymptotic variance $V_{n,\delta}(x)\xi^2(x)$ is affected by both grouping and missingness. See Appendix A.4 for a short discussion about the non grouped i.i.d. case. Note that $B_\delta(x)$ can vanish at some x (see Appendix A.4), in which case to find the dominating part of the bias, one would need to investigate the higher order terms in $o_p(\delta^{d+1})$.

When $B_\delta(x) \neq 0$, we see from (4.2) that $B_\delta(x) \asymp \delta^{d+1}$. Combining with (4.3), in that case the fastest rate of convergence of our estimator is $n^{-(d+1)/(2d+3)}$, obtained by taking $\delta^{d+1} \asymp (n\delta)^{-1/2}$, i.e. $\delta \asymp n^{-1/(2d+3)}$, so that $K \asymp n^{1/(2d+3)}$ (recall Assumption (A7)). For some d and x such that $B_\delta(x) = 0$, our estimator converges faster than the $n^{-(d+1)/(2d+3)}$ rate if we take δ^{d+1} an order of magnitude larger than $(n\delta)^{-1/2}$, but the exact rate and order depend on the exact rate of the $o_p(\delta^{d+1})$ term. Comparing our results with those of Stone (1980), we can see that our estimator reaches the optimal rate of the standard nonparametric regression estimation problem with non grouped and non missing data, which implies that our estimator is rate-optimal for our problem too.

4.2 Choosing the φ_j's

We deduce from Theorem 1 that, to first order, the φ_j's influence only the variance term $\xi^2(x)V_{n,\delta}(x)$. Together with the fact that we wish to use the same weights for all x, this motivates us to choose the φ_j's that minimise $\int_a^b \xi^2(x)V_{n,\delta}(x)\,dx$. In Appendix A.5, we show that these are given by $\varphi_j^*(q_0) = 1/V_j$, where $V_j = \int_a^b \pi^4(x)\boldsymbol{N}^\top(x)\tilde{\boldsymbol{H}}_{n,\lambda}^{-1}\tilde{\boldsymbol{G}}_{g,j}\tilde{\boldsymbol{H}}_{n,\lambda}^{-1}\boldsymbol{N}(x)\,dx$, with π defined at the end of section 2, $\tilde{\boldsymbol{H}}_{n,\lambda} = \lambda\boldsymbol{D}_\ell/n + \int_a^b \boldsymbol{N}(y)\boldsymbol{N}^\top(y)\pi(y)f_X(y)\,dy$ and $\tilde{\boldsymbol{G}}_{g,j} = \int_a^b g(y)\{q_0^{1-n_j} - g(y)\}\boldsymbol{N}(y)\boldsymbol{N}^\top(y)\pi(y)f_X(y)\,dy$.

In practice, q_0, g, π and f_X in the definition of the V_j's are unknown and need to be estimated to produce an estimator $\widehat{\varphi}_j^*$ of each φ_j^*. We can estimate q_0 by $\widehat{q}_0$ in Appendix A, and g by a pilot estimator $\widehat{g}_{\text{pilot}}$, which we take equal to $\widehat{g}_s$ with all the φ_j's equal to 1. To estimate π, note from (3.6) and (3.7) that $\pi(x) = p_1/\{1 + (p_1/p_0 - 1)g(x)\}$. This motivates us to estimate $\pi(x)$ by $\widehat{\pi}(x) = \widehat{p}_1/\{1 + (\widehat{p}_1/\widehat{p}_0 - 1)\widehat{g}_{\text{pilot}}(x)\}$. Estimating f_X is more difficult. Recall that when there are no missing data, this density can be consistently estimated by the kernel density estimator $(nh)^{-1}\sum_{j=1}^J\sum_{i=1}^{n_j} L\{(x - X_{i,j})/h\}$, where L is a kernel function and $h > 0$ is a bandwidth (see e.g. Fan and Gijbels, 1996). In our case, all we can calculate is

$$\widehat{f}_{X,\mathrm{obs}}(x) = (nh)^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} L\{(x - X_{i,j})/h\} \Delta_{i,j} \, , \tag{4.7}$$

which, using a Taylor expansion, is easily seen to be a consistent estimator of $\pi(x) f_X(x)$, but not of $f_X(x)$. Instead of being an issue, this is actually an advantage in our case because looking at the definition of $\tilde{\boldsymbol{G}}_{g,j}$ above, where f_X and π appear, we can see that all we need to estimate is precisely $\pi(x) f_X(x)$.

Combining the above calculations, we take

$$\widehat{\varphi}_j^*(\widehat{q}_0) = 1 \Big/ \int_a^b \widehat{\pi}^4(x) \boldsymbol{N}^\top(x) \widehat{\boldsymbol{H}}^{-1} \widehat{\boldsymbol{G}}_j \widehat{\boldsymbol{H}}^{-1} \boldsymbol{N}(x) \, dx \, ,$$

with $\widehat{\boldsymbol{H}} = \lambda \boldsymbol{D}_\ell / n + \int_a^b \boldsymbol{N}(y) \boldsymbol{N}^\top(y) \widehat{f}_{X,\mathrm{obs}}(y) \, dy$ and $\widehat{\boldsymbol{G}}_j = \int_a^b \widehat{g}_{\mathrm{pilot}}(y) \{\widehat{q}_0^{1-n_j} - \widehat{g}_{\mathrm{pilot}}(y)\} \boldsymbol{N}(y) \boldsymbol{N}^\top(y) \widehat{f}_{X,\mathrm{obs}}(y) \, dy$.

Of course when the groups sizes n_j are all equal, the $\widehat{\varphi}_j^*(q_0)$'s are all equal, and since they are invariant to scale, we set them equal to 1.

4.3 Confidence interval for $p(x)$

It is well known in the nonparametric regression literature that constructing genuine data-driven confidence intervals for a regression curve is extremely complex. In particular, no matter what approach is used, at some stage one needs to estimate quantities that are not easy to estimate and/or select several tuning parameters in a very subtle way (see e.g. Krivobokova et al., 2010 and the discussion in Appendices C1 and C2 of Delaigle et al., 2015).

To understand the difficulty, note that it follows from Theorem 1 that, for $x \in [a, b]$,

$$\frac{\widehat{p}_s(x) - p(x) - B_\delta(x)\xi(x)}{V_{n,\delta}^{1/2}(x)\xi(x)} \xrightarrow{D} N(0,1) \, , \text{ as } n \to \infty \, .$$

We can use this result to construct asymptotic confidence intervals for $p(x)$ based on the asymptotic normal distribution, but since $B_\delta(x), \xi(x)$ and $V_{n,\delta}(x)$ are unknown, they have to be estimated nonparametrically for this expression to be practical. However, estimating $B_\delta(x)$ impacts the coverage rate of the confidence interval, and removing this coverage error is difficult (see e.g. Delaigle et al., 2015).

The impact of estimating $B_\delta(x)\xi(x)$ can be avoided if the estimator $\widehat{p}_s(x)$ is undersmoothed, so that the term $B_\delta(x)\xi(x)$ can be neglected in the asymptotic normality result displayed in the previous paragraph. Specifically, if we take $K \asymp n^r$ with $r > 1/(2d+3)$, then we see from Theorem 1 that $\{\widehat{p}_s(x) - p(x)\}/\{V_{n,\delta}^{1/2}(x)\xi(x)\} \overset{D}{\to} N(0,1)$, as $n \to \infty$, and to construct an asymptotic confidence interval for $p(x)$ from this result we only need to estimate $\xi(x)$ and $V_{n,\delta}(x)$. However, in practice, proposing a genuine data-driven method to choose the smoothing parameters involved while maintaining the coverage rate of the confidence intervals is again a very hard problem, even in standard settings with non grouped data.

A third possibility for constructing a confidence interval is to use bootstrap, but again this requires the delicate choice of genuinely data-driven tuning parameters (see e.g. Delaigle et al., 2015). In summary, while it is possible to construct nonparametric confidence intervals for $p(x)$, the issue is complex and typically requires whole papers dedicated entirely to the issue (see e.g. Krivobokova et al., 2010 and Delaigle et al., 2015).

5 Other methods

In the group testing literature without missing data, two estimators of p are often in use: a nonparametric local polynomial estimator and a parametric likelihood-based one. Although the main focus of this paper is on nonparametric spline-based methods, which have never been studied rigorously before in the group testing literature, here we show how the two usual estimators can be adapted to our setting. We also extend our nonparametric procedure to the multivariate setting.

5.1 Local polynomial estimator

We start by showing how to construct an estimator of p based on local polynomial techniques. Before we introduce the general local polynomial estimators that can be used in our case, we start by showing how to construct a local constant estimator, also referred to as a Nadaraya-Watson estimator. If the $(X_{i,j}, Y_{i,j})$'s were available, we could compute the Nadaraya-Watson estimator (Fan and Gijbels, 1996) of the regression curve $p(x) = \mathbb{E}(Y|X = x)$, defined by

$$\widetilde{p}_{\mathrm{NW}}(x) = \frac{n^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j} Y_{ij} L_h(X_{i,j} - x)}{n^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j} L_h(X_{i,j} - x)}\,,$$

where $h > 0$ denotes a bandwidth, L is a kernel function and $L_h(\cdot) = h^{-1}L(\cdot/h)$. Recalling the definition of g at (3.3), this suggests estimating $g(x)$ by

$$\widehat{g}_{\mathrm{NW}}(x) = \frac{n^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j} \widehat{q}_0^{\,1-n_j} Z_j^* \Delta_{i,j} L_h(X_{i,j} - x)}{n^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j} \Delta_{i,j} L_h(X_{i,j} - x)}\,. \tag{5.1}$$

Using (3.7), we estimate $p(x)$ by $\widehat{p}_{\mathrm{NW}}(x) = \{1 - \widehat{g}_{\mathrm{NW}}(x)\}/\{1 + (\widehat{p}_1\widehat{p_0}^{-1} - 1)\widehat{g}_{\mathrm{NW}}(x)\}$.

To understand why this estimator is consistent, consider the version $\widetilde{g}_{\mathrm{NW}}(x)$ of $\widehat{g}_{\mathrm{NW}}(x)$, where $\widehat{q}_0$ is replaced by q_0. Under standard smoothness assumptions and if $h \to 0$ and $nh \to \infty$ as $n \to \infty$, it is easy to check that the variance of the numerator and of the denominator of $\widetilde{g}_{\mathrm{NW}}(x)$, denoted below by Num and Den, tends to zero. For their mean, using a Taylor expansion we have $\mathbb{E}(\mathrm{Den}) = \mathbb{E}\big\{E\big(\Delta_{i,j}\big|X_{ij}\big)L_h(X_{i,j} - x)\big\} = \mathbb{E}\big\{\pi(X_{i,j})L_h(X_{i,j} - x)\big\} \sim \pi(x)f_X(x)$ and $\mathbb{E}(\mathrm{Num}) = \mathbb{E}\big\{\mathbb{E}(q_0^{1-n_j}Z_j^*|X_{i,j}, \Delta_{i,j})\Delta_{i,j}L_h(X_{i,j} - x)\big\} = \int \mathbb{E}(q_0^{1-n_j}Z_j^*|X_{i,j} = y, \Delta_{i,j} = 1)P(\Delta_{i,j} = 1|X_{ij} = y)L_h(y - x)f_X(y)\,dy \sim g(x)\pi(x)f_X(x)$. We deduce that under standard smoothness assumptions, $\widetilde{g}_{NW}(x) \xrightarrow{P} g(x)$ as $n \to \infty$.

More generally, if the $(X_{i,j}, Y_{i,j})$'s were available, we could estimate $p(x)$ by the standard local polynomial estimator of order ℓ (Fan and Gijbels, 1996), defined by $\widetilde{p}_{\text{LP}}(x) = \mathbf{e}_1^\top \mathbf{U}_n^{-1} \mathbf{V}_n$, where $\mathbf{e}_1^\top = (1, 0, \ldots, 0)$ with 1 in the first position and 0 elsewhere, and where $\mathbf{U}_n = (U_{n,k,k'})_{0 \le k,k' \le \ell}$ and $\mathbf{V}_n = (V_{n,0}, \ldots, V_{n,\ell})^\top$, with $U_{n,k,k'} = (nh^{k+k'})^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} L_h(X_{i,j} - x)(X_{i,j} - x)^{k+k'}$ and $V_{n,k} = (nh^k)^{-1} \sum_{j=1}^{J} Y_{i,j} \sum_{i=1}^{n_j} L_h(X_{i,j} - x)(X_{i,j} - x)^k$ for $k, k' = 0, \ldots, \ell$. Following the above construction of the estimator $\widehat{g}_{NW}$, in our case this suggests defining a local polynomial estimator of $g(x)$, of order ℓ, by

$$\widehat{g}_{\text{LP}}(x) = \mathbf{e}_1^\top \mathbf{S}_n^{-1} \mathbf{T}_n, \tag{5.2}$$

where $\mathbf{S}_n = (S_{n,k,k'})_{0 \le k,k' \le \ell}$ and $\mathbf{T}_n = (T_{n,0}, \ldots, T_{n,\ell})^\top$, with, for $k, k' = 0, \ldots, \ell$, $S_{n,k,k'} = (nh^{k+k'})^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \Delta_{i,j} L_h(X_{i,j} - x)(X_{i,j} - x)^{k+k'}$ and $T_{n,k} = (nh^k)^{-1} \sum_{j=1}^{J} \widehat{q}_0^{\,1-n_j} Z_j^* \sum_{i=1}^{n_j} \Delta_{i,j} L_h(X_{i,j} - x)(X_{i,j} - x)^k$. Then, using (3.7), we can estimate $p(x)$ by $\widehat{p}_{\text{LP}}(x) = \{1 - \widehat{g}_{\text{LP}}(x)\}/\{1 + (\widehat{p}_1 \widehat{p}_0^{-1} - 1)\widehat{g}_{\text{LP}}(x)\}$.

5.2 Parametric estimator

Sometimes we have a parametric model for p, i.e. we know that $p(x) = p_{\boldsymbol{\theta}}(x)$, where the function $p_{\boldsymbol{\theta}}$ is known up to the value of a parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{d_p}$, with $d_p \in \mathbb{Z}^+$ finite and Θ a compact set. Let $\boldsymbol{\theta}_0 \in \Theta$ denote the true value of $\boldsymbol{\theta}$. In this section we construct an estimator of $\boldsymbol{\theta}_0$ without making parametric assumptions on f_X.

Recall that we observe Y_j^* and $\Delta_{i,j}$ for all $i = 1, \ldots, n_j$, $j = 1, \ldots, J$, and $X_{i,j}$ when $\Delta_{i,j} = 1$. Let $\mathcal{Y} = \{Y_j^*, j = 1 \ldots, J\}$, $\mathit{\Delta} = \{\Delta_{i,j}, i = 1, \ldots, n_j, j = 1, \ldots, J\}$ and $\mathcal{X} = \{\mathcal{X}_{\text{obs}}, \mathcal{X}_{\text{mis}}\}$, where $\mathcal{X}_{\text{obs}}$ and $\mathcal{X}_{\text{mis}}$ denote respectively the observed $X_{i,j}$'s and the missing ones. In missing values problems, since we cannot compute the standard likelihood function depending on the unobserved full data, a common approach for estimating $\boldsymbol{\theta}$ (see Little and Rubin, 2014) is to maximise the likelihood function obtained by integrating the full likelihood with respect to the unobserved data, i.e. in our case,

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{Y}, \mathit{\Delta}, \mathcal{X}_{\text{obs}}) = \int_{\Omega_{\text{mis}}} f(\mathcal{Y}, \mathcal{X}, \mathit{\Delta}; \boldsymbol{\theta})\, d\mathcal{X}_{\text{mis}}\,, \tag{5.3}$$

where $f(\mathcal{Y}, \mathcal{X}, \mathit{\Delta}; \boldsymbol{\theta}) = P(\mathcal{Y}, \mathit{\Delta}|\mathcal{X}; \boldsymbol{\theta}) f(\mathcal{X})$, with $P(\cdot|\cdot; \boldsymbol{\theta})$ and $f(\cdot)$ the conditional probability mass and joint density function, and where Ω_{mis} denotes the outcome space for the missing values. In standard non grouped settings, this method gives a consistent estimator of $\boldsymbol{\theta}_0$.

To compute (5.3), note that $f(\mathcal{Y}, \mathcal{X}, \mathit{\Delta}; \boldsymbol{\theta}) = \prod_{j=1}^{J} P(Y_j^*, \mathit{\Delta}_j|\mathcal{X}_j; \boldsymbol{\theta}) \prod_{i=1}^{n_j} f_X(X_{i,j})$, where $\mathcal{X}_j = \{X_{i,j}, i = 1, \ldots, n_j\}$ and $\mathit{\Delta}_j = \{\Delta_{i,j}, i = 1, \ldots, n_j\}$. Thus we need to express $P(Y_j^*, \mathit{\Delta}_j|\mathcal{X}_j; \boldsymbol{\theta})$ in terms of $p_{\boldsymbol{\theta}}$, and integrate the resulting expression for $f(\mathcal{Y}, \mathcal{X}, \mathit{\Delta}; \boldsymbol{\theta})$ w.r.t. the missing $X_{i,j}$'s. The calculations are quite technical and are relegated to Appendix A.6, where we show that

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{Y}, \mathit{\Delta}, \mathcal{X}_{\text{obs}}) = \prod_{j=1}^{J} f_{\boldsymbol{\theta}}(Y_j^*, \mathcal{X}_{j,\text{obs}}, \mathit{\Delta}_j)\,, \tag{5.4}$$

with $f_{\boldsymbol{\theta}}(Y_j^*, \mathcal{X}_{j,\text{obs}}, \mathit{\Delta}_j) = (1 - Y_j^*) f_{0,\boldsymbol{\theta}}(\mathcal{X}_{j,\text{obs}}, \mathit{\Delta}_j) + Y_j^*\{f_{X\Delta,\boldsymbol{\theta}}(\mathcal{X}_{j,\text{obs}}, \mathit{\Delta}_j) - f_{0,\boldsymbol{\theta}}(\mathcal{X}_{j,\text{obs}}, \mathit{\Delta}_j)\}$, $\mathcal{X}_{j,\text{obs}}$ denotes the observed $X_{i,j}$'s for group j, and where we used the notation

$$\begin{aligned} f_{0,\boldsymbol{\theta}}(\mathcal{X}_{j,\text{obs}}, \mathit{\Delta}_j) = \prod_{i=1}^{n_j} \Bigg[&\Delta_{i,j} p_0\{1 - p_{\boldsymbol{\theta}}(X_{i,j})\} f_X(X_{i,j}) \\ &+ (1 - \Delta_{i,j})(1 - p_0) \int_{-\infty}^{\infty} \{1 - p_{\boldsymbol{\theta}}(x)\} f_X(x)\, dx \Bigg]\,, \end{aligned} \tag{5.5}$$

$$\begin{aligned} f_{X\Delta,\boldsymbol{\theta}}(\mathcal{X}_{j,\text{obs}}, \mathit{\Delta}_j) = \prod_{i=1}^{n_j} \Bigg(&\Delta_{i,j} [p_0\{1 - p_{\boldsymbol{\theta}}(X_{i,j})\} + p_1 p_{\boldsymbol{\theta}}(X_{i,j})] f_X(X_{i,j}) \\ + (1 - \Delta_{i,j}) &\int_{-\infty}^{\infty} (1 - p_0)\{1 - p_{\boldsymbol{\theta}}(x)\} f_X(x) + (1 - p_1) p_{\boldsymbol{\theta}}(x) f_X(x)\, dx \Bigg)\,. \end{aligned} \tag{5.6}$$

Ideally, we could estimate $\boldsymbol{\theta}_0$ by maximising $\ell(\boldsymbol{\theta}) = \log\big\{\mathcal{L}(\boldsymbol{\theta}|\mathcal{Y}, \mathit{\Delta}, \mathcal{X}_{\text{obs}})\big\}/J$. However, p_0, p_1 and f_X are unknown. While p_0 and p_1 can be estimated by $\widehat{p}_0$ and $\widehat{p}_1$ as in section 3.1, estimating f_X nonparametrically is less easy. We have already seen in section 4.2 that instead of f_X, what the kernel estimator $\widehat{f}_{X,\text{obs}}$ at (4.7) consistently estimates from our partially observed data is πf_X. To deduce an estimator of f_X, we use (3.6) to write $f_X(x) = \pi(x) f_X(x)/[p_1 p(x) + p_0\{1 - p(x)\}]$. Then we estimate p_0, p_1 and πf_X by $\widehat{p}_0$, $\widehat{p}_1$ and $\widehat{f}_{X,\text{obs}}$, respectively. However, p is unknown (recall that our goal is precisely to estimate it), and on this occasion we cannot replace it by $p_{\boldsymbol{\theta}}$ because to estimate πf_X we have used $\widehat{f}_{X,\text{obs}}$, which, because it is consistent, corresponds to $\boldsymbol{\theta}_0$ and not $\boldsymbol{\theta}$. Instead, we suggest using the Nadaraya-Watson estimator, $\widehat{p}_{\text{NW}}$ from section 5.1, which leads to the following estimator of $f_X(x)$:

$$\widehat{f}_X(x) = \widehat{f}_{X,\text{obs}}(x)/[\widehat{p}_1 \widehat{p}_{\text{NW}}(x) + \widehat{p}_0\{1 - \widehat{p}_{\text{NW}}(x)\}] \,. \tag{5.7}$$

Now, let $\widehat{f}_{n,\boldsymbol{\theta}}(Y_j^*, \mathcal{X}_{j,\text{obs}}, \varDelta_j) = (1 - Y_j^*)\widehat{f}_{0,n,\boldsymbol{\theta}}(\mathcal{X}_{j,\text{obs}}, \varDelta_j) + Y_j^*\{\widehat{f}_{X\Delta,n,\boldsymbol{\theta}}(\mathcal{X}_{j,\text{obs}}, \varDelta_j) - \widehat{f}_{0,n,\boldsymbol{\theta}}(\mathcal{X}_{j,\text{obs}}, \varDelta_j)\}$, where $\widehat{f}_{0,n,\boldsymbol{\theta}}$ and $\widehat{f}_{X\Delta,n,\boldsymbol{\theta}}$ are defined as at (5.5) and (5.6), respectively, but with p_0, p_1 and f_X there replaced by respectively $\widehat{p}_0, \widehat{p}_1$ and $\widehat{f}_X$. We propose to estimate $\boldsymbol{\theta}_0$ by $\widehat{\boldsymbol{\theta}}_n$ that maximises the following estimated log-likelihood function:

$$\widehat{\ell}_n(\boldsymbol{\theta}) \;=\; \frac{1}{J}\sum_{j=1}^{J} \log\left\{\widehat{f}_{n,\boldsymbol{\theta}}(Y_j^*, \mathcal{X}_{j,\text{obs}}, \varDelta_j)\right\}, \tag{5.8}$$

under the constraint $\widehat{f}_{n,\boldsymbol{\theta}}(Y_j^*, \mathcal{X}_{j,\text{obs}}, \varDelta_j) > 0$.

To establish asymptotic normality of the estimator $\widehat{\boldsymbol{\theta}}_n$, we need Assumption **P** below, which we discuss in Appendix A.7. Throughout, we let $\nabla_{\boldsymbol{\theta}}$ denote the gradient operator, $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}$ be the Hessian operator, $\|\cdot\|_2$ denote the Euclidean norm, and $\|f\|_{L_q}$ be the L_q norm of a function f for $1 \le q \le \infty$.

Assumption P

(P1) For all $x \in \mathbb{R}$, $p_{\boldsymbol{\theta}}(x)$ is twice continuously differentiable w.r.t. $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \Theta$, $0 < p_{\min} = \inf_{\boldsymbol{\theta}\in\Theta}\inf_{x\in\mathbb{R}} p_{\boldsymbol{\theta}}(x) \le \sup_{\boldsymbol{\theta}\in\Theta}\sup_{x\in\mathbb{R}} p_{\boldsymbol{\theta}}(x) = p_{\max} < 1$, $\sup_{\boldsymbol{\theta}\in\Theta}\sup_{x\in\mathbb{R}} \|\nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(x)\|_2 < \infty$ and $\sup_{\boldsymbol{\theta}\in\Theta}\sup_{x\in\mathbb{R}} \|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(x)\|_2 < \infty$.

(P2) Θ is a compact set and $\boldsymbol{\theta}_0 \in \text{interior}(\Theta)$.

(P3) f_X is a uniformly continuous and continuously differentiable density function such that $\|f_X'\|_{L_\infty} < \infty$ and $f_{\max} \equiv \|f_X\|_{L_\infty} < \infty$.

(P4) For all $x \in \mathbb{R}$, and $P_{\boldsymbol{\theta}} = p_{\boldsymbol{\theta}}$, $P_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}$ and $P_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta\theta}} p_{\boldsymbol{\theta}}$, there exist constants $0 < C_{1,P}, C_{2,P}, \alpha_{1,P}, \alpha_{2,P} < \infty$ such that for all $\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}} \in \Theta$, $\|P_{\boldsymbol{\theta}}(x) f_X(x) - P_{\widetilde{\boldsymbol{\theta}}}(x) f_X(x)\|_2 \leq C_{1,P} \|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\|_2^{\alpha_{1,P}}$ and $\| \int_{\mathbb{R}} \{P_{\boldsymbol{\theta}}(x) f_X(x) - P_{\widetilde{\boldsymbol{\theta}}}(x) f_X(x)\}\, dx\|_2 \leq C_{2,P} \|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\|_2^{\alpha_{2,P}}$.

(P5) π is uniformly continuous and continuously differentiable such that $\|\pi'\|_{L_\infty} < \infty$ and $0 < \inf_{x\in\mathbb{R}} \pi(x) \leq \sup_{x\in\mathbb{R}} \pi(x) < 1$.

(P6) g is uniformly continuous and continuously differentiable such that $\|g'\|_{L_\infty} < \infty$ and $0 < \inf_{x\in\mathbb{R}} g(x) \leq \sup_{x\in\mathbb{R}} g(x) < 1$.

(P7) L is a symmetric, bounded and continuous probability density function such that $\int_{\mathbb{R}} |x| L(x)\, dx < \infty$.

(P8) $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

(P9) For $F = f_X \pi$ and $F = g f_X \pi$, we have $\int_{\mathbb{R}} |F'(x)|\, dx < \infty$ and, for all $x \in \mathbb{R}$ and all $c_x \in [-1, 1]$, we have $\int_{\mathbb{R}} \sqrt{F(x - c_x)}\, dx < \infty$.

(P10) There exists a constant $c_p > c_0 > 0$ such that $\min(p_0, p_1) > c_p$, where c_0 is defined at page 8.

(P11) For $j = 1, \ldots, J$, $\mathbb{E}\big[\nabla_{\boldsymbol{\theta}} \log\{f_j(\boldsymbol{\theta}_0)\} \nabla_{\boldsymbol{\theta}} \log\{f_j(\boldsymbol{\theta}_0)\}^\top\big]$ is nonsingular.

The next theorem establishes asymptotic properties of $\widehat{\boldsymbol{\theta}}_n$. Its proof is given in Appendix F of the supplemental file.

Theorem 2. *Let $\ell_0(\boldsymbol{\theta}) = \mathbb{E}\{\ell(\boldsymbol{\theta})\}$, where $\ell(\boldsymbol{\theta})$ is defined at page 19. Under Assumptions (A9) and $\mathbf{P}$, if $\boldsymbol{\theta}_0$ uniquely maximises $\ell_0(\boldsymbol{\theta})$ subject to $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{d_p}$, then $\sqrt{J}\Sigma_n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{D} N(0, I_{d_p})$ as $n \to \infty$, where $\Sigma_n = J^{-1} \sum_{j=1}^J \mathbb{E}[\nabla_{\boldsymbol{\theta}} \log\{f_j(\boldsymbol{\theta}_0)\} \nabla_{\boldsymbol{\theta}} \log\{f_j(\boldsymbol{\theta}_0)\}^\top]$, with $f_j(\boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(Y_j^*, \mathcal{X}_{j,obs}, \mathit{\Delta}_j)$ for $j = 1, \ldots, J$, and where I_{d_p} is the identity matrix of dimension d_p.*

Theorem 2 establishes asymptotic normality of our estimator and thus can be used to construct confidence intervals for the components $(\boldsymbol{\theta}_0)_i$ of $\boldsymbol{\theta}_0$, for $i = 1, \dots, d_p$. Since, in practice, the covariance matrix Σ_n is unknown, it needs to be estimated, for example by

$$\widehat{\Sigma}_n = \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \widehat{\ell}_n(\widehat{\boldsymbol{\theta}}) , \tag{5.9}$$

where $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}$ is defined above Assumption P and $\widehat{\ell}_n$ is defined at (5.8). Using Lemma F.1, (F.29) and (F.30) in Appendix F, one can check that $\sqrt{J}\widehat{\Sigma}_n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \overset{D}{\to} N(0, I_{d_p})$ as $n \to \infty$. Therefore, we can compute an asymptotic $1 - \alpha$ confidence interval for $(\boldsymbol{\theta}_0)_i$, for $i = 1, \dots, d_p$, by taking

$$(\widehat{\boldsymbol{\theta}}_n)_i \pm z_{\alpha/2} J^{-1/2} \sqrt{(\widehat{\Sigma}_n^{-1})_{ii}} \, . \tag{5.10}$$

Using similar arguments, we can make simultaneous inference on $d_q < d_p$ components of $\boldsymbol{\theta}_0$, say $(\boldsymbol{\theta}_0)_{\boldsymbol{i}_{d_q}}$ with $\boldsymbol{i}_{d_q}$ a set of indices in $\{1, \dots, d_p\}$. We can obtain a simultaneous asymptotic $1 - \alpha$ confidence region for $(\boldsymbol{\theta}_0)_{\boldsymbol{i}_{d_q}}$ by taking the set of all $\boldsymbol{\theta}_{d_q} \in \mathbb{R}^{d_q}$ satisfying

$$J\{(\widehat{\boldsymbol{\theta}}_n)_{\boldsymbol{i}_{d_q}} - \boldsymbol{\theta}_{d_q}\}^\top (\widehat{\Sigma}_n)_{\boldsymbol{i}_{d_q}} \{(\widehat{\boldsymbol{\theta}}_n)_{\boldsymbol{i}_{d_q}} - \boldsymbol{\theta}_{d_q}\} \le \chi^2_{d_q}(1 - \alpha) , \tag{5.11}$$

where $(\widehat{\Sigma}_n)_{\boldsymbol{i}_{d_q}}$ denotes the submatrix of $\widehat{\Sigma}_n$ taking the rows and columns corresponding to the index set $\boldsymbol{i}_{d_q}$. Another possibility, which is easier to compute, is to use a confidence box $\prod_{i \in \boldsymbol{i}_{d_p}} \left\{(\widehat{\boldsymbol{\theta}}_n)_i \pm z_{\alpha/(2d_q)} J^{-1/2} \sqrt{(\widehat{\Sigma}_n^{-1})_{ii}}\right\}$, but this generally provides conservative confidence intervals (see e.g. Galambos and Simonelli, 1996).

5.3 Multivariate techniques

Our nonparametric procedures can be extended to the case where X is multivariate. Let $\boldsymbol{X}^\top = (\boldsymbol{W}^\top, \boldsymbol{V}^\top) \in \mathbb{R}^\kappa$, $\kappa > 1$, denote the vector of covariates, where $\boldsymbol{W} \in \mathbb{R}^{\kappa_1}$, $\kappa_1 \ge 0$, denotes the covariates that are always fully observed, $\boldsymbol{V} \in \mathbb{R}^{\kappa_2}$, $\kappa_2 > 1$, denotes the covariates that are subject to missingness and $\kappa_1 + \kappa_2 = \kappa$. The case where $\kappa_2 = 0$ is the standard multivariate group testing problem with no missing data (Delaigle and Meister, 2011 and Delaigle et al., 2014). One way to generalise the assumption (2.2) on the missing mechanism to the multivariate case is to assume that the probability that $\boldsymbol{X}$ is observed depends on the response variable Y and all the fully observed covariates (e.g. Robins et al., 1994, Chen, 2004, Efromovich, 2011, Liang et al., 2011 and Wei et al., 2012), that is, if we let $\Delta = 1\{\boldsymbol{V}$ is observed$\}$,

$$\mathbb{P}(\Delta = r|\boldsymbol{X}, Y) = \mathbb{P}(\Delta = r|\boldsymbol{W}, Y)\,, \quad \text{for} \quad r = 0, 1\,. \tag{5.12}$$

We show how to extend our nonparametric methodology to this setting.

Under (5.12), using arguments similar to those used to derive (3.4), it can be proved that, for $i = 1, \ldots, n_j$, $j = 1, \ldots, J$, we have

$$\{1 - p(\boldsymbol{x})\}\mathbb{P}(\Delta_{i,j} = 1|\boldsymbol{X}_{i,j} = \boldsymbol{x}, Z_j^* = 1) = g(\boldsymbol{x})\pi(\boldsymbol{x})\,, \tag{5.13}$$

where $\boldsymbol{x}^\top = (\boldsymbol{w}^\top, \boldsymbol{v}^\top) \in \mathbb{R}^\kappa$, $g(\boldsymbol{x}) = \mathbb{E}(q_0^{1-n_j} Z_j^*|\boldsymbol{X}_{i,j} = \boldsymbol{x}, \Delta_{i,j} = 1)$ and

$$\pi(\boldsymbol{x}) = \mathbb{P}(\Delta_{i,j} = 1|\boldsymbol{X}_{i,j} = \boldsymbol{x}) = P_0(\boldsymbol{w})\{1 - p(\boldsymbol{x})\} + P_1(\boldsymbol{w})p(\boldsymbol{x})\,, \tag{5.14}$$

with $P_0(\boldsymbol{w}) = \mathbb{P}(\Delta_{i,j} = 1|\boldsymbol{W}_{i,j} = \boldsymbol{w}, Y_{i,j} = 0)$ and $P_1(\boldsymbol{w}) = \mathbb{P}(\Delta_{i,j} = 1|\boldsymbol{W}_{i,j} = \boldsymbol{w}, Y_{i,j} = 1)$. Now, using (5.12) and arguments similar to those used in Appendix A.2, we have

$$\mathbb{P}(\Delta_{i,j} = 1|\boldsymbol{X}_{i,j} = \boldsymbol{x}, Z_j^* = 1) = \mathbb{P}(\Delta_{i,j} = 1|\boldsymbol{W}_{i,j} = \boldsymbol{w}, Z_j^* = 1) = P_0(\boldsymbol{w})\,. \tag{5.15}$$

Combining (5.13), (5.14) and (5.15) and assuming that $P_0(\boldsymbol{w}) > 0$ and $P_1(\boldsymbol{w}) > 0$, we get

$$p(\boldsymbol{x}) = \big\{1 - g(\boldsymbol{x})\big\}\big/\big[\{P_1(\boldsymbol{w})/P_0(\boldsymbol{w}) - 1\}g(\boldsymbol{x}) + 1\big]\,, \tag{5.16}$$

and to estimate p nonparametrically, we need to estimate g, P_1 and P_0 nonparametrically.

When $\kappa_1 = 0$, $\boldsymbol{W}$ is empty and $P_0(\boldsymbol{w}) = p_0$ and $P_1(\boldsymbol{w}) = p_1$, where p_0 and p_1 are defined at page 8 and can be estimated using the methods described there. When $\kappa_1 > 0$, things are more involved than in the univariate case because P_0 and P_1 depend also on $\boldsymbol{W}$. Now, P_0 depends only on observed variables, but P_1 depends on $Y_{i,j}$ which is not available. However, using a decomposition similar to the one we used in Appendix A.2 for p_1, we can express P_1 in a form that involves quantities that depend only on the observed data:

$$P_1(\boldsymbol{w}) = \big\{\Pi(\boldsymbol{w}) - Q_0(\boldsymbol{w})P_0(\boldsymbol{w})\big\}/\big\{1 - Q_0(\boldsymbol{w})\big\}\,, \tag{5.17}$$

where $\Pi(\boldsymbol{w}) = \mathbb{E}(\Delta_{i,j}|\boldsymbol{W}_{i,j} = \boldsymbol{w})$ and $Q_0(\boldsymbol{w}) = \mathbb{E}(q_0^{1-n_j} Z_j^*|\boldsymbol{W}_{i,j} = \boldsymbol{w})$.

It follows from (5.16) and (5.17) that to estimate p we need to estimate g, P_0, Π and Q_0. The latter three are standard regression curves which can be estimated by standard nonparametric multivariate regression techniques, such as tensor product splines (De Boor, 2001 and Ruppert et al., 2003), local polynomial estimators (Fan and Gijbels, 1996), or a dimension reduced version of them (Fan and Gijbels, 1996). We can estimate P_0 by applying such techniques to the pairs $(\boldsymbol{W}_{i,j}, \Delta_{i,j})$, where $i = 1, \ldots, n_j$ with $j = 1, \ldots, J$ such that $Z_j^* = 1$. For Q_0 and Π, respectively, we apply those techniques to, respectively, the pairs $(\boldsymbol{W}_{i,j}, q_0^{1-n_j} Z_j^*)$ and $(\boldsymbol{W}_{i,j}, \Delta_{i,j})$, for $i = 1, \ldots, n_j$, $j = 1, \ldots, J$. Below we use $\widehat{P}_0, \widehat{\Pi}$ and $\widehat{Q}_0$ to denote the resulting estimators of P_0, Π and Q_0.

It remains to see how we can estimate g. A first approach is to use a fully nonparametric estimator, for example a cubic tensor-product spline estimator, as follows. Recall the definition of B-splines in Appendix E.1 and let $\otimes$ denote the usual tensor product. Using notations similar to those under (3.14), for $\ell = 1, \ldots, \kappa$, let $\boldsymbol{N}_{K_\ell}(x_\ell)$ be the vector of cubic B-splines and K_ℓ knots for the ℓth component of $\boldsymbol{X}$, and let $\boldsymbol{\mathcal{N}}_{K_\ell} = \big(\boldsymbol{N}_{K_\ell}(X_{1,1,\ell}), \ldots, \boldsymbol{N}_{K_\ell}(X_{n_J,J,\ell})\big)^\top$. Our univariate spline estimator of g at (3.14) can be extended to the following cubic tensor-product spline estimator of $g(\boldsymbol{x})$:

$$\widehat{g}(\boldsymbol{x}) = \phi(\widehat{q}_0)\boldsymbol{B}^\top(\boldsymbol{x})\{\phi(\widehat{q}_0)\boldsymbol{\mathcal{B}}^\top\boldsymbol{\Phi}(\widehat{q}_0)\boldsymbol{\Delta}\boldsymbol{\mathcal{B}} + \lambda\boldsymbol{\mathcal{D}}\}^{-1}\boldsymbol{\mathcal{B}}^\top\boldsymbol{\Phi}(\widehat{q}_0)\boldsymbol{\Delta}\widehat{\boldsymbol{Q}}\boldsymbol{Z}^*\,, \tag{5.18}$$

where $\boldsymbol{B}(\boldsymbol{x}) = \boldsymbol{N}_{K_1}(x_1) \otimes \cdots \otimes \boldsymbol{N}_{K_\kappa}(x_\kappa)$, $\boldsymbol{\mathcal{B}} = \big(\boldsymbol{B}(\boldsymbol{X}_{1,1}), \ldots, \boldsymbol{B}(\boldsymbol{X}_{n_J,J})\big)^\top$ and $\boldsymbol{\mathcal{D}} =$

$\sum_{\ell,k=1}^{\kappa}\int_A \left\{\partial^2\boldsymbol{B}(\boldsymbol{x})/(\partial x_\ell \partial x_k)\right\}\left\{\partial^2\boldsymbol{B}(\boldsymbol{x})/(\partial x_\ell \partial x_k)\right\}^\top d\boldsymbol{x}$, with $A \subset \mathbb{R}^\kappa$ the range where we want to estimate p, and where ϕ, $\widehat{q}_0$, $\boldsymbol{\Phi}$, $\boldsymbol{\Delta}$, $\widehat{\boldsymbol{Q}}$ and $\boldsymbol{Z}^*$ are defined in section 3.2. See Appendix A.8 for how to compute the ϕ_j's needed for ϕ and Φ in this case.

When $\kappa > 3$, estimating g fully nonparametrically as above may not be a good choice because nonparametric estimators suffer from the so-called curse of dimensionality (their convergence rates degrade fast as dimension increases). Instead it is common to use methods that combine parametric and nonparametric components. A popular approach to this is the partially linear model, which can handle discrete and continuous covariates simultaneously. See Delaigle et al. (2014) in the group testing case without missing data.

In the partially linear model, $\boldsymbol{X}$ is decomposed as $\boldsymbol{X}^\top = (\boldsymbol{T}^\top, \boldsymbol{U}^\top)$, where the contribution from $\boldsymbol{T} \in \mathbb{R}^{\kappa_3}$ is modelled nonparametrically and that from $\boldsymbol{U} \in \mathbb{R}^{\kappa-\kappa_3}$ is modelled parametrically. For example, a partially linear model on g assumes that, for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$, $g(\boldsymbol{X}_{i,j}) = m(\boldsymbol{T}_{i,j}) + \boldsymbol{U}_{i,j}^\top \boldsymbol{\beta}$, where m is an unknown smooth function and $\boldsymbol{\beta} \in \mathbb{R}^{\kappa-\kappa_3}$ is a vector of unknown parameters. In the case where some of the covariates are discrete and others are continuous, the discrete covariates typically contribute only to the linear part; see e.g. Delaigle et al. (2014). Moreover, to avoid the curse of dimensionality, we typically choose $\boldsymbol{T}$ so that $\kappa_3 \leq 3$.

Using the tensor product B-spline technique described above to estimate m, and taking an approach similar to the one at (3.13), we can estimate m and $\boldsymbol{\beta}$ by

$$(\widehat{m}, \widehat{\boldsymbol{\beta}}) = \underset{s\in\mathcal{S}, \boldsymbol{\beta}\in\mathbb{R}^{\kappa-\kappa_3}}{\operatorname{argmin}} \left[\sum_{j=1}^{J} \sum_{i=1}^{n_j} \{\widehat{q}_0^{1-n_j} Z_j^* - s(\boldsymbol{T}_{i,j}) - \boldsymbol{U}_{i,j}^\top \boldsymbol{\beta}\}^2 \Delta_{i,j} + \lambda \int_{A_T} \sum_{\ell,k=1}^{\kappa_3} \left\{ \frac{\partial^2 s(\boldsymbol{t})}{\partial t_\ell \partial t_k} \right\}^2 d\boldsymbol{t} \right],$$

where $\mathcal{S}$ is the κ_3-dimensional cubic tensor-product spline space and $A_T \subset \mathbb{R}^{\kappa_3}$ denotes the range of $\boldsymbol{t}$ corresponding to the range of $\boldsymbol{x}^\top = (\boldsymbol{t}^\top, \boldsymbol{u}^\top) \in \mathbb{R}^\kappa$ where we want to estimate $p(\boldsymbol{x})$. See Holland (2017) for tensor product spline estimators in the partially linear model in the standard i.i.d. case without grouped nor missing data. It follows from standard calculations that $\widehat{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\mathcal{U}}}^\top \tilde{\boldsymbol{\mathcal{U}}})^{-1} \tilde{\boldsymbol{\mathcal{U}}}^\top \tilde{\boldsymbol{R}}$ and $\widehat{m}(\boldsymbol{t}) = \boldsymbol{B}^\top(\boldsymbol{t}) \boldsymbol{C}(\boldsymbol{R} - \boldsymbol{\mathcal{U}}\widehat{\boldsymbol{\beta}})$, where $\boldsymbol{\mathcal{U}} = (\boldsymbol{U}_{1,1}, \ldots, \boldsymbol{U}_{n_J,J})^\top$, $\boldsymbol{R} = (\widehat{q}_0^{1-n_1} Z_1^* \overset{\times n_1}{\cdots}, \cdots, \widehat{q}_0^{1-n_J} Z_J^* \overset{\times n_J}{\cdots})^\top$, $\tilde{\boldsymbol{\mathcal{U}}} = \boldsymbol{\mathcal{U}} - \boldsymbol{\mathcal{B}} \boldsymbol{C} \boldsymbol{\mathcal{U}}$ and $\tilde{\boldsymbol{R}} = \boldsymbol{R} - \boldsymbol{\mathcal{B}} \boldsymbol{C} \boldsymbol{R}$, and where $\boldsymbol{C} = \phi(\widehat{q}_0)\{\phi(\widehat{q}_0) \boldsymbol{\mathcal{B}}^\top \boldsymbol{\Phi}(\widehat{q}_0) \boldsymbol{\Delta} \boldsymbol{\mathcal{B}} + \lambda \boldsymbol{\mathcal{D}}\}^{-1} \boldsymbol{\mathcal{B}}^\top \boldsymbol{\Phi}(\widehat{q}_0) \boldsymbol{\Delta}$, and $\boldsymbol{\mathcal{B}}$, $\boldsymbol{B}$, $\boldsymbol{\mathcal{D}}$ are as above, but with $\boldsymbol{X}$ and $\boldsymbol{x}$ replaced by $\boldsymbol{T}$ and $\boldsymbol{t}$. Finally the partially linear estimator of $g(\boldsymbol{x})$ is $\widehat{g}(\boldsymbol{x}) = \widehat{m}(\boldsymbol{t}) + \boldsymbol{u}^\top \widehat{\beta}$.

Let $\widehat{g}$ denote the fully nonparametric or the partially linear estimator of g derived above. Using (5.16), we deduce that, in the multivariate case, we can estimate $p(\boldsymbol{x})$ by

$$\widehat{p}(\boldsymbol{x}) = \begin{cases} \big\{1 - \widehat{g}(\boldsymbol{x})\big\} / \big\{(\widehat{p}_1/\widehat{p}_0 - 1)\widehat{g}(\boldsymbol{x}) + 1\big\} & \text{if } \kappa_1 = 0\,, \\ \big\{1 - \widehat{g}(\boldsymbol{x})\big\} / \big[\{\widehat{P}_1(\boldsymbol{w})/\widehat{P}_0(\boldsymbol{w}) - 1\}\widehat{g}(\boldsymbol{x}) + 1\big] & \text{if } \kappa_1 > 0\,, \end{cases} \tag{5.19}$$

where $\widehat{p}_0$ and $\widehat{p}_1$ are defined at page 8 and $\widehat{P}_1(\boldsymbol{w}) = \big\{\widehat{\Pi}(\boldsymbol{w}) - \widehat{Q}_0(\boldsymbol{w})\widehat{P}_0(\boldsymbol{w})\big\} / \big\{1 - \widehat{Q}_0(\boldsymbol{w})\big\}$.

6 Numerical properties

6.1 Choosing K, λ and h by cross-validation

We suggest choosing the number of knots K and the smoothing parameter λ of our univariate spline estimator by cross-validation. Specifically, motivated by (3.3), we choose K and λ as the values which minimise the following criterion w.r.t. K and λ:

$$CV(K, \lambda) = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \Delta_{i,j} \{\widehat{q}_0^{1-n_j} Z_j^* - \widehat{g}_s^{(-j)}(X_{i,j}, K, \lambda)\}^2 \mathbb{1}_{[\tilde{a}, \tilde{b}]}(X_{i,j})\,, \tag{6.1}$$

where $\tilde{a}$ and $\tilde{b}$ are empirical quantiles of the $X_{i,j}$'s, and $\widehat{g}_s^{(-j)}$ denotes the spline estimator obtained by minimising (3.13), but without using the observations from group j. Likewise, to choose the bandwidth h of the local polynomial estimator, we take h that minimises

$$CV(h) = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \Delta_{i,j} \big\{\widehat{q}_0^{1-n_j} Z_j^* - \hat{g}_{\mathrm{LP}}^{(-j)}(X_{i,j})\big\}^2 \mathbb{1}_{[\tilde{a}, \tilde{b}]}(X_{i,j})\,,$$

where $\hat{g}_{\mathrm{LP}}^{(-j)}$ denotes the local polynomial estimator at (5.2) computed without using the observations from group j. For both methods, in our simulations we took $\tilde{a}$ and $\tilde{b}$ to be the 0.1 and 0.9 empirical quantiles.

6.2 Simulation results

We applied our new univariate estimators to data $(\Delta_{i,j}, X_{i,j}, Y_j^*)$, where $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$, generated according the model described in section 2, where we considered three settings: (i) $p(x) = \exp(-4+2x)/\{8+8\exp(-4+2x)\}$ and $X \sim N(2, 1.5^2)$; (ii) $p(x) = \min\{1, \max(0, 0.03-0.05x+0.04x^2)\}$, $X = 4-Z/4$ and $Z \sim \chi^2(8)$; (iii) $p(x) = \{\sin(\pi x/2)+1.2\}/[20 + 40x^2\{\mathrm{sign}(x) + 1\}]$ and $X \sim N(0, 1.5^2)$, with two missing mechanisms: (1) $\mathbb{P}(\Delta_{i,j} = 1|Y_{i,j} = 0) = 0.7$ and $\mathbb{P}(\Delta_{i,j} = 1|Y_{i,j} = 1) = 0.9$; or (2) $\mathbb{P}(\Delta_{i,j} = 1|Y_{i,j} = 0) = 0.4$ and $\mathbb{P}(\Delta_{i,j} = 1|Y_{i,j} = 1) = 0.6$. In the parametric case, we used the following parametric forms (a), (b) and (c) for p, for models (i), (ii) and (iii) respectively: (a) $p_{\boldsymbol{\theta}}(x) = \exp(\theta_1 + \theta_2 x)/[\theta_3\{1 + \exp(\theta_1 + \theta_2 x)\}]$; (b) $p_{\boldsymbol{\theta}}(x) = \min\{1, \max(0, \theta_1 + \theta_2 x + \theta_3 x^2)\}$; (c) $p_{\boldsymbol{\theta}}(x) = \{\sin(\pi x/2) + \theta_1\}/[\theta_2 + \theta_3 x^2\{\mathrm{sign}(x) + 1\}]$.

We generated samples of size $n = 2000$ and 5000, and pooled the data into J groups of equal sizes n_j equal to 4 or 8. Following section 4.2, in that case we took all φ_j's equal to 1 but we also did simulations for unequal group sizes; see our results in Appendix B. For each combination of n, n_j and model setting, we generated 200 samples. For each sample, we estimated p using each of the penalised spline estimator $\widehat{p}_s$ from section 3.2, the local linear estimator $\widehat{p}_{\mathrm{LL}}$, that is $\widehat{p}_{\mathrm{LP}}$ with $\ell = 1$ from section 5.1 and the parametric estimator $p_{\widehat{\theta}}$ from section 5.2. To compute $\widehat{p}_0$ and $\widehat{p}_1$ used by our procedures (see section 3.1), we need to choose the small constant c_0 used to ensure that $\widehat{p}_0$ and $\widehat{p}_1$ are bounded away from zero. While this constant is required in our theoretical development, in practice it is largely unnecessary because, as p_0 and p_1 are usually far away from zero, $\widehat{p}_0$ and $\widehat{p}_1$ are almost always positive even if $c_0 = 0$. Therefore, we can take c_0 equal to any small constant and we took $c_0 = 0.001$.

To highlight the importance of addressing missingness in the right way, we computed the naive complete-case spline estimator that does not correct for the bias introduced by the missing data, i.e. $\widehat{p}_{\text{naive},s} = 1 - \widehat{g}_s$ where $\widehat{g}_s$ is defined at (3.12). Finally, to illustrate the loss incurred by the missing data, before deleting some $X_{i,j}$'s, we also computed two estimators from the full data $(X_{i,j}, Y_j^*)$, where $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$: the local linear estimator $\widehat{p}_{\text{oracle,LL}}$ of Delaigle and Meister (2011) and the spline estimator $\widehat{p}_{\text{oracle},s}$ which corresponds to (3.15) with $\widehat{p}_0 = \widehat{p}_1 \equiv 1$ and all $\Delta_{i,j}$'s equal to 1 (we call them oracle since they require the full data which are not available in our missing data context).

Throughout, for all methods involving a kernel function, we took it equal to the standard normal density. We selected the smoothing parameters of our new nonparametric methods as in section 6.1, and for the parametric estimator, we selected the bandwidth in (4.7) by the method of Sheather and Jones (1991). We chose the bandwidth h for $\widehat{p}_{\text{oracle,LL}}$ using the plug-in method of Delaigle and Meister (2011).

To summarise, we applied six methods: our three new consistent estimators based on the incomplete data, a naive inconsistent estimator computed from the incomplete data, and two consistent estimators computed from the complete data before some $X_{i,j}$'s were removed. For each method, we computed 200 estimators, each corresponding to one of the 200 samples. We measured the quality of each estimator $\widehat{p}$ of p by computing the integrated squared error ISE $= \int_a^b \{\widehat{p}(x) - p(x)\}^2\, dx$, where a and b were, respectively, the 5th and the 95th percentile of the range of X.

Table 1 presents, for each estimator and for models (i) to (iii) and missing mechanism (1), the mean and standard deviation of the 200 values of ISE obtained from the 200 simulated samples. We see that the behaviour of our nonparametric estimators $\widehat{p}_{\text{LL}}$ and $\widehat{p}_s$

Table 1: $10^4 \times$ Mean (Standard deviation) of 200 ISE values for six estimators of p, obtained from 200 samples simulated from models (i) to (iii) and missing mechanism (1), when the group sizes are equal to $n_j = 4$ or $n_j = 8$ and when $n = 2000$ or 5000.

		$n_j = 4$					
Model	n	$p_{\widehat{\theta}}$	$\widehat{p}_{\text{LL}}$	$\widehat{p}_s$	$\widehat{p}_{\text{naive},s}$	$\widehat{p}_{\text{oracle,LL}}$	$\widehat{p}_{\text{oracle},s}$
(i)	$2 \cdot 10^3$	9.42(6.86)	18.62(19.24)	12.71(11.44)	36.85(24.61)	13.07(11.29)	11.41(10.39)
	$5 \cdot 10^3$	5.91(4.64)	9.61(7.59)	6.27(5.52)	26.46(15.87)	7.34(6.23)	6..52(5.94)
(ii)	$2 \cdot 10^3$	10.33(9.65)	25.87(22.69)	16.78(16.12)	53.72(38.24)	19.73(14.17)	14.28(12.48)
	$5 \cdot 10^3$	5.62(4.51)	11.32(8.78)	7.26(6.75)	37.45(29.41)	9.60(6.57)	7.13(5.60)
(iii)	$2 \cdot 10^3$	1.95(1.54)	17.76(11.42)	16.60(8.95)	24.82(13.65)	14.27(7.31)	15.93(8.01)
	$5 \cdot 10^3$	1.19(0.90)	7.93(4.93)	8.23(4.75)	14.80(8.41)	8.99(3.44)	8.49(4.69)

		$n_j = 8$					
Model	n	$p_{\widehat{\theta}}$	$\widehat{p}_{\text{LL}}$	$\widehat{p}_s$	$\widehat{p}_{\text{naive},s}$	$\widehat{p}_{\text{oracle,LL}}$	$\widehat{p}_{\text{oracle},s}$
(i)	$2 \cdot 10^3$	14.27(11.63)	29.91(32.01)	24.39(23.67)	54.01(47.28)	29.27(27.37)	25.09(26.72)
	$5 \cdot 10^3$	10.24(7.19)	16.76(18.04)	11.14(9.92)	35.24(23.53)	13.62(11.28)	12.23(12.15)
(ii)	$2 \cdot 10^3$	16.12(14.58)	53.51(52.69)	33.52(36.77)	80.21(64.00)	53.35(47.37)	34.13(38.19)
	$5 \cdot 10^3$	11.58(8.58)	30.16(24.19)	17.74(16.76)	51.18(32.06)	23.82(17.23)	17.19(14.34)
(iii)	$2 \cdot 10^3$	2.23(1.73)	27.33(20.67)	24.75(14.61)	35.57(22.51)	21.80(14.74)	24.45(12.65)
	$5 \cdot 10^3$	1.62(1.02)	14.38(7.93)	14.32(7.17)	22.96(12.39)	12.52(6.28)	14.39(7.20)

computed from the grouped and incomplete data is reasonably close to that of the estimators $\widehat{p}_{\text{oracle,LL}}$ and $\widehat{p}_{\text{oracle},s}$ computed from the grouped but full data. This illustrates the results from Theorem 1, which imply that having missing $X_{i,j}$'s at random does not affect the convergence rates of nonparametric estimators. As expected, our new estimators worked much better than the non consistent one, $\widehat{p}_{\text{naive},s}$. Confirming the theory, the performance of our consistent estimators improved as sample size increased and degraded when the group size increased. Overall the spline estimator worked better than the local linear estimator. Finally, also unsurprisingly, our parametric estimator $p_{\widehat{\theta}}$ worked significantly better than the nonparametric ones. However, in real data settings, it can only be applied when we know the correct parametric form for p, which is not always possible.

Of course the more data are missing, the worse all estimators perform. This is illustrated in Table B.2 in Appendix B, where we compare the results of our three estimators and

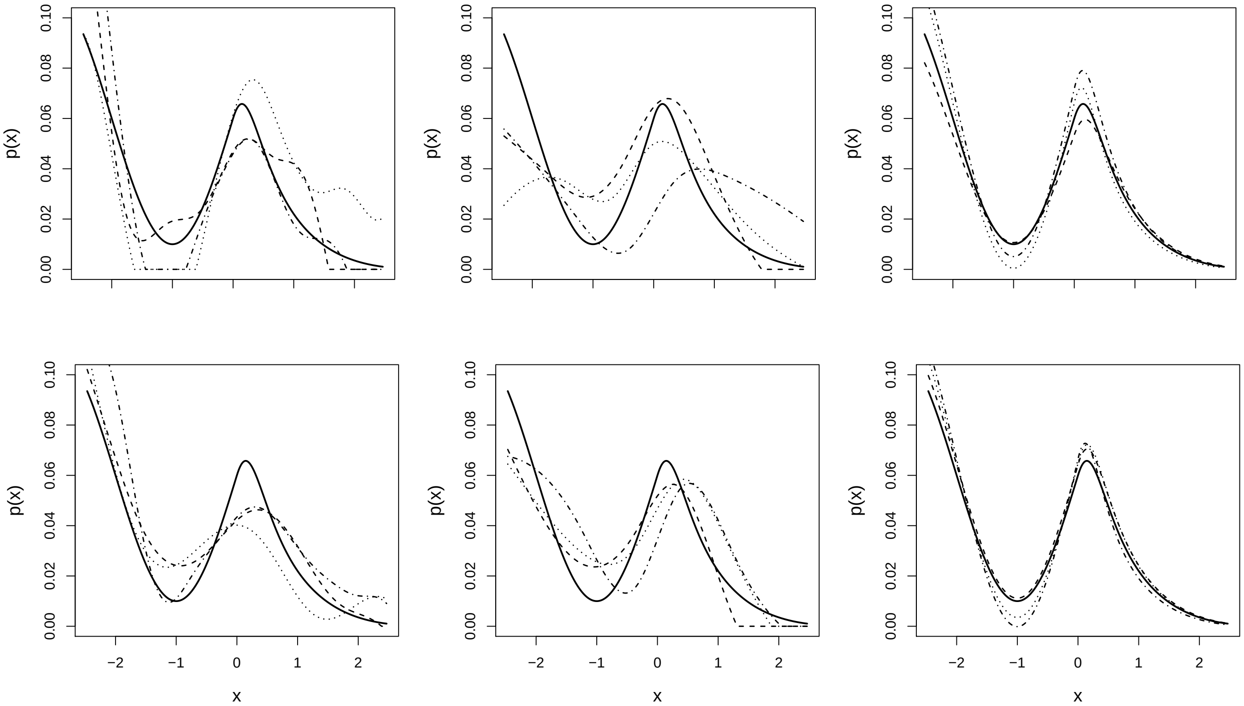

Figure 1: True curve (—) and the 1st (- - -), 2nd ($\cdots$) and 3rd ($- \cdot -\cdot$) quartile curves for $\widehat{p}_{\mathrm{LL}}$ (left), $\widehat{p}_s$ (center) and $p_{\widehat{\theta}}$ (right), obtained from 200 samples coming from model (iii) and missing mechanism (1) when $n = 2000$ (row 1) and $n = 5000$ (row 2), with $n_j = 4$.

the naive estimator under missing mechanisms (1) and (2). However, the estimator that degrades the most is the non consistent naive estimator since its bias increases with the missing rate.

Next we illustrate our estimators graphically by showing what we call quartile estimated curves. For a given estimator and a given setting, these are the three estimated curves corresponding to the three samples that gave the first, second and third quartiles of the 200 computed values of the ISE. In Figure 1 we illustrate the effect of increasing n by depicting the quartile curves obtained for model (iii) and missing mechanism (1) when $n_j = 4$ with $n = 2000$ and $n = 5000$. The quartiles curves for model (ii) with sample

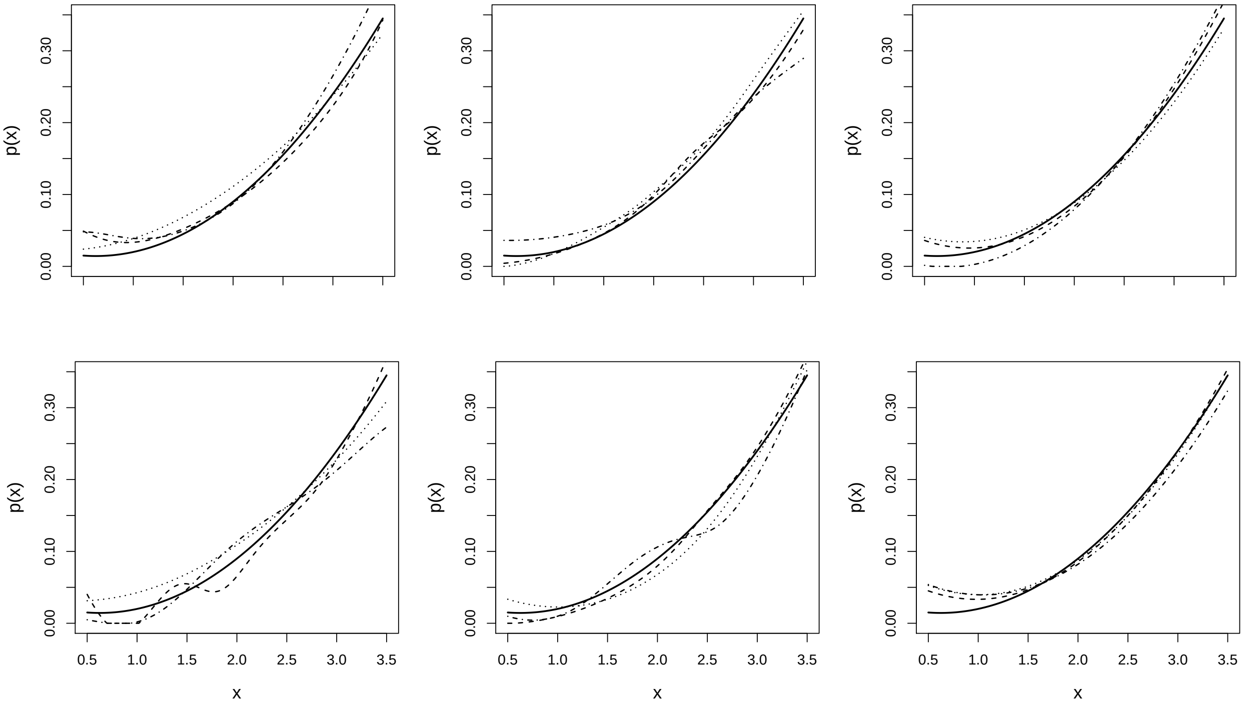

Figure 2: True curve (—) and the 1st (- - -), 2nd ($\cdots$) and 3rd ($- \cdot -\cdot$) quartile curves for $\widehat{p}_{\mathrm{LL}}$ (left), $\widehat{p}_s$ (center) and $p_{\widehat{\theta}}$ (right), obtained from 200 samples coming from model (ii) when $n = 5000$ and $n_j = 4$, under missing mechanism (1) (row 1) or missing mechanism (2) (row 2).

size $n = 5000$ and group size $n_j = 4$ and $n_j = 8$ displayed in Figure 2 illustrate the deterioration of all estimators when the missing rate increases. Finally, Figure 3 illustrates the deterioration of our nonparametric estimators when the group sizes increase; there we present the quartile curves obtained, in model (i), for $\widehat{p}_{\mathrm{naive},s}$, $\widehat{p}_{\mathrm{LL}}$ and $\widehat{p}_s$ when $n = 5000$, $n_j = 4$ and $n_j = 8$. We see that as n is not too small, our nonparametric estimators can work particularly well and significantly improve the naive estimator $\widehat{p}_{\mathrm{naive},s}$.

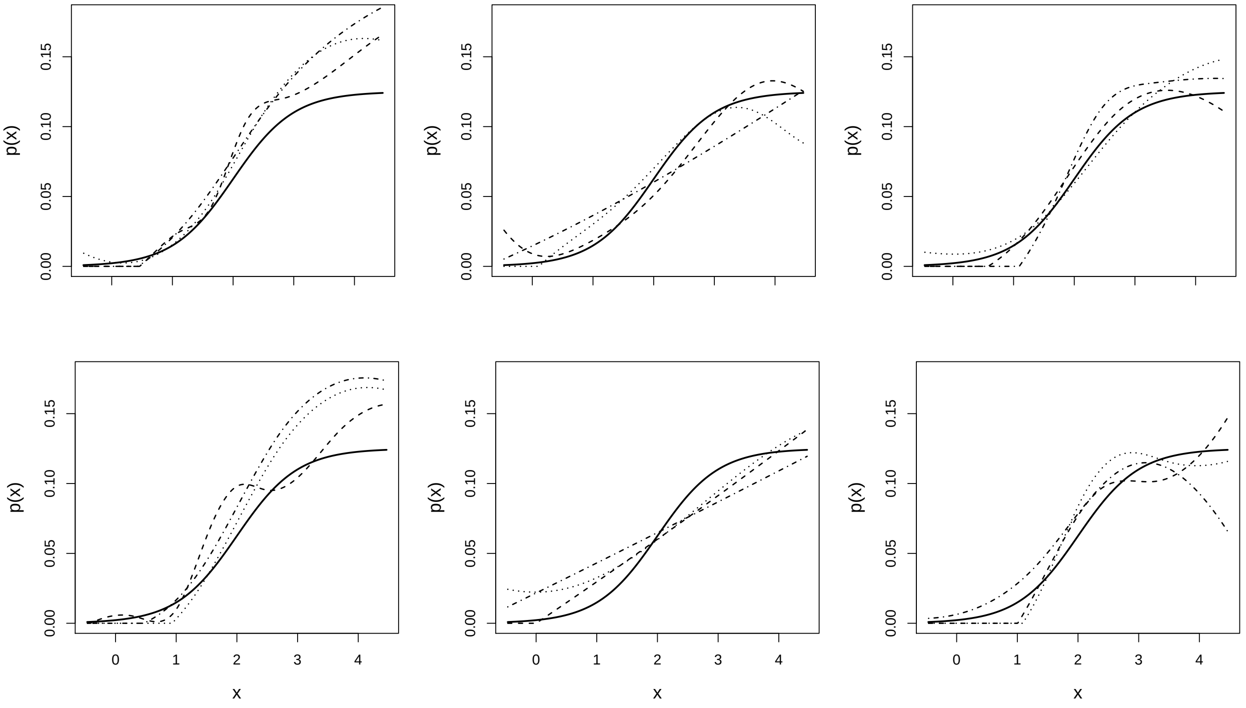

Figure 3: True curve (—) and the 1st (- - -), 2nd ($\cdots$) and 3rd ($- \cdot -\cdot$) quartile curves for $\widehat{p}_{\text{naive},s}$ (left), $\widehat{p}_{\text{LL}}$ (center) and $\widehat{p}_s$ (right), obtained from 200 samples coming from model (i) and missing mechanism (1) when $n = 5000$, $n_j = 4$ (row 1) and $n_j = 8$ (row 2).

6.3 Illustration with real data

Since group testing data do not usually come naturally grouped (pooling the data is often a choice driven by time and cost saving issues), typically the goal of real analyses in the group testing literature is to illustrate how new procedures designed for group testing data work when real, non grouped data, are pooled artificially, and compare their performance with that of standard estimators applied to individual data; in addition to references in the introduction, see Delaigle and Meister (2011), Zhang et al. (2013), Wang et al. (2014) and Delaigle and Zhou (2015). The idea of those analyses is to demonstrate that pooling can produce reasonable results, which could convince more labs worldwide to use this technique. In this spirit, we applied our estimators to real data coming from Demographics and Health Survey (DHS) available from `https://dhsprogram.com/data`. This very large survey was carried out in over 90 countries. It collected many variables, among which HIV data. To illustrate our method, we use a subset of the data, namely demographics and HIV data collected in Rwanda between 2014 and 2015.

In this example, we take Y to be the indicator of HIV infection, and $X = \log(Z + 10)$ where Z is the age at first sexual intercourse. As is often the case for real data illustrations from the group testing literature, in this example the individual observations Y are available, and we group the data artificially to see the effect that grouping has on estimators. In this dataset, only X is subject to missingness but some of the individuals had not had sex when the survey was conducted, and since such cases are not of interest here, we discarded those individuals, which leaves us with a sample of size $n = 4979$.

In this example, 57.9% of the individuals whose $Y = 1$ reported their X values, while only 39.2% of those with $Y = 0$ reported it, which suggests that the missingness of X does depend on the value of Y. Moreover, the missing rate is high, and the effective sample size is much smaller than 4979.

We randomly pooled the data in groups of size $n_j = 4$ and computed our local linear estimator $\widehat{p}_{\mathrm{LL}}$, our penalised spline estimator $\widehat{p}_s$ and our parametric estimator $p_{\widehat{\theta}}$ from the grouped data, where, for the parametric model, we took the logistic curve $p_{\boldsymbol{\theta}}(x) = \exp(\theta_1 + \theta_2 x)/\{1 + \exp(\theta_1 + \theta_2 x)\}$, which is often employed in prevalence studies. Finally we also computed, from the grouped data, the naive spline, local linear and parametric estimators, $\widehat{p}_{\mathrm{naive,LL}}$, $\widehat{p}_{\mathrm{naive},s}$ and $p_{\mathrm{naive},\widehat{\theta}}$ which use the complete cases without any correction for missingness. We repeated this process 200 times, that is, we grouped the data randomly in groups of size $n_j = 4$ in 200 different random ways, obtaining in this way 200 samples,

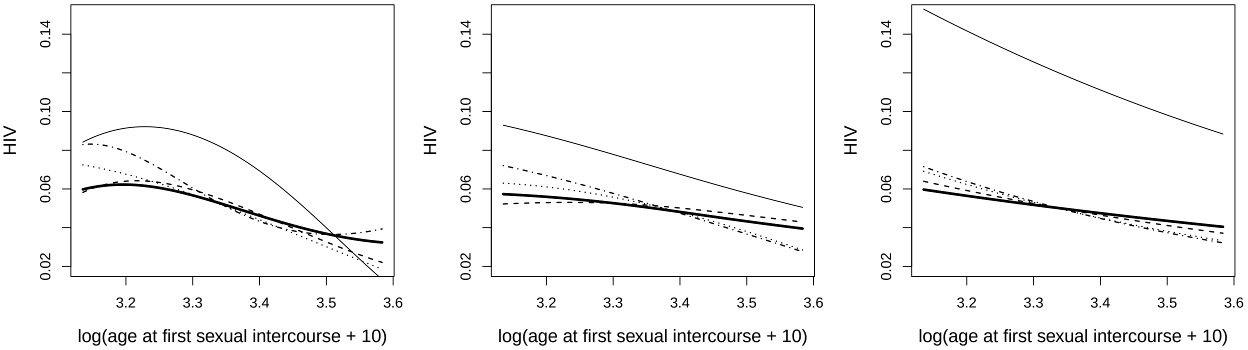

Figure 4: DHS study: $\widehat{p}_{\text{oracle}}$ (**——**), 2nd quartile curve for $\widehat{p}_{\text{naive}}$ (—), and 1st (- - -), 2nd ($\cdots$) and 3rd ($- \cdot -\cdot$) quartile curves for $\widehat{p}_{\text{new}}$, when using a local linear estimator (left), a spline estimator (center) and a parametric estimator (right).

and thus, for each method, 200 estimated curves.

Unlike in the simulation section, we cannot compare our estimators with the true curve p since we do not know it. However we have access to non grouped data, and from there we can compute estimators of p that are close to the true curve: $\widehat{p}_{\text{oracle,LL}}$, $\widehat{p}_{\text{oracle},s}$ and $p_{\text{oracle},\widehat{\theta}}$, which correspond to our local linear estimator $\widehat{p}_{\text{LL}}$, our penalised spline estimator $\widehat{p}_s$ and our parametric estimator $p_{\widehat{\theta}}$, computed with the ungrouped data with $n_j \equiv 1$. When assessing the quality of estimators, we treat those oracle curves as the true curve p. Specifically, for each method (local linear, spline or parametric, with naive and consistent versions), denoted here generically by $\widehat{p}$ for the estimator computed from grouped data, and $\widehat{p}_{\text{oracle}}$ for the estimator computed from non grouped data which plays the role of the truth, we calculated the 200 corresponding integrated squared differences ISD $= \int_a^b \{\widehat{p}(x) - \widehat{p}_{\text{oracle}}(x)\}^2\, dx$, where $a = 3.136$ and $b = 3.584$, which are the 5th and 95th percentile of the range of X, respectively and correspond to age 13 to 26.

In Figure 4, for the three estimation techniques, we depict the oracle estimator, our consistent estimator and the the naive estimator computed from the grouped and incomplete data, denoted generically by $\widehat{p}_{\text{oracle}}$, $\widehat{p}_{\text{new}}$ and $\widehat{p}_{\text{naive}}$, respectively, in the caption. For each version of $\widehat{p}_{\rm new}$, we show the three estimators that represent the first, second and third quartiles of the 200 ISDs. For $\widehat{p}_{\rm naive}$ we show only the second quartile curve.

As one could have expected, the conditional prevalence of HIV is a decreasing function of age at first sexual intercourse, and this trend is captured by all estimators. Compared to our consistent methods, the naive estimators tended to largely overestimate the prevalence of HIV because, in this example, the individuals without HIV have a higher missing probability. Thus, by deleting them without any adjustment, the ratio of HIV infection for each X value raises artificially.

Acknowledgments

Delaigle's research was supported by a Discovery Project and a Future Fellowship from the Australian Research Council. Huang's research is supported by Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS).

References

Chen, H. Y. (2004). Nonparametric and semiparametric models for missing covariates in parametric regression. *J. Amer. Statist. Assoc.*, 99:1176–1189.

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

Claeskens, G., Krivobokova, T., and Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96:529–544.

De Boor, C. (2001). *A Practical Guide to Splines.* Springer-Verlag, Berlin.

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

Delaigle, A. and Hall, P. (2015). Nonparametric methods for group testing data, taking dilution into account. *Biometrika*, 102:871–887.

Delaigle, A., Hall, P., and Jamshidi, F. (2015). Confidence bands in nonparametric errors-in-variables regression. *J. Royal Statist. Society Ser. B*, 77:149–169.

Delaigle, A., Hall, P., and Wishart, J. (2014). New approaches to nonparametric and semiparametric regression for univariate and multivariate group testing data. *Biometrika*, 101:567–585.

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

Delaigle, A. and Zhou, W.-X. (2015). Nonparametric and parametric estimators of prevalence from group testing data with aggregated covariates. *J. Amer. Statist. Assoc.*, 110:1785–1796.

Dempster, A. and Rubin, D. (1983). Introduction. In *Incomplete Data in Sample Surveys (Volume 2): Theory and Bibliography*, pages 3–10. Academic Press, New York.

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

Efromovich, S. (2011). Nonparametric regression with predictors missing at random. *J. Amer. Statist. Assoc.*, 106:306–319.

Fahey, J. W., Ourisson, P. J., and Degnan, F. H. (2006). Pathogen detection, testing, and control in fresh broccoli sprouts. *Nutr. J.*, 5:1–6.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability.* Chapman & Hall, London.

Fletcher, J. D., Russell, A. C., and Butler, R. C. (1999). Seed-borne cucumber mosaic virus in new zealand lentil crops: Yield effects and disease incidence. *New Zeal. J. Crop Hort.*, 27:197–204.

Galambos, J. and Simonelli, I. (1996). *Bonferroni-type Inequalities with Applications.* Springer-Verlag, New York.

Holland, A. D. (2017). Penalized spline estimation in the partially linear model. *J. Multivar. Anal.*, 153:211–235.

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

Jiang, D., Zhao, P., and Tang, N. (2016). A propensity score adjustment method for regression models with nonignorable missing covariates. *Comput. Stat. Data Anal.*, 94:98–119.

Kim, J. K. and Yu, C. L. (2012). A semiparametric estimation of mean functionals with nonignorable missing data. *J. Amer. Statist. Assoc.*, 106:157–165.

Krivobokova, T., Kneib, T., and Claeskens, G. (2010). Simultaneous confidence bands for penalized spline estimators. *J. Amer. Statist. Assoc.*, 105:852–863.

Lennon, J. T. (2007). Diversity and metabolism of marine bacteria cultivated on dissolved dna. *Appl. Environ. Microb.*, 73:2799–2805.

Lewis, J. L., Lockary, V. M., and Kobic, S. (2012). Cost savings and increased efficiency using a stratified specimen pooling strategy for chlamydia trachomatis and neisseria gonorrhoeae. *Sexually transmitted diseases*, 39:46–48.

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

Liang, H., Wang, S., and Carroll, R. J. (2007). Partially linear models with missing response variables and error-prone covariates. *Biometrika*, 94:185–198.

Liang, H., Wang, S., Robins, J. M., and Carroll, R. J. (2011). Estimation in partially linear models with missing covariates. *J. Amer. Statist. Assoc.*, 99:357–367.

Lindan, C., Mathur, M., Kumta, S., Jerajani, H., Gogate, A., Schachter, J., and Moncada, J. (2005). Utility of pooled urine specimens for detection of chlamydia trachomatis and neisseria gonorrhoeae in men attending public sexually transmitted infection clinics in mumbai, india, by pcr. *J. Clin. Microbiol*, 43:1674–1677.

Little, R. J. (1986). Survey nonresponse adjustments for estimates of means. *Int. Stat. Rev.*, 54:139–157.

Little, R. J. and Rubin, D. B. (2014). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, A., and Verbeke, G. (2014). *Handbook of Missing Data Methodology*. Handbooks of Modern Statistical Methods. Boca Raton, Florida.

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., and Eskridge, K. (2012). Sample size under inverse negative binomial group testing for accuracy in parameter estimation. *PloS one*, 7:1–11.

Montesinos-López, O. A., Montesinos-Lopez, A., Crossa, J., and Eskridge, K. (2013). Sample size for detecting transgenic plants using inverse binomial group testing with dilution effect. *Seed Sci. Res.*, 23:279–288.

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

Oh, H. L. and Scheuren, F. J. (1983). Weighting adjustment for unit nonresponse. In

Incomplete Data in Sample Surveys (Volume 2): Theory and Bibliography, pages 143–184. Academic Press, New York.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.*, 89:846–866.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Rubin, D. B., Stern, H. S., and Vehovar, V. (1995). Handling "don't know" survey responses: The case of the slovenian plebiscite. *J. Amer. Statist. Assoc.*, 90:822–828.

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

Sarov, B., Novack, L., Beer, N., Safi, J., Soliman, H., Pliskin, J., Litvak, E., Yaari, A., and Shinar, E. (2007). Feasibility and cost-benefit of implementing pooled screening for hcvag in small blood bank settings. *Transfusion Medicine*, 17:479–487.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory.* Cambridge University Press, Cambridge.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B*, 53:683–690.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8:1348–1360.

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.

Verstraeten, T., Farah, B., Duchateau, L., and Matu, R. (1998). Pooling sera to reduce the cost of hiv surveillance: a feasibility study in a rural kenyan district. *Trop. Med. Int. Health*, 3:747–750.

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

Wang, C., Wang, S., Gutierrez, R. G., and Carroll, R. J. (1998). Local linear regression for generalized linear models with missing data. *Ann. Statist.*, 26:1028–1050.

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

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

Wei, Y., Ma, Y., and Carroll, R. J. (2012). Multiple imputation in quantile regression. *Biometrika*, 99:423–438.

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

Zhang, B., Bilder, C. R., and Tebbs, J. M. (2013). Regression analysis for multiple-disease group testing data. *Stat. Med.*, 32:4954–4966.