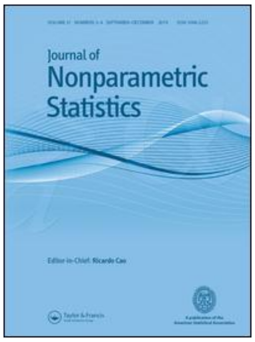


Taylor & Francis
Taylor & Francis Group

Journal of Nonparametric Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gnst20

Local polynomial regression for pooled response data

Dewei Wang , Xichen Mou , Xiang Li & Xianzheng Huang

To cite this article: Dewei Wang , Xichen Mou , Xiang Li & Xianzheng Huang (2020): Local polynomial regression for pooled response data, Journal of Nonparametric Statistics, DOI: 10.1080/10485252.2020.1834104

To link to this article: https://doi.org/10.1080/10485252.2020.1834104

JOURNAL OF NONPARAMETRIC STATISTICS
https://doi.org/10.1080/10485252.2020.1834104

Taylor & Francis
Taylor & Francis Group

Local polynomial regression for pooled response data

Dewei Wang[a], Xichen Mou[b], Xiang Li[c] and Xianzheng Huang[a]

[a]Department of Statistics, University of South Carolina, Columbia, SC, USA; [b]Division of Epidemiology, Biostatistics, and Environmental Health, University of Memphis, Memphis, TN, USA; [c]JPMorgan Chase, Jersey City, NJ, USA

ABSTRACT
We propose local polynomial estimators for the conditional mean of a continuous response when only pooled response data are collected under different pooling designs. Asymptotic properties of these estimators are investigated and compared. Extensive simulation studies are carried out to compare finite sample performance of the proposed estimators under various model settings and pooling strategies. We apply the proposed local polynomial regression methods to two real-life applications to illustrate practical implementation and performance of the estimators for the mean function.

ARTICLE HISTORY
Received 5 February 2020
Accepted 3 October 2020

KEYWORDS
Cross-validation; homogeneous pooling; random pooling

2010 MATHEMATICS SUBJECT CLASSIFICATIONS
62G08; 62G20

1. Introduction

Instead of measuring individual specimens to collect data for biomarkers or analytes of interest, collecting such data on pools of specimens has become increasingly common in epidemiological studies (Kendziorski, Zhang, Lan, and Attie 2003; Shih et al. 2004) and environmental studies (Kärrman et al. 2006; Kato et al. 2009; Heffernan et al. 2016; Mosites, Rodriguez, Caudill, Hennessy, and Berner 2020). Collecting pooled data can reduce information loss when there is a detecting limit and offer a more timely manner to gather information, in addition to the obvious benefit of reducing cost of laboratory assays and preserving irreplaceable specimens. In some econometrics applications, pooled data are all that is available to researchers, such as data aggregated by family or by region (Martinez-Espineira 2003; Fukuda 2006; Jiang, Manchanda, and Rossi 2009). In these applications, data of other attributes at the individual level are often also recorded, and researchers are interested in associations between quantities at the individual level even though some data are collected at the pool level. Our study is motivated by these research questions that require methodologies for regression analysis based on pooled continuous response data and individual-level covariate data.

CONTACT Xianzheng Huang huang@stat.sc.edu JPMorgan Chase, Jersey City, NJ 07310, USA

Supplemental data for this article can be accessed here. https://doi.org/10.1080/10485252.2020.1834104

Traditional regression methodology applicable to individual response data cannot be directly used to analyse pooled response data, and there exist some research on regression analysis for pooled continuous responses. Under the parametric framework, Malinovsky, Albert, and Schisterman (2012) considered Gaussian random effects models for pooled repeated measures and studied inference for variance components under different pooling strategies. Mitchell et al. (2014) proposed a Monte Carlo expectation maximisation algorithm to carry out regression analyses of pooled biomarker assessments assuming that the biomarker follows a log-normal distribution given covariates. McMahan, McLain, Gallagher, and Schisterman (2016) developed methods to infer receiver-operating characteristic curves using pooled biomarker measurements. Liu, McMahan, and Gallagher (2017) provided a general strategy based on Monte Carlo maximum likelihood for regression analysis of pooled data under generic parametric models assumed for the individual response given covariates. Under the semiparametric framework, Mitchell et al. (2015) proposed a semiparametric method for regression analysis of a right-skewed and positive response when data for the response are taken from pooled specimens. Without imposing parametric assumptions on the biomarker distribution, Lin and Wang (2018) developed a semiparametric approach for analysing pooled biomarker measurements originating from a single-index model for the individual response. Different from these works, we develop nonparametric estimation methods without imposing a functional form for the conditional mean of the response or a distribution family on the response given covariates. The estimation methods proposed in this article are thus more generally applicable, even though prediction based on a nonparametrically estimated mean response is less convenient than when one employs a parametric estimation method. Also under the nonparametric framework, Linton and Whang (2002) proposed a kernel-based estimator for regression function for pooled data when covariate data are also aggregated, with both aggregated response data and covariate data subject to additive measurement error. Instead of the framework of kernel regression adopted in their work, we consider in this study local polynomial regression with kernel weights that has been shown to have advantages both asymptotically and in finite sample performance over kernel regression (Fan, Heckman, and Wand 1995).

Among the existing works on regression analysis of pooled response data, many consider various pooling designs. For example, Ma, Vexler, Schisterman, and Tian (2011) compared two pooling designs in the context of linear regression analysis for a pooled continuous response and aggregated covariates, one being random pooling where pools are randomly formed without taking into account covariate information, and the other termed as optimal pooling by the authors, where pools are formed by gathering specimens corresponding to similar covariate values. This latter strategy is better known as homogeneous pooling in the pool/group testing literature (Shu and Burn 2003; Bilder and Tebbs 2009; Deckert, Brnighausen, and Kyei 2020), and many researchers have shown efficiency gain in prediction and covariate effects estimation when homogeneous pooled data are used than when random pooled data are used (Vansteelandt, Goetghebeur, and Verstraeten 2000; Ma et al. 2011). Mitchell et al. (2014) developed a regression methodology for log-normal response data subject to a special form of homogeneous pooling where covariate values within a pool are identical. Like the regression analysis discussed in Ma et al. (2011), Mitchell et al. (2014) also regressed the pooled continuous response on aggregated covariates to infer the association between the response and covariates at the individual level.

In this article, we propose local polynomial estimators for the mean of a continuous response given covariates using pooled response data and individual-level covariate data. More specifically, the proposed estimators are for the mean function $m(x) = E(Y \mid X = x)$, where Y is a continuous response of an experimental unit, X is the covariate that can be vector-valued and relate to attributes of the experimental unit or individual. For ease of exposition, we consider a scalar covariate in this article. Observed data available for inferring $m(x)$ include pooled responses from J groups of individuals, $\mathbf{Z} = (Z_1, \ldots, Z_J)^{\mathrm{T}}$, where $Z_j = c_j^{-1} \sum_{k=1}^{c_j} Y_{jk}$, in which c_j is the number of individuals in pool j, and Y_{jk} is the unobserved response of individual k in that pool, for $j = 1, \ldots, J$, $k = 1, \ldots, c_j$. Also observed are covariate data $\mathbb{X} = \{\tilde{\mathbf{X}}_j, j = 1, \ldots, J\}$, where $\tilde{\mathbf{X}}_j = (X_{j1}, \ldots, X_{j,c_j})^{\mathrm{T}}$, with X_{jk} being the covariate associated with individual k in pool j, for $k = 1, \ldots, c_j$ and $j = 1, \ldots, J$. Three proposed local polynomial estimators for $m(x)$ based on data $(\mathbf{Z}, \mathbb{X})$ are presented in Section 2 next, where we assume that data arise from random pooling. Section 3 presents local polynomial estimators based on homogeneous pooled data. Asymptotic properties of these estimators are investigated and compared in Section 4 under each of the two pooling designs. Section 5 describes bandwidth selection methods tailored for the proposed estimators. Section 6 presents a simulation study where we compare finite sample performance of the proposed estimators under different model settings and various pooling designs. We further illustrate the implementation and performance of the proposed methods in two real-life applications in Section 7. Finally, in Section 8, we summarise contributions of our study and discuss follow-up research directions.

2. Local polynomial estimators under random pooling

Local polynomial regression has been a well-received and widely applicable nonparametric strategy for estimating $m(x)$ when individual data are available (Fan and Gijbels 1996). To estimate the regression function $m(x)$ based on individual data $\{(Y_{jk}, X_{jk}), k = 1, \ldots, c_j\}_{j=1}^{J}$, this strategy exploits the weighted least squares method to construct an objective function following a pth-order Taylor expansion of $m(s)$ around x, $m(s) \approx \sum_{\ell=0}^{p} \{m^{(\ell)}(x)/\ell!\}(s-x)^{\ell}$, with $m^{(\ell)}(x)$ equal to $(\partial^{\ell}/\partial s^{\ell})m(s)$ evaluated at $s = x$. In particular, the objective function is given by

$$Q_0(\boldsymbol{\beta}) = \sum_{j=1}^{J} \sum_{k=1}^{c_j} \left\{ Y_{jk} - \sum_{\ell=0}^{p} \beta_{\ell} (X_{jk} - x)^{\ell} \right\}^2 K_h(X_{jk} - x), \qquad (1)$$

where $K_h(t) = K(t/h)/h$, $K(t)$ is a symmetric kernel, h is a bandwidth, $\beta_{\ell} = m^{(\ell)}(x)/\ell!$, for $\ell = 0, 1, \ldots, p$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$. Minimising $Q_0(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ yields an estimate of $m(x) (= \beta_0)$, along with estimates of $m^{(\ell)}(x) (= \ell! \beta_{\ell})$, for $\ell = 1, \ldots, p$. Denote by $\hat{m}_0(x)$ the so-obtained estimator for $m(x)$.

In what follows, we revise $Q_0(\boldsymbol{\beta})$ to construct new objective functions to adapt the local polynomial regression strategy to pooled response data from random pooling. Despite the pooling design considered, it is assumed that the individual data, $\{(Y_{jk}, X_{jk}), k = 1, \ldots, c_j\}_{j=1}^{J}$, consist of $N = \sum_{j=1}^{J} c_j$ independent copies multivariate random variable (Y, X) from a common distribution.

2.1. The average-weighted estimator

Now that individual responses $\{Y_{jk}, k = 1, \ldots, c_j\}_{j=1}^J$ in (1) are unobserved but pooled responses $\{Z_j\}_{j=1}^J$ are instead, it is natural to switch attention from $E(Y_i \mid X_i)$ to $E(Z_j \mid \tilde{\mathbf{X}}_j) = c_j^{-1} \sum_{k=1}^{c_j} m(X_{jk})$, as if one were regressing Z on the accompanying covariates in a pool collectively. This motivates the following weighted least squares objective function,

$$Q_1(\boldsymbol{\beta}) = \sum_{j=1}^{J} \left\{ Z_j - \sum_{\ell=0}^{p} \beta_\ell c_j^{-1} \sum_{k=1}^{c_j} (X_{jk} - x)^\ell \right\}^2 \left\{ c_j^{-1} \sum_{k=1}^{c_j} K_h(X_{jk} - x) \right\}. \qquad (2)$$

In (1), the weight function $K_h(X_i - x)$ quantifies the proximity of the ith covariate data point to x, producing a larger weight for an individual whose covariate value is closer to x. In (2), the average of such proximity measures associated with c_j covariate data points in pool j is used to assess the overall closeness of this collection of covariate values to x.

Minimising $Q_1(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ and extracting the first element of the resultant minimiser gives a pth-order local polynomial estimator for $m(x)$. This estimator can be explicitly expressed as $\hat{m}_1(x) = \boldsymbol{e}_1^{\mathrm{T}} \mathbf{S}_1^{-1}(x) \mathbf{T}_1(x)$, where $\boldsymbol{e}_1^{\mathrm{T}} = (1, 0, \ldots, 0)_{1\times(p+1)}$, $\mathbf{S}_1(x) = \mathbf{D}_1(x)^{\mathrm{T}} \mathbf{K}_1(x) \mathbf{D}_1(x)$, and $\mathbf{T}_1(x) = \mathbf{D}_1(x)^{\mathrm{T}} \mathbf{K}_1(x) \mathbf{Z}$, in which, $\mathbf{D}_1(x)$ is a $J \times (p+1)$ matrix with $\mathbf{D}_1(x)[j, \ell+1] = c_j^{-1} \sum_{k=1}^{c_j} (X_{jk} - x)^\ell$, for $j = 1, \ldots, J$, $\ell = 0, 1, \ldots, p$, and $\mathbf{K}_1(x) = \mathrm{diag}\{c_1^{-1} \sum_{k=1}^{c_1} K_h(X_{1k} - x), \ldots, c_J^{-1} \sum_{k=1}^{c_J} K_h(X_{Jk} - x)\}$. Elaborated expressions of entries in $\mathbf{S}_1(x)$ and $\mathbf{T}_1(x)$ are given in Appendix A of the supplementary materials. To highlight the weight function construction in (2), $\hat{m}_1(x)$ is referred to as the average-weighted estimator in this article.

2.2. The product-weighted estimator

Instead of averaging individual-level weights to construct a weight function as in $Q_1(\boldsymbol{\beta})$, one may view $\tilde{\mathbf{X}}_j$ as a multivariate covariate resulting from stacking the c_j individual-level covariates in pool j on top of each other, and an alternative weight function can be formulated to measure the nearness of this multivariate covariate to $x\mathbf{1}_{c_j}$, where $\mathbf{1}_{c_j}$ denotes the $c_j \times 1$ vector of one's. Mimicking the product kernel used in multivariate kernel density estimation, we propose the following weighted least squares objective function with a different weight function,

$$Q_2(\boldsymbol{\beta}) = \sum_{j=1}^{J} \left\{ Z_j - \sum_{\ell=0}^{p} \beta_\ell c_j^{-1} \sum_{k=1}^{c_j} (X_{jk} - x)^\ell \right\}^2 \left\{ \prod_{k=1}^{c_j} K_h(X_{jk} - x) \right\}. \qquad (3)$$

More succinctly, the estimator for $m(x)$ resulting from minimising $Q_2(\boldsymbol{\beta})$ is given by $\hat{m}_2(x) = \boldsymbol{e}_1^{\mathrm{T}} \mathbf{S}_2^{-1}(x) \mathbf{T}_2(x)$, where $\mathbf{S}_2(x) = \mathbf{D}_1(x)^{\mathrm{T}} \mathbf{K}_2(x) \mathbf{D}_1(x)$ and $\mathbf{T}_2(x) = \mathbf{D}_1(x)^{\mathrm{T}} \mathbf{K}_2(x) \mathbf{Z}$, in which $\mathbf{K}_2(x) = \mathrm{diag}\{\prod_{k=1}^{c_1} K_h(X_{1k} - x), \ldots, \prod_{k=1}^{c_J} K_h(X_{Jk} - x)\}$. Detailed expressions of entries in $\mathbf{S}_2(x)$ and $\mathbf{T}_2(x)$ are provided in Appendix B in the supplementary materials. Due to the construction of the weight function in (3), we call $\hat{m}_2(x)$ the product-weighted estimator in the sequel.

2.3. The marginal-integration estimator

The first two estimators are motivated by the mean of Z_j given all covariate data in pool j. The third estimator is inspired by the mean of c_jZ_j given one arbitrary individual's covariate in pool j derived next under the assumption that $Y_{jk'} \perp X_{jk}$ for $k' \neq k$ and the pools are formed randomly independent of covariate information. By the definition of Z_j, we have

$$E(c_jZ_j \mid X_{jk} = x) = \sum_{k'=1,k'\neq k}^{c_j} E(Y_{jk'} \mid X_{jk} = x) + E(Y_{jk} \mid X_{jk} = x)$$

$$= \sum_{k'=1,k'\neq k}^{c_j} E(Y_{jk'}) + m(x) = (c_j - 1)\mu + m(x),$$

where $\mu = E(Y_{jk'})$ for $k' = 1, \ldots, c_j$ and $j = 1, \ldots, J$. Hence,

$$E\{c_jZ_j - (c_j - 1)\mu \mid X_{jk} = x\} = m(x). \tag{4}$$

If one views $c_jZ_j - (c_j - 1)\mu$ as a pseudo response, (4) is reminiscent of the conditional mean model for individual-level data, $E(Y_i \mid X_i = x) = m(x)$, except for the dependence of the pseudo response on the unknown parameter μ. Since μ is the marginal mean of Y, one may use the overall sample mean response, $\hat{\mu} = N^{-1}\sum_{j=1}^{J} c_jZ_j$, to estimate μ. This yields a surrogate of the pseudo response defined by $R_j = c_jZ_j - (c_j - 1)\hat{\mu}$, for $j = 1, \ldots, J$. However, $E(R_j \mid X_{jk} = x) \neq m(x)$ due to the estimation of μ in R_j. In fact, one can show that $E(R_j \mid X_{jk} = x) = m(x) + \{\mu - m(x)\}(c_j - 1)/N$. Suggested by an anonymous referee, in each R_j, we replace $\hat{\mu}$ by $\hat{\mu}_j = \sum_{s\neq j,s=1}^{J} c_sZ_s/(N - c_j)$ and define $\hat{Y}_{jk} = c_jZ_j - (c_j - 1)\hat{\mu}_j$. One can view $\hat{Y}_{jk}$ as a bias-corrected version of R_j and also as an 'estimator' for Y_{jk} that satisfies $E(\hat{Y}_{jk} \mid X_{jk} = x) = m(x)$.

Using the surrogate of the pseudo response and (4), we formulate the following weighted least squares objective function,

$$Q_3(\boldsymbol{\beta}) = \sum_{j=1}^{J}\sum_{k=1}^{c_j}\left\{\hat{Y}_{jk} - \sum_{\ell=0}^{p}\beta_\ell(X_{jk} - x)^\ell\right\}^2 K_h(X_{jk} - x). \tag{5}$$

Minimising $Q_3(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ yields our third proposed pth-order local polynomial estimator for $m(x)$, denoted by $\hat{m}_3(x)$. As one can see from the elaborated expression of it given in Appendix C of the supplementary materials that $\hat{m}_3(x)$ is simply $\hat{m}_0(x)$ with Y_{jk} replaced by $\hat{Y}_{jk}$, for $j = 1, \ldots, J$, $k = 1, \ldots, c_j$. The construction of $\hat{m}_3(x)$ stems from the marginal integration result (4). For this reason, we refer to $\hat{m}_3(x)$ as the marginal-integration estimator henceforth. Using marginal integration is not new in the pooling literature. Lin and Wang (2018) used it to estimate a single-index model with a focus on the parametric part of their model. The asymptotic properties of $\hat{m}_3(x)$ in Section 4 do not follow their derivation directly, and additionally, they used R_j while we use $\hat{Y}_{jk}$ to correct the bias induced by R_j.

All three estimators reduce to $\hat{m}_0(x)$ when $c_j = 1$ for $j = 1, \ldots, J$ but are otherwise typically very different from each other. In-depth comparisons between the three estimators that go beyond their formulations demand more systematic investigation on their theoretical properties. This is the content of Section 4, where we look into the asymptotic bias and variance of these estimators under each of the two considered pooling designs.

3. Local polynomial estimators under homogeneous pooling

When pooled data result from homogeneous pooling, it is no longer sensible to consider the mean of c_jZ_j given one 'arbitrary' covariate data point in pool j as we just did to construct $\hat{m}_3(x)$, since individuals' covariates within a pool are not that 'arbitrary' now after all, and $E(Y_{jk'} \mid X_{jk} = x)$ is typically not equal to $E(Y_{jk'})$ for $k' \neq k$. But it is still meaningful to consider the mean of Z_j given all covariate data in pool j as we did under random pooling that leads to $\hat{m}_1(x)$ and $\hat{m}_2(x)$.

To be more concrete, consider the homogeneous pooling design following which pools of individuals are created according to the sorted covariate data in $\mathbb{X}$. This yields covariate data associated with pool j given by $\tilde{\mathbf{X}}_{(j)} = (X_{(j1)}, \ldots, X_{(jc_j)})^{\mathrm{T}}$, for $j = 1, \ldots, J$, where $X_{(11)} \leq X_{(12)} \leq \ldots \leq X_{(1c_1)} \leq X_{(21)} \leq \ldots \leq X_{(2c_2)} \leq \ldots \leq X_{(J1)} \leq \ldots \leq X_{(Jc_J)}$. Even though the response data are not sorted, we use $Z_{(j)} = c_j^{-1} \sum_{k=1}^{c_j} Y_{(jk)}$ to denote the corresponding pooled response, where $Y_{(jk)}$ is the response of the individual whose covariate value is $X_{(jk)}$, for $k = 1, \ldots, c_j$, and $j = 1, \ldots, J$. Evaluating the objective functions in (2) and (3) at $\{(Z_{(j)}, \tilde{\mathbf{X}}_{(j)})\}_{j=1}^{J}$ give the following objective functions one maximises with respect to $\boldsymbol{\beta}$ in order to obtain the average-weighted estimator, $\hat{m}_1(x)$, and the product-weighted estimator, $\hat{m}_2(x)$, respectively, under homogeneous pooling,

$$Q_1(\boldsymbol{\beta}) = \sum_{j=1}^{J} \left\{ Z_{(j)} - \sum_{\ell=0}^{p} \beta_\ell c_j^{-1} \sum_{k=1}^{c_j} (X_{(jk)} - x)^\ell \right\}^2 \left\{ c_j^{-1} \sum_{k=1}^{c_j} K_h(X_{(jk)} - x) \right\},$$

$$Q_2(\boldsymbol{\beta}) = \sum_{j=1}^{J} \left\{ Z_{(j)} - \sum_{\ell=0}^{p} \beta_\ell c_j^{-1} \sum_{k=1}^{c_j} (X_{(jk)} - x)^\ell \right\}^2 \left\{ \prod_{k=1}^{c_j} K_h(X_{(jk)} - x) \right\}.$$

4. Comparisons between different estimators

4.1. Asymptotic properties

Under certain regularity conditions listed in the supplementary materials, we derive asymptotic means and variances of the proposed estimators for $\boldsymbol{\beta}$ as $J \to \infty$ with $\max_{1 \leq j \leq J} c_j$ bounded. Conditions listed there relate to $m(x)$, the variance function $\sigma^2(x) = Var(Y \mid X = x)$, the density function of X, $f_X(x)$, and the kernel $K(t)$, which are mostly common conditions seen in the context of local polynomial regression using individual-level data. In what follows, we summarise findings from these derivations (with details provided in the supplementary materials) in two theorems that highlight some interesting contrasts between different estimators for $m(x)$ when pools are of equal size with $c_j = c$, for $j = 1, \ldots, J$, with additional conditions imposed in each theorem when needed. Several quantities appearing in these theorems are defined next for ease of reference:

$$\boldsymbol{\mu}_\ell^* = (\mu_\ell, \mu_{\ell+1}, \ldots, \mu_{\ell+p})^{\mathrm{T}}, \quad \tilde{\boldsymbol{\mu}}_\ell = [\mu_{\ell_1+\ell_2+\ell}]_{\ell_1,\ell_2=0,1,\ldots,p},$$

$$\tilde{\boldsymbol{\nu}}_0 = [\nu_{\ell_1+\ell_2}]_{\ell_1,\ell_2=0,1,\ldots,p}, \quad \mathbf{R}^*_p = (R_{0,p}(x), R_{1,p}(x), \ldots, R_{p,p}(x))^{\mathrm{T}},$$
$$\boldsymbol{\Delta}^*_0(x) = (1, \delta_1(x), \ldots, \delta_p(x))^{\mathrm{T}}, \quad \tilde{\boldsymbol{\Delta}}_0(x) = [\delta_{\ell_1+\ell_2}(x)]_{\ell_1,\ell_2=0,1,\ldots,p}, \tag{6}$$

where $R_{\ell,p}(x) = E[(X-x)^\ell\{m(X) - \sum_{\ell=0}^p \beta_\ell (X-x)^\ell\}]$ and $\delta_\ell(x) = E\{(X-x)^\ell\}$, for $\ell = 0, 1, \ldots, 2p$.

The first theorem concerns the three estimators under random pooling. Appendices A, B, and C in the supplementary materials provide the proof for the three parts of this theorem that allow unequal pool sizes.

Theorem 4.1: *As $J \to \infty$ and $h \to 0$, one has the following results regarding the difference between an estimator for $m(x)$ and $m(x)$.*

(i) *If the ℓ-th moment of X exists, for $\ell = 1, \ldots, 2p$, then*

$$\hat{m}_1(x) - m(x) = \boldsymbol{e}_1^{\mathrm{T}} \mathbf{M}_0^{-1}(x) \left\{ \mathbf{L}_0(x) - h f_X^{-1}(x) f_X'(x) \mathbf{M}_1(x) \mathbf{M}_0^{-1}(x) \mathbf{L}_0(x) + O\left(h^2\right) \right\} + \sqrt{c} \times O_P\left(\frac{1}{\sqrt{Nh}}\right), \tag{7}$$

where

$$\mathbf{L}_0(x) = \frac{c-1}{c^2}\left\{R_{0,p}(x)\boldsymbol{e}_1 + \mathbf{R}^*_p(x)\right\} + \frac{(c-1)(c-2)_+}{c^2} R_{0,p}(x)\boldsymbol{\Delta}^*_0(x),$$
$$\mathbf{M}_0(x) = \frac{\tilde{\boldsymbol{\mu}}_0}{c^2} + \frac{c-1}{c^2}\left\{\tilde{\boldsymbol{\Delta}}_0(x) + \boldsymbol{\Delta}^*_0(x)\boldsymbol{\mu}_0^{*\mathrm{T}} + \boldsymbol{\mu}^*_0\boldsymbol{\Delta}_0^{*\mathrm{T}}(x)\right\} + \frac{(c-1)(c-2)_+}{c^2}\boldsymbol{\Delta}^*_0(x)\boldsymbol{\Delta}_0^{*\mathrm{T}}(x),$$
$$\mathbf{M}_1(x) = \frac{\tilde{\boldsymbol{\mu}}_1}{c^2} + \frac{c-1}{c^2}\left\{\boldsymbol{\Delta}^*_0(x)\boldsymbol{\mu}_1^{*\mathrm{T}} + \boldsymbol{\mu}^*_1\boldsymbol{\Delta}_0^{*\mathrm{T}}(x)\right\},$$

in which $(t)_+ = \max(t, 0)$.

(ii) *If $m(x)$ is $(p+3)$th-order continuously differentiable, then*

$$\begin{aligned}
&\hat{m}_2(x) - m(x) \\
&= \boldsymbol{e}_1^{\mathrm{T}} h^{p+1} \Bigg\{ \beta_{p+1}\left\{\tilde{\boldsymbol{\mu}}_0 + (c-1)\boldsymbol{\mu}^*_0\boldsymbol{\mu}_0^{*\mathrm{T}}\right\}^{-1}\left\{\boldsymbol{\mu}^*_{p+1} + (c-1)\mu_{p+1}\boldsymbol{\mu}^*_0\right\} \\
&\quad + h f_X^{-1}(x)\Big[\left\{\beta_{p+2}f_X(x) + \beta_{p+1}f_X'(x)\right\}\left\{\tilde{\boldsymbol{\mu}}_0 + (c-1)\boldsymbol{\mu}^*_0\boldsymbol{\mu}_0^{*\mathrm{T}}\right\}^{-1} \\
&\quad \times \left\{\boldsymbol{\mu}^*_{p+2} + (c-1)\mu_{p+2}\boldsymbol{\mu}^*_0\right\} - \beta_{p+1}f_X'(x)\left\{\tilde{\boldsymbol{\mu}}_0 + (c-1)\boldsymbol{\mu}^*_0\boldsymbol{\mu}_0^{*\mathrm{T}}\right\}^{-1} \\
&\quad \times \left\{\tilde{\boldsymbol{\mu}}_1 + (c-1)\left(\boldsymbol{\mu}^*_0\boldsymbol{\mu}_1^{*\mathrm{T}} + \boldsymbol{\mu}^*_1\boldsymbol{\mu}_0^{*\mathrm{T}}\right)\right\}\left\{\tilde{\boldsymbol{\mu}}_0 + (c-1)\boldsymbol{\mu}^*_0\boldsymbol{\mu}_0^{*\mathrm{T}}\right\}^{-1} \\
&\quad \times \left\{\boldsymbol{\mu}^*_{p+1} + (c-1)\mu_{p+1}\boldsymbol{\mu}^*_0\right\}\Big] + O(h^2)\Bigg\} + \sqrt{c} \times O_P\left(\frac{1}{\sqrt{Nh^c}}\right).
\end{aligned}$$

(iii) *Let $\bar{\sigma}^2 = E\{\sigma^2(X)\}$. If $Var(Y)$ exists, then*

$$\hat{m}_3(x) - m(x) = \boldsymbol{e}_1^{\mathrm{T}} h^{p+1}\Big\{\beta_{p+1}\tilde{\boldsymbol{\mu}}_0^{-1}\boldsymbol{\mu}^*_{p+1} + h f_X^{-1}(x)\left[\left\{\beta_{p+2}f_X(x)\right.\right.$$

$$+\beta_{p+1} f'_X(x)\big\}\, \tilde{\boldsymbol{\mu}}_0^{-1} \boldsymbol{\mu}^*_{p+2} - \beta_{p+1} f'_X(x) \tilde{\boldsymbol{\mu}}_0^{-1} \tilde{\boldsymbol{\mu}}_1 \tilde{\boldsymbol{\mu}}_0^{-1} \boldsymbol{\mu}^*_{p+1}\Big]$$

$$+O(h^2)\big\} + \sqrt{\sigma^2(x) + (c-1)\bar{\sigma}^2} \times O_P\left(\frac{1}{\sqrt{Nh}}\right). \tag{8}$$

Theorem 4.1-(i) indicates that $\hat{m}_1(x)$ is an inconsistent estimator for $m(x)$, with the dominating bias given by $\boldsymbol{e}_1^{\mathrm{T}} \mathbf{M}_0^{-1}(x) \mathbf{L}_0(x)$ that does not depend on h, and thus does not diminish as $h \to 0$, but it does vanish when $c = 1$. Considering a local constant estimator by setting $p = 0$ in (7), we show in Appendix A in the supplementary material that

$$\begin{aligned}
&\hat{m}_1(x) - m(x) \\
&\quad = \frac{c-1}{c} E\{m(X) - m(x)\} + \frac{h^2 \mu_2}{c} \left\{\beta_1 \frac{f'_X(x)}{f_X(x)} + \beta_2\right\} + O(h^4) + O_P\left(\frac{1}{\sqrt{Jh}}\right),
\end{aligned} \tag{9}$$

of which the second term (of order h^2) is c^{-1} times the dominating bias of the Nadaraya-Watson estimator based on individual-level data. Observing that the dominating bias in (9) is equal to $c^{-1}(c-1)\{\mu - m(x)\}$, one can easily derive an improved local constant estimator by correcting $\hat{m}_1(x)$ for this dominating bias. This leads to a consistent local constant estimator given by $c\hat{m}_1(x) - (c-1)\hat{\mu}$, of which the bias is of order $O_P(h^2)$. For $p > 0$, correcting $\hat{m}_1(x)$ for its dominating bias requires estimating functionals of $m(x)$ more involved than $\mu = E\{m(X)\}$ that appear in $\mathbf{R}_p^*$ in (6).

Theorem 4.1-(ii) suggests that $\hat{m}_2(x)$ is a consistent estimator for $m(x)$ with the asymptotic variance of order $O\{1/(Jh^c)\}$, which inflates quickly as c increases. It is worth pointing out that, $Q_2(\boldsymbol{\beta})$ in (3) is essentially a special form of the objective function associated with the regular multivariate local polynomial estimator for the c-variate conditional mean $m^*(x_1, \ldots, x_c) \triangleq c^{-1} \sum_{k=1}^{c} m(x_k)$ based on individual-level multivariate covariate data of c dimensional. Hence, following Masry (1996) and Gu, Li, and Yang (2015), under the same set of regularity conditions listed in the supplementary materials, $\hat{m}^*(x\mathbf{1}_c) = \hat{m}_2(x)$ is asymptotically normal when $Jh^{c+2p} \to \infty$ and $Jh^{c+2p+6} \to 0$ as $J \to \infty$.

Comparing Theorem 4.1-(ii) and (iii) reveals that $\hat{m}_2(x)$ and $\hat{m}_3(x)$ typically do not share the same dominating bias except when $c = 1$, and $\hat{m}_3(x)$ exhibits the same asymptotic bias as that of $\hat{m}_0(x)$ regardless of the pool size. The variability of $\hat{m}_3(x)$ is understandably higher than that of $\hat{m}_0(x)$, but it only grows linearly in c and thus is much less inflated than the variance of $\hat{m}_2(x)$. More specifically, (8) implies that the amount of variance inflation of $\hat{m}_3(x)$ depends linearly on the pool size and $\bar{\sigma}^2$.

Summarising the above remarks on Theorem 4.1, we conclude that the marginal-integration estimator $\hat{m}_3(x)$ is the preferred estimator among the three proposed under random pooling. It outperforms the average-weighted estimator $\hat{m}_1(x)$ for its consistency, and it surpasses the product-weighted estimator $\hat{m}_2(x)$ for its much less inflated variance when compared with $\hat{m}_0(x)$. In practice, we recommend using the local linear version of $\hat{m}_3(x)$ which corresponds to $p = 1$. For this case, Theorem 4.1-(iii) yields that

$$\mathrm{Bias}\{\hat{m}_3(x) \mid \mathbb{X}\} = \left\{\frac{1}{2} m''(x) \mu_2 h^2 + \sqrt{c-1}\, O_P\left(\frac{1}{\sqrt{Nh}}\right)\right\} \{1 + o_P(1)\},$$

$$\mathrm{Var}\{\hat{m}_3(x) \mid \mathbb{X}\} = \frac{\nu_0}{Nh} \frac{\sigma^2(x) + \bar{\sigma}^2(c-1)}{f_X(x)} \{1 + o_P(1)\}.$$

Thus the mean squared error of $\hat{m}_3(x)$ is $O_P(h^4 + 1/Nh)$ which attains the optimal nonparametric rate $O_P(N^{-4/5})$ when $h = O(N^{-1/5})$ is used. Furthermore, if there exists $\eta > 0$ such that $E\{|cZ_j - m(x) - (c-1)\mu|^{2+\eta} \,|\, X_{j1} = x_1, \ldots, X_{jc} = x_c\}$ is bounded for all $x_1, \ldots, x_c$, then we have $[\hat{m}_3(x) - m(x) - \text{Bias}\{\hat{m}_3(x)\}]/\sqrt{\text{Var}\{\hat{m}_3(x)\}}$ converges in distribution to $N(0,1)$ as N goes to infinity.

Despite these virtues of $\hat{m}_3(x)$, it is no longer well justified under homogeneous pooling as pointed out in Section 3. The following theorem is regarding the average-weighted estimator and the product-weighted estimator applied to data from the homogeneous pooling design. Appendix D in the supplementary material provides the proof for this theorem.

Theorem 4.2: *Assume that x is an interior point of a compact and nondegenerate interval $\mathcal{I}$, the pdf of X, $f_X(\cdot)$, is bounded away from zero on an interval $\mathcal{J}$, where $\mathcal{I} \subset \mathcal{J}$, and $K(|t|) = 0$ for $|t| > 1$, with $K'(t)$ bounded. Then, as $J \to \infty$, $h \to 0$, and $Jh^4 \to \infty$,*

$$\hat{m}_1(x) - m(x) = \text{Bias}(\hat{m}_1(x)|\mathbb{X}) + O_P(\sqrt{Var\left\{\hat{m}_1(x)|\mathbb{X}\right\}})$$

$$\begin{aligned}
&\text{Bias}(\hat{m}_1(x) \,|\, \mathbb{X}) \\
&\quad = \boldsymbol{e}_1^{\mathrm{T}} h^{p+1} \Big\{ \beta_{p+1} \tilde{\boldsymbol{\mu}}_0^{-1} \boldsymbol{\mu}_{p+1}^* + h f_X^{-1}(x) \Big[\left\{ \beta_{p+2} f_X(x) + \beta_{p+1} f_X'(x) \right\} \tilde{\boldsymbol{\mu}}_0^{-1} \boldsymbol{\mu}_{p+2}^* \\
&\qquad - \beta_{p+1} f_X'(x) \tilde{\boldsymbol{\mu}}_0^{-1} \tilde{\boldsymbol{\mu}}_1 \tilde{\boldsymbol{\mu}}_0^{-1} \boldsymbol{\mu}_{p+1}^* \Big] + O(h^2) \Big\},
\end{aligned} \tag{10}$$

and

$$Var\left\{\hat{m}_1(x) \,|\, \mathbb{X}\right\} = \frac{\sigma^2(x)}{Nhf_X(x)} \boldsymbol{e}_1^{\mathrm{T}} \tilde{\boldsymbol{\mu}}_0^{-1} \tilde{\boldsymbol{\nu}}_0 \tilde{\boldsymbol{\mu}}_0^{-1} \left\{1 + o_P(1)\right\}. \tag{11}$$

If Condition (C5) is satisfied for the kernel defined by $K^\dagger(t) = K^c(t)$, then

$$\hat{m}_2(x) - m(x) = Bias(\hat{m}_2(x) \,|\, \mathbb{X}) + O_P(\sqrt{Var\left\{\hat{m}_2(x) \,|\, \mathbb{X}\right\}})$$

$$\begin{aligned}
&Bias(\hat{m}_2(x) \,|\, \mathbb{X}) \\
&\quad = \boldsymbol{e}_1^{\mathrm{T}} h^{p+1} \Big\{ \beta_{p+1} \tilde{\boldsymbol{\mu}}_{\dagger,0}^{-1} \boldsymbol{\mu}_{\dagger,p+1}^* + h f_X^{-1}(x) \Big[\left\{ \beta_{p+2} f_X(x) + \beta_{p+1} f_X'(x) \right\} \tilde{\boldsymbol{\mu}}_{\dagger,0}^{-1} \boldsymbol{\mu}_{\dagger,p+2}^* \\
&\qquad - \beta_{p+1} f_X'(x) \tilde{\boldsymbol{\mu}}_{\dagger,0}^{-1} \tilde{\boldsymbol{\mu}}_{\dagger,1} \tilde{\boldsymbol{\mu}}_{\dagger,0}^{-1} \boldsymbol{\mu}_{\dagger,p+1}^* \Big] + O(h^2) \Big\},
\end{aligned} \tag{12}$$

and

$$\text{Var}\left\{\hat{m}_2(x) \,|\, \mathbb{X}\right\} = \frac{\sigma^2(x)}{Nhf_X(x)} \boldsymbol{e}_1^{\mathrm{T}} \tilde{\boldsymbol{\mu}}_{\dagger,0}^{-1} \tilde{\boldsymbol{\nu}}_{\dagger,0} \tilde{\boldsymbol{\mu}}_{\dagger,0}^{-1} \left\{1 + o_P(1)\right\}, \tag{13}$$

where $\boldsymbol{\mu}_{\dagger,\ell}^*$, $\tilde{\boldsymbol{\mu}}_{\dagger,\ell}$, and $\tilde{\boldsymbol{\nu}}_{\dagger,0}$ are the counterparts of $\boldsymbol{\mu}_\ell^*$, $\tilde{\boldsymbol{\mu}}_\ell$, and $\tilde{\boldsymbol{\nu}}_0$, respectively, with $K(t)$ replaced by $K^\dagger(t)$.

Among the additional assumptions imposed in Theorem 4.2, the one on x and the assumption on $K(t)$ are similar to Conditions (T1) and (T5) in Delaigle and Hall (2012), respectively. Theorem 4.2 indicates that both $\hat{m}_1(x)$ and $\hat{m}_2(x)$ are consistent estimators for $m(x)$ under homogeneous pooling, with the former sharing the same dominating bias as that of $\hat{m}_0(x)$, and the latter exhibiting the same form of dominating bias with a redefined kernel that depends on c. Moreover, the asymptotic variances of both estimators are of the same order as that of $\hat{m}_0(x)$ despite the pool size. The practical implication of Theorem 4.2 is that, if one uses homogeneous pooled data to infer $m(x)$ via either one of the two proposed local polynomial estimators, one only needs J assays without losing accuracy or efficiency asymptotically compared with when un-pooled data are used that require $N = cJ$ assays.

In practice, we recommend using $\hat{m}_1(x)$ or $\hat{m}_2(x)$ with $p = 1$. For $\hat{m}_1(x)$ when $p = 1$, Theorem 4.2 implies that

$$\mathrm{Bias}\{\hat{m}_1(x)\,|\,\mathbb{X}\} = \frac{1}{2}m''(x)\mu_2 h^2\{1 + o_P(1)\},$$

$$\mathrm{Var}\{\hat{m}_1(x)\,|\,\mathbb{X}\} = \frac{\nu_0}{Nh}\frac{\sigma^2(x)}{f_X(x)}\{1 + o_P(1)\}.$$

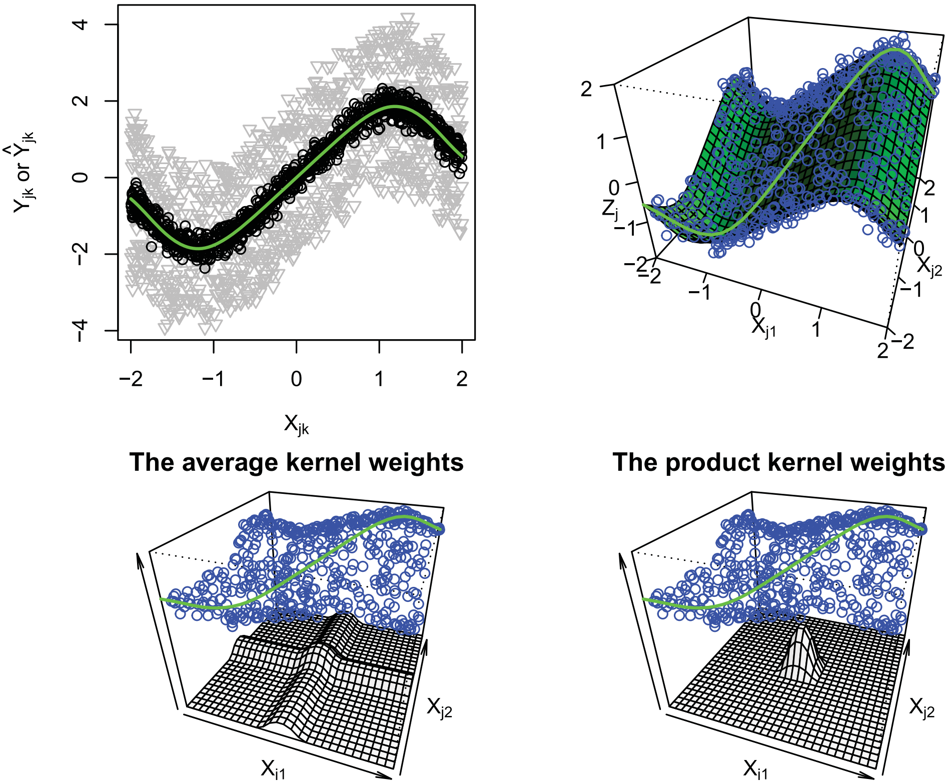
The mean squared error of $\hat{m}_1(x)$ is also $O_P(h^4 + 1/Nh)$ which attains the optimal nonparametric rate $O_P(N^{-4/5})$ when $h = O(N^{-1/5})$ is used. Furthermore, if there exists $\eta > 0$ such that $E\{|Y_{jk} - m(x)|^{2+\eta}\,|\,X_{jk} = x\}$ is bounded for all x, we have $[\hat{m}_1(x) - m(x) - \mathrm{Bias}\{\hat{m}_1(x)\}]/\sqrt{\mathrm{Var}\{\hat{m}_1(x)\}}$ converges in distribution to $N(0, 1)$ as $N \to \infty$. Similar properties hold for $\hat{m}_2(x)$.

4.2. Further remarks

We are now in the position to reflect on the findings in Theorems 4.1 and 4.2 to gain a deeper understanding of the three proposed estimators for $m(x)$ using pooled data.

The stark contrast between properties of the average-weighted estimator under the two pooling designs may seem peculiar at first glance. As natural as it initially appears to be, the use of average weights is the root cause for the persistent bias of $\hat{m}_1(x)$ under random pooling. For ease of exposition, assume for the time being $c_j = 2$, for $j = 1, \ldots, J$. The objective function $Q_1(\boldsymbol{\beta})$ in (2) associated with $\hat{m}_1(x)$ is essentially constructed for estimating $m^*(x_1, x_2) \triangleq \{m(x_1) + m(x_2)\}/2$ evaluated at $(x_1, x_2) = x\mathbf{1}_2$. The same weight, $\{K_h(X_{j1} - x) + K_h(X_{j2} - x)\}/2$, is assigned to both individuals in pool j whose covariate values are $\tilde{\mathbf{X}}_j = (X_{j1}, X_{j2})^{\mathrm{T}}$. This can yield misleading weight when, for example, X_{j1} is close to x but X_{j2} is far away from x, which can often happen under random pooling. In contrast, the product weight in $Q_2(\boldsymbol{\beta})$ in (3) associated with $\hat{m}_2(x)$ avoids such misleading weighting scheme because $K_h(X_{j1} - x)K_h(X_{j2} - x)$ is small if either one of the two individual weights is small, and thus $\tilde{\mathbf{X}}_j$ will only contribute more in estimating $m^*(x, x) = m(x)$ when both X_{j1} and X_{j2} are closer to x. In particular, when $K(t)$ is the Gaussian kernel, the product weight function amounts to evaluating the bivariate Gaussian density function at the Euclidean distance between $\tilde{\mathbf{X}}_j$ and $x\mathbf{1}_2$, whereas the average weight function lacks such connection with a meaningful distance measure between the two points in $\mathbb{R}^2$.

Even though $\hat{m}_2(x)$ exploits a more sensible weight function when comparing with $\hat{m}_1(x)$ under random pooling, downplaying X_{j1} even when it is close to x simply because the covariate value of the other individual in the same pool is far away from x is not an efficient use of data. And such waste of data information is more severe when the pool size is bigger, which is essentially the curse-of-dimensionality when one estimates the multivariate function $m^*(x\mathbf{1}_c)$ based on a response along with a c-dimensional covariate. It is such inefficient use of data that causes the much inflated variance concluded in Theorem 4.1 for $\hat{m}_2(x)$. Figure 1 illustrates the average weight function and the product weight function (in bottom panels) under random pooling when $c = 2$ and $K(t)$ is the Epanechnikov kernel. Also shown in Figure 1 (see the top-left panel) are individual-level data generated according to the model specified in (D1) described in Section 6, overlaid with the pseudo response data from random pooling, which are used for the construction of $\hat{m}_3(x)$. From there one can see that the pseudo data, $\{(\hat{Y}_{jk}, X_{jk}), k = 1, 2\}_{j=1}^{J}$, are much more variable than the original data used to obtain $\hat{m}_0(x)$, and thus the increased variance of $\hat{m}_3(x)$ is expected when



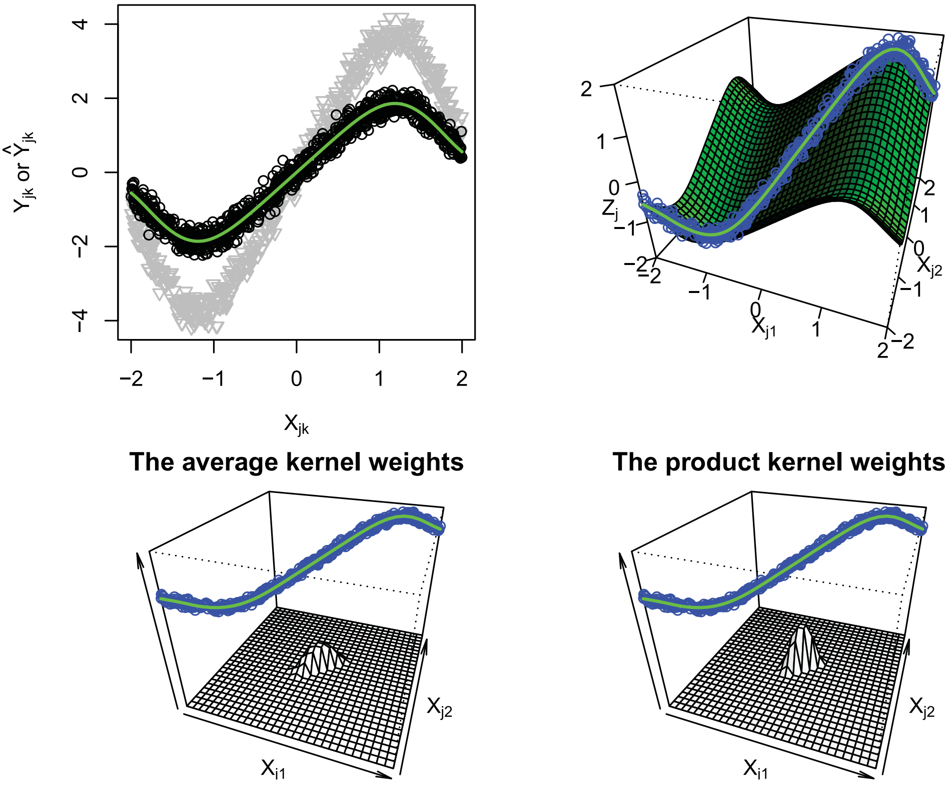
The average kernel weights

The product kernel weights

Figure 1. Plots under random pooling. Top-left panel: Individual-level data $\{(Y_{jk}, X_{jk}), k = 1, 2\}_{j=1}^{J}$ as circles, pseudo individual-level responses and covariate data $\{(\hat{Y}_{jk}, X_{jk}), k = 1, 2\}_{j=1}^{J}$ as triangles, overlaid with the true $m(x)$ as the curve running through circles. Top-right panel: the bivariate function $m^*(x_1, x_2) = \{m(x_1) + m(x_2)\}/2$ as the curved surface, with its value evaluated at (x, x), i.e. $m^*(x, x) = m(x)$, highlighted as the curve running through the surface, overlaid with the pool-level data $\{(X_{j1}, X_{j2}, Z_j)\}_{j=1}^{J}$ as circles. Bottom-left panel: the shape of the average kernel weights $\{[K(\{X_{j1} - x\}/h) + K(\{X_{j2} - x\}/h)]/2\}_{j=1}^{J}$ when $x = 0$, along with $m^*(x, x)$ and the pool-level data. Bottom-right panel: the shape of the product kernel weights $\{[K(\{X_{j1} - x\}/h)K(\{X_{j2} - x\}/h)]/2\}_{j=1}^{J}$ when $x = 0$, along with $m^*(x, x)$ and the pool-level data.

compared with $\hat{m}_0(x)$. Despite the higher variability, the pseudo data cloud does preserve the overall pattern of the original data cloud, which explains the common dominating bias shared between $\hat{m}_3(x)$ and $\hat{m}_0(x)$. Unlike $Q_1(\boldsymbol{\beta})$ and $Q_2(\boldsymbol{\beta})$, the construction of $Q_3(\boldsymbol{\beta})$ in (5) is directly designed for estimating the univariate function $m(x)$ instead of $m^*(x\mathbf{1}_c)$, and thus $\hat{m}_3(x)$ overcomes the pitfall of misleading weight assignment in $\hat{m}_1(x)$, as well as the curse-of-dimensionality that $\hat{m}_2(x)$ suffers.

Figure 2 is the counterpart of Figure 1 under homogeneous pooling. Now one can see (in the top-left panel) in Figure 2 that the pseudo data, $\{(\hat{Y}_{(jk)}, X_{(jk)}), k = 1, 2\}_{j=1}^{N}$, clearly distort the original data pattern, and thus are inappropriate for estimating $m(x)$. With individuals sharing similar covariates values gathering in the same pool, the concern relating to $\hat{m}_1(x)$ of assigning inadequate weight no longer exists, neither does the concern relating to $\hat{m}_2(x)$ of inefficient use of data. The bottom panels of Figure 2 depict the average weight



The average kernel weights

The product kernel weights

Figure 2. Plots under homogeneous pooling. Top-left panel: Individual-level data $\{(Y_{(jk)}, X_{(jk)}), k = 1, 2\}_{j=1}^{J}$ as circles, pseudo individual-level responses and covariate data $\{(\hat{Y}_{(jk)}, X_{(jk)}), k = 1, 2\}_{j=1}^{J}$ as triangles, overlaid with the true $m(x)$ as the curve running through circles. Top-right panel: the bivariate function $m^*(x_1, x_2) = \{m(x_1) + m(x_2)\}/2$ as the curved surface, with its value evaluated at (x, x), i.e. $m^*(x, x) = m(x)$, highlighted as the curve running through the surface, overlaid with pool-level data $\{(X_{j1}, X_{j2}, Z_j)\}_{j=1}^{J}$ as circles. Bottom-left panel: the shape of the average kernel weights $\{[K(\{X_{j1} - x\}/h) + K(\{X_{j2} - x\}/h)]/2\}_{j=1}^{J}$ when $x = 0$, along with $m^*(x, x)$ and the pool-level data. Bottom-right panel: the shape of the product kernel weights $\{[K(\{X_{j1} - x\}/h)K(\{X_{j2} - x\}/h)]/2\}_{j=1}^{J}$ when $x = 0$, along with $m^*(x, x)$ and the pool-level data.

function and the product weight function, both are reminiscent of some symmetric kernel function.

5. Bandwidth selection

The choice of bandwidths in local polynomial estimators plays a key role in the performance of these estimators. Besides the usual challenges encountered in bandwidth selection in local polynomial regression, a unique complication we face here is the lack of individual-level response data, which makes loss functions used for bandwidth selection that are based on individual-level residuals (or prediction errors) inapplicable in our context. Next we develop leave-one-pool-out cross-validation (CV) procedures to choose bandwidths in three proposed local polynomial estimators for $m(x)$ using random pooled data.

For the average-weighted estimator, $\hat{m}_1(x)$, we choose the bandwidth h that minimises the following pool-level residual sum of squares,

$$\mathrm{RSS}_1(h) = \sum_{j=1}^{J} \sum_{k=1}^{c_j} \left\{ Z_j - c_j^{-1} \sum_{k=1}^{c_j} \hat{m}_{1,h}^{(-j)}(X_{jk}) \right\}^2, \tag{14}$$

where $\hat{m}_{1,h}^{(-j)}(X_{jk})$ is the realisation of $\hat{m}_1(X_{jk})$ based on the observed data $(\mathbf{Z}, \mathbb{X})$ excluding data from pool j, $(Z_j, \tilde{\mathbf{X}}_j)$, with the bandwidth set at h. The bandwidth in the product-weighted estimator, $\hat{m}_2(x)$, is chosen by minimising a CV criterion similarly defined as (14),

$$\mathrm{RSS}_2(h) = \sum_{j=1}^{J} \sum_{k=1}^{c_j} \left\{ Z_j - c_j^{-1} \sum_{k=1}^{c_j} \hat{m}_{2,h}^{(-j)}(X_{jk}) \right\}^2. \tag{15}$$

Admittedly, CV criteria or loss functions constructed based on prediction errors at the pool level may not be sensitive to the influence of h on prediction power at the individual level and thus may not serve as effective model criteria for the purpose for choosing bandwidths.

Given (14) and (15), one can easily envision a similar CV criterion, denoted by $\mathrm{RSS}_3(h)$, defined for choosing h in $\hat{m}_3(x)$. We however take into account the close tie between $\hat{m}_3(x)$ and local polynomial estimators designed for individual-level data and propose a new and more effective CV criterion. This new criterion tailored for $\hat{m}_3(x)$ is mostly thanks to the pseudo individual-level observations, $\{(\hat{Y}_{jk}, X_{jk}),\ k = 1, \ldots, c\}_{j=1}^{J}$, used in $\hat{m}_3(x)$. In particular, we choose h used in $\hat{m}_3(x)$ that minimises the following pseudo (individual-level) residual sum of squares,

$$\mathrm{PRSS}_3(h) = \sum_{j=1}^{J} \sum_{k=1}^{c_j} \{\hat{Y}_{jk} - \hat{m}_{3,h}^{(-j)}(X_{jk})\}^2, \tag{16}$$

where $\hat{m}_{3,h}^{(-j)}(X_{jk})$ is the realisation of $\hat{m}_3(X_{jk})$ based on the pseudo individual-level data excluding data from pool j, $(Z_j, X_{j1}, \ldots, X_{jc_j})$, with bandwidth set at h. Empirical evidence suggest that $\mathrm{PRSS}_3(h)$ is a more effective CV criterion for bandwidth selection than $\mathrm{RSS}_3(h)$.

6. Simulation study

6.1. Design of simulation experiments

To compare different estimators of $m(x)$ in regard to their finite sample performance and to explore other factors that may influence the estimation, we carry out an empirical study using synthetic data. More specifically, we adopt the following data generating processes reported in Delaigle, Fan, and Carroll (2009) to generate individual-level response data:

(D1) $m(x) = x^3 \exp(x^4/1000) \cos x$, $\epsilon \sim N(0, 0.6^2)$, $X \sim 0.8X_1 + 0.2X_2$, where X_1 follows a distribution with pdf given by $0.1875x^2 I(-2 \le x \le 2)$ and $X_2 \sim$ uniform $(-1, 1)$;
(D2) $m(x) = 2x \exp(-10x^4/81)$, $\epsilon \sim (0, 0.2^2)$, $X \sim 0.8X_1 + 0.2X_2$, where the distributions of X_1 and X_2 are as those specified in (D1);
(D3) $m(x) = x^3, \epsilon \sim N(0, 1.2^2), X \sim N(0, 1)$;
(D4) $m(x) = x^4, \epsilon \sim N(0, 4^2), X \sim N(0, 1)$.

Under each generating process, we generate individual-level data, $\{(Y_i, X_i) : i = 1, \ldots, N\}$, where $N \in \{600, 1200\}$. Given an individual-level data set, we create pooled data, first using random pooling and then using homogeneous pooling, with a common pool size $c = 2, 3, 4, 5, 6$ across all J pools. Given each pooled data set, we obtain three local linear estimates for the mean function, $\hat{m}_1(x)$, $\hat{m}_2(x)$, and $\hat{m}_3(x)$. In addition, we also compute the local linear estimate using individual-level data, $\hat{m}_0(x)$, as a benchmark estimate. In all four estimators, we set $K(t)$ as the Epanechnikov kernel. The empirical integrated squared error (ISE) is the metric we use to assess the overall quality of an estimated mean function, defined by ISE $= N^{-1} \sum_{j=1}^{J} \sum_{k=1}^{c} \{m(X_{jk}) - \hat{m}(X_{jk})\}^2$ for an estimator $\hat{m}(\cdot)$. Additionally, we monitor in the simulation the pointwise empirical bias and standard error of each estimate for $m(\cdot)$.

6.2. Simulation results

We summarise in this section simulation results when individual-level data are generated according to (D2) with $N = 600$. Counterparts results relating to (D1), (D3), and (D4) are provided in Appendix E in the supplementary materials.

More specifically, Figure 3 shows boxplots of 500 realisations of ISE associated with the two proposed consistent estimators based on random pooling data, $\hat{m}_2(x)$ and $\hat{m}_3(x)$, and those corresponding to the two proposed consistent estimators based on homogeneous pooling data, $\hat{m}_1(x)$ and $\hat{m}_2(x)$, at each considered pool size, all comparing with ISEs of $\hat{m}_0(x)$. Evidently, when homogeneous pooling data are used, the overall performances of the two proposed estimators are similar to the benchmark estimator based on individual-level data, $\hat{m}_0(x)$. In contrast, when random pooling data are used, albeit consistent, both $\hat{m}_1(x)$ and $\hat{m}_2(x)$ exhibit much higher ISE than $\hat{m}_0(x)$ does, especially when the pool size is larger.

Instead of the overall performance of a consistent estimator over a range of covaraite values, Figures 4 and 5 depict the pointwise performance of all four estimators, $\hat{m}_0(x)$, $\hat{m}_1(x)$, $\hat{m}_2(x)$, and $\hat{m}_3(x)$, in regard to bias, variance, and mean squared error (MSE), when $c = 2$ and $c = 5$, respectively. Under random pooling (see upper panels of Figures 4

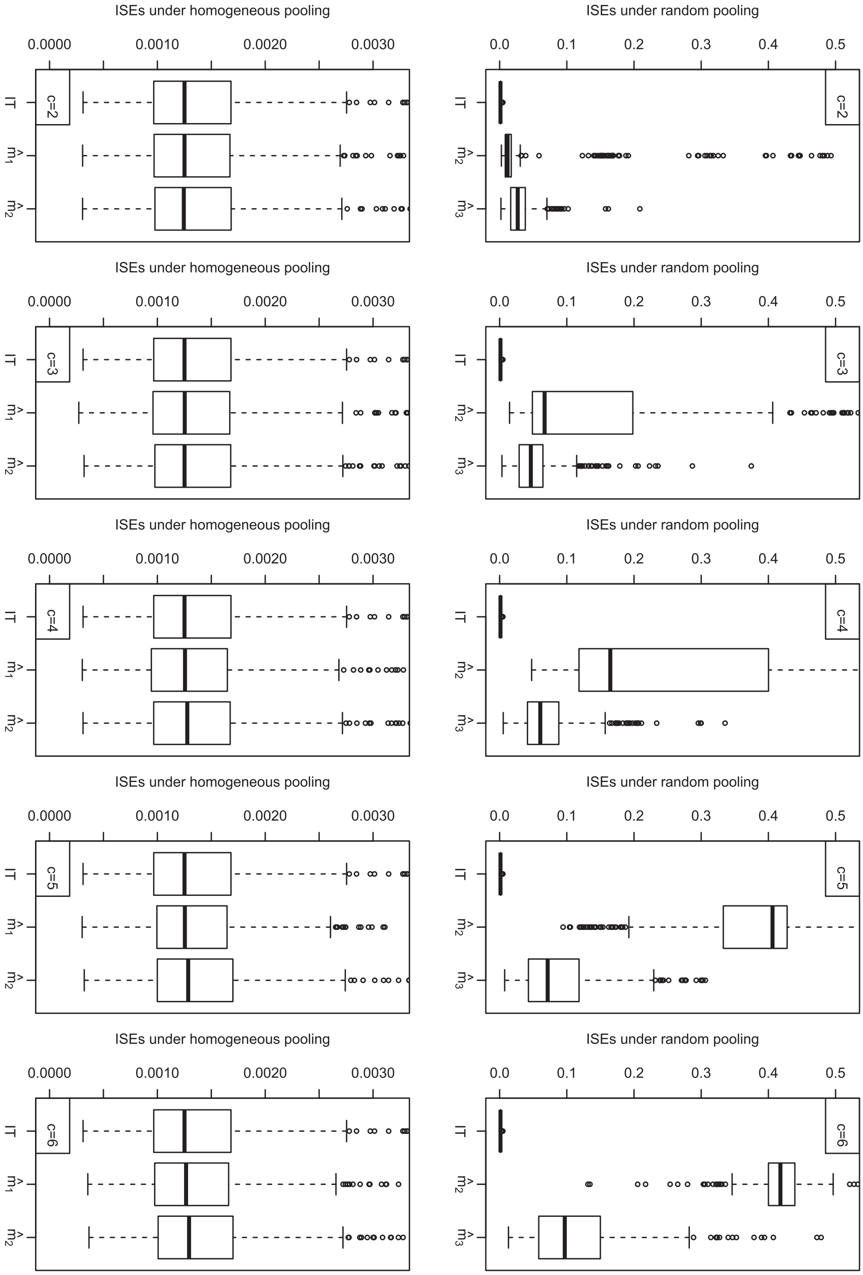


Figure 3. Boxplots of 500 ISEs associated with each of the consistent local linear estimators for $m(x)$ based on random pooling data (upper panels) and those based on homogeneous pooling data (lower panels) under (D2) at each pool size configuration, all comparing with boxplots of ISEs associated with the local linear estimator based on individual-level data (IT). Consistent estimators based on random pooling data include the product-weighted estimator, $\hat{m}_2(x)$, and the marginal-integration estimator, $\hat{m}_3(x)$. Consistent estimators based on homogeneous pooling data include the average-weight estimator, $\hat{m}_1(x)$, and $\hat{m}_2(x)$.

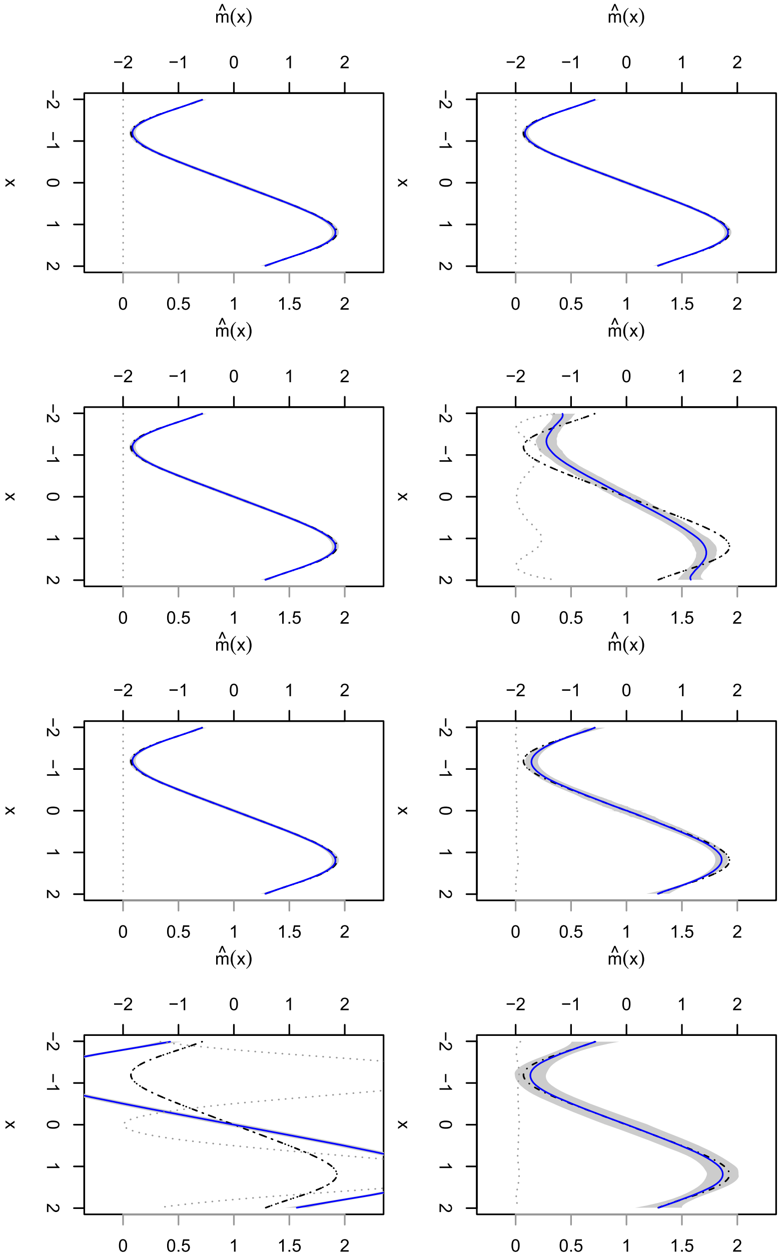


Figure 4. Four estimates for $m(x)$ under (D2) when $c = 2$: the local linear estimate based on individual-level data (the first column), $\hat{m}_0(x)$, the average-weighted estimate (the second column), $\hat{m}_1(x)$, the product-weighted estimate (the third column), $\hat{m}_2(x)$, and the marginal-integration estimate (the fourth column), $\hat{m}_3(x)$. The latter three estimates are based on random pooled data in the upper panels, and are based on homogeneous pooled data in the lower panels. Within each panel, the dot-dashed curve is the true function $m(x)$, the blue curve is the pointwise mean curve based on the 500 function estimates, the grey band is constructed by the mean curve plus and minus 1.96 times the pointwise standard deviation curve, and the dotted lines provides a comparison of the pointwise mean squared error curve across different estimates plotted with respect to the right axis of each subfigure.

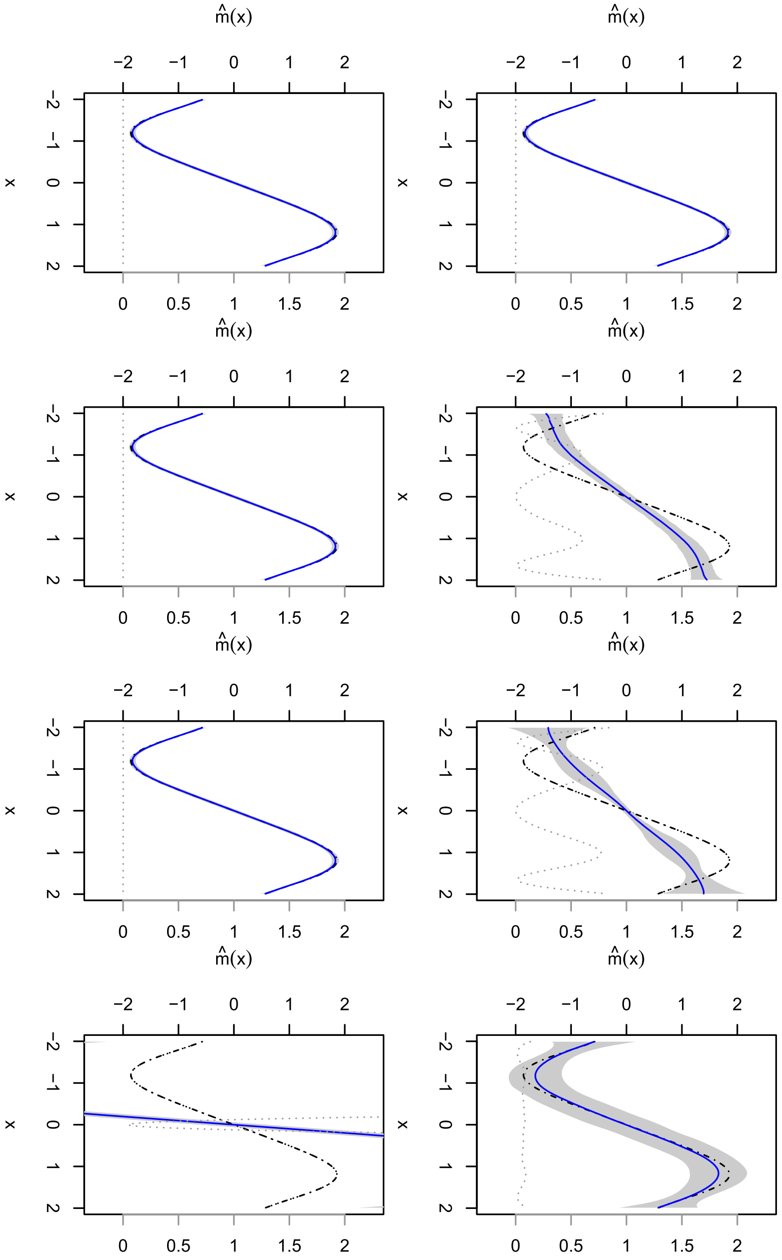


Figure 5. Four estimates for $m(x)$ under (D2) when $c = 5$: the local linear estimate based on individual-level data (the first column), $\hat{m}_0(x)$, the average-weighted estimate (the second column), $\hat{m}_1(x)$, the product-weighted estimate (the third column), $\hat{m}_2(x)$, and the marginal-integration estimate (the fourth column), $\hat{m}_3(x)$. The latter three estimates are based on random pooled data in the upper panels, and are based on homogeneous pooled data in the lower panels. Within each panel, the dot-dashed curve is the true function $m(x)$, the blue curve is the pointwise mean curve based on the 500 function estimates, the grey band is constructed by the mean curve plus and minus 1.96 times the pointwise standard deviation curve, and the dotted lines provides a comparison of the pointwise mean squared error curve across different estimates plotted with respect to the right axis of each subfigure.

and 5), the average-weighted estimator $\hat{m}_1(x)$ is unable to capture the shape of $m(x)$, and it fails more miserably around regions with more curvature. The product-weighted estimator $\hat{m}_2(x)$ is able to recover the overall shape of $m(x)$, although more variable than $\hat{m}_0(x)$, especially around the inflection points of $m(x)$. With $c = 2$ (as in Figure 4), the marginal-integration estimator $\hat{m}_3(x)$ performs similarly as $\hat{m}_2(x)$. When $c = 5$ (as in Figure 5), $\hat{m}_3(x)$ outperforms $\hat{m}_2(x)$ substantially in every regard. This is in line with the implication of Theorem 4.1 that the variance of $\hat{m}_2(x)$ inflates faster as the pool size increases than the variance of $\hat{m}_3(x)$ does. Under homogeneous pooling (see lower panels of Figures 4 and 5), the marginal-integration estimator $\hat{m}_3(x)$ distorts the functional form of $m(x)$, whereas both $\hat{m}_1(x)$ and $\hat{m}_2(x)$ perform similarly as $\hat{m}_0(x)$, in regard to both accuracy and precision.

7. Real-life applications

In this section, we analyse data from two real-life applications to illustrate the proposed local linear estimators for a conditional mean function. The individual-level observations are available in both applications, making it feasible to compute the local linear estimate based on individual-level data, $\hat{m}_0(x)$, which we compare our proposed estimates based on pooled data with. In all considered estimators, we set $K(t)$ as the Epanechnikov kernel.

Example 7.1 (Perfluorinated chemicals): The first data set is from the National Health and Nutrition Examination Survey, relating to a study of the bioaccumulation of perfluorinated chemicals (PFCs) in human bodies. PFCs are widely used in the coating of industrial products, such as food packaging foams and non-stick cookware surfaces, many of which are toxic and accumulate in human bodies. Kärrman et al. (2006) studied the relationship between the concentration levels of PFCs in an individual's blood and one's age, gender, and geographic region using pooled serum samples of individuals in Australia. The particular data we entertain here include concentration levels of multiple PFCs in the serum samples of 1904 residents in the United States between 2011 and 2012, along with their demographic information. The goal of our analysis is to infer the relationship between the concentration level of one particular type of PFCs, perfluorohexane sulfonic acid (PFHxS, Y), in an individual's blood and his/her age (X).

To assess the uncertainty of each estimation procedure, we generate 500 bootstrap samples from the raw individual-level data. Based on each bootstrap version of the individual-level data, we compute the local linear estimate, $\hat{m}_0(x)$, for the mean concentration level of PFHxS given one's age. Additionally, using the original data, we randomly create 952 pools, each of size two, producing a set of random pooled data; and we also create 952 pools of equal size based on the sorted data for age, producing a set of homogeneous pooled data. With the pool composition under each pooling design fixed, 500 bootstrap versions of random pooled data, and 500 bootstrap versions of homogeneous pooled data are generated by resampling pools with replacement. Using each pooled data set, we compute $\hat{m}_1(x)$, $\hat{m}_2(x)$, and $\hat{m}_3(x)$, resulting in 500 realisations of each estimator.

Figure 6 depicts the average of each estimate across 500 bootstrap samples and two quantiles of selected estimates. When random pooled data are used, the marginal-integration estimate $\hat{m}_3(x)$ matches closely with the benchmark estimate based on individual-level data, $\hat{m}_0(x)$, both indicating a relatively stable level of PFHxS with a slight decrease as one approaches age 40, and then a steep increase of the concentration level once one passes around age 50. This pattern can be explained by the fact that PFHxS can be partly eliminated from the human body via, for instance, gastrointestinal activities, menstrual bleeding, and breast feeding (Genuis, Curtis, and Birkholz 2013), but many of these pathways of PFCs elimination become less proactive or are completely lost (such as due to menopause) after one reaches certain age. In contrast, the average-weighted estimate, $\hat{m}_1(x)$, and the product-weighted estimate, $\hat{m}_2(x)$, suggest a much slower and nearly a constant increase in the concentration level as one gets older across the entire observed age

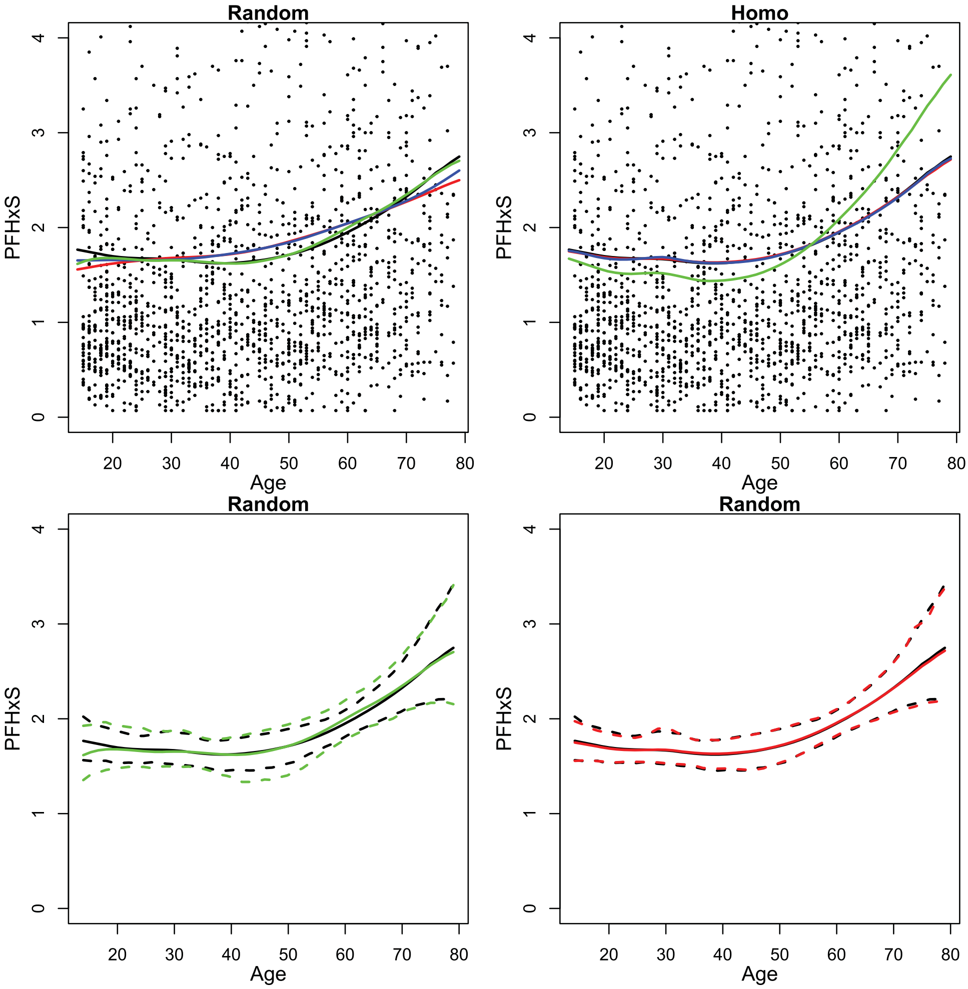


Figure 6. Results from Example 7.1 (Perfluorinated chemicals). Top panels depict the average of each considered estimate across 500 bootstraps. The black dots are individual observations, with observations far larger than 4 omitted. Within each panel, the solid black line corresponds to the local linear estimate based on individual-level data, $\hat{m}_0(x)$; the solid red, blue, and green lines correspond to the average-weight estimate $\hat{m}_1(x)$, the product-weighted estimate $\hat{m}_2(x)$, and the marginal-integration estimate $\hat{m}_3(x)$, respectively. Bottom panels show two quantiles of the estimates across 500 bootstraps. The dashed black, red, and green lines are 5% and 95% quantiles of $\hat{m}_0(x)$, $\hat{m}_1(x)$, and $\hat{m}_3(x)$, respectively.

range. We believe that this is one case where $\hat{m}_1(x)$ fails to capture the underlying pattern of $m(x)$ due to its inherent inconsistency in estimation, and $\hat{m}_2(x)$ also misses this pattern due to its high uncertainty in estimation. In conclusion, when only random pooled data are available, $\hat{m}_3(x)$ provides a more reliable estimate for the underlying relationship between one's PFHxS level in blood and age than the other two proposed estimates, although its variability is slightly higher than that of $\hat{m}_0(x)$ according to the bootstrap quantiles of the two estimates.

When homogeneous pooled data are used (see the top-right panel of Figure 6), $\hat{m}_3(x)$ appears to exaggerate the curvature of the conditional mean function, resulting in a much faster increase in the concentration level once one passes age 50, compared to the rate of increase indicated by the same estimate under random pooling. Despite the use of pooled data, $\hat{m}_1(x)$ and $\hat{m}_2(x)$ are nearly indistinguishable from $\hat{m}_0(x)$, and these three estimates mostly preserve the earlier estimated pattern of $m(x)$ that can be justified on scientific grounds. Moreover, the variability of $\hat{m}_1(x)$ is comparable with that of $\hat{m}_0(x)$ according to the comparison of the bootstrap quantiles associated with these two estimates. In conclusion, the marginal-integration estimate $\hat{m}_3(x)$ based on homogeneous pooled data leads to misleading inference for the underlying truth, whereas the other two estimates based on pooled data provide inference similar to those from the estimate based on individual-level data without noticeable efficiency loss.

Example 7.2 (Chemokines): The second data set we use to illustrate local linear estimation using different types of data is from the Collaborative Perinatal Project (CPP), which is a long-standing, collaborative project on maternal and child health in the United States. More specifically, this data include chemokine levels collected from 388 pregnant females recruited in CPP, with measurements taken at the individual level as well as the pool level, with 194 non-overlapping pools of size two randomly formed. Chemokines are a family of small proteins related to the homeostatic and inflammatory process in the human body. Medical researchers have studied extensively the role that chemokines play in the immune system. For example, regarding to two particular chemokines, MCP-3 and GRO-α, Dhawan and Richmond (2002) investigated the role of the former in tumorigenesis, and Tsou et al. (2007) studied the latter in monocyte mobilisation.

Based on the observed individual-level data and the random pooled data available in CPP, we infer the conditional mean concentration of GRO-α (Y) given MCP-3 (X). For illustration purpose, we generate another pooled data set, with a common pool size of two, following the homogeneous pooling design based on sorted MCP-3 levels. To assess the uncertainty of each estimation method, we generate 500 bootstrap samples for each of the three data types, individual-level data, random pooled data, and homogeneous pooled data, following the same resampling process described in the first example. Figure 7 shows the average of each considered estimate across 500 bootstrap samples and two quantiles of selected estimates.

Similar to the phenomena in the first example, the marginal-integration estimate $\hat{m}_3(x)$ yields a similar estimate for the mean concentration level of GRO-α given the level of MCP-3 as that of $\hat{m}_0(x)$ when random pooled data are used; but it grossly deviates from this benchmark estimate when homogeneous pooled data are used. In contrast, the other two proposed local linear estimates based on random pooled data go through an obviously

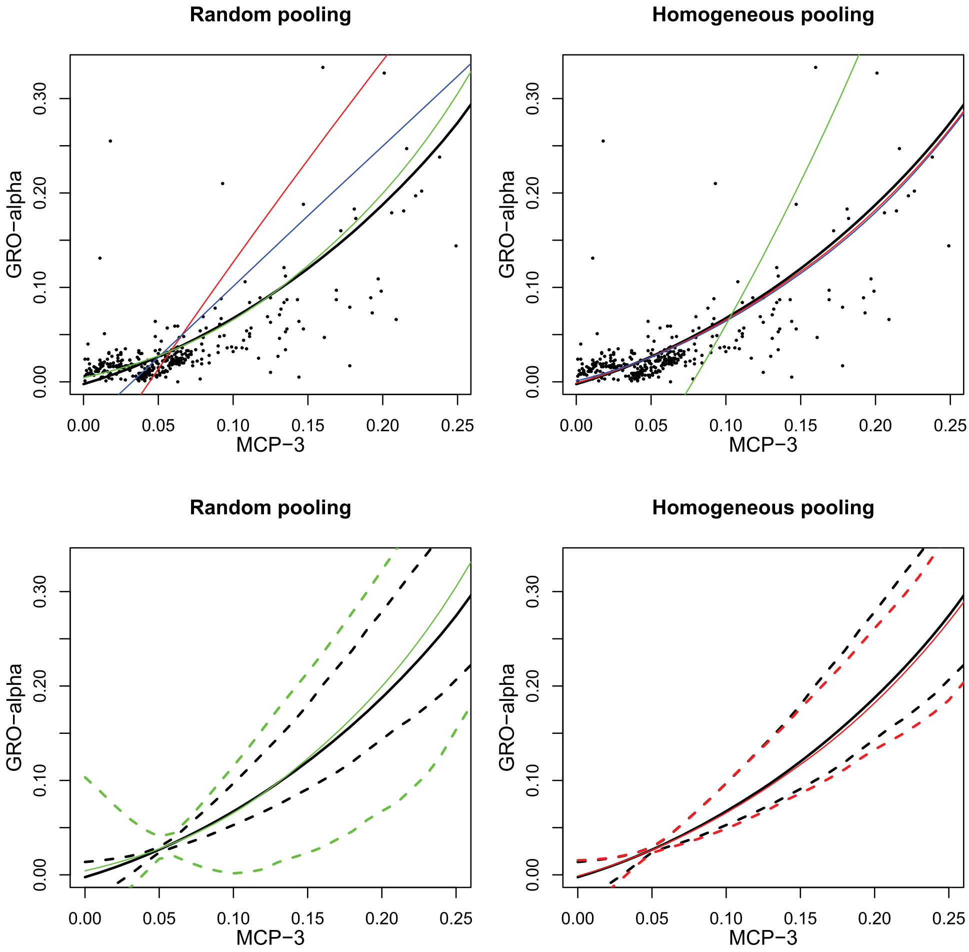


Figure 7. Results from Example 7.2 (Chemokines). Top panels depict the average of each considered estimate across 500 bootstraps. The black dots are individual observations. Within each panel, the solid black line corresponds to the local linear estimate based on individual-level data, $\hat{m}_0(x)$; the solid red, blue, and green lines correspond to the average-weight estimate $\hat{m}_1(x)$, the product-weighted estimate $\hat{m}_2(x)$, and the marginal-integration estimate $\hat{m}_3(x)$, respectively. Bottom panels show two quantiles of the estimates across 500 bootstraps. The dashed black, red, and green lines are 5% and 95% quantiles of $\hat{m}_0(x)$, $\hat{m}_1(x)$, and $\hat{m}_3(x)$, respectively.

uninteresting region of the observed data, yet both estimates applied to homogeneous pooled data follow closely the benchmark estimate $\hat{m}_0(x)$, and they only show slight discrepancy from it around the region where data are relatively scarce.

8. Discussion

We present in this article methods for estimating the mean of a continuous response given covariates via local polynomial regression when only pooled response data are observed along with individual-level covariates. Two commonly adopted pooling designs in practice are considered when formulating the local polynomial estimators, and properties of the proposed estimators are compared under each of the pooling designs. We use two real-life applications to illustrate the implementation and performance of the proposed estimators in comparison with their counterpart estimator when individual response data are available. Findings from the two applications are in line with observations on their finite sample performance using synthetic data from the simulation study, which agree with the theoretical implications of the large-sample properties derived for the proposed estimators.

In summary, the marginal-integration estimator $\hat{m}_3(x)$ is the winner among the three proposed when pooled data are from a random pooling design, but it fails when pools are not formed randomly; the average-weighted estimator $\hat{m}_1(x)$ performs the best when homogeneous pooled data are used, but it is an inconsistent estimator for the mean function when pools are formed randomly; the product-weighted estimator $\hat{m}_2(x)$ is a consistent estimator under both pooling designs but is subject to high variability under random pooling. Between the two winners, i.e. $\hat{m}_3(x)$ under random pooling and $\hat{m}_1(x)$ under homogeneous pooling, they share the same bias asymptotically. A closer look at their asymptotic variances (see Appendix C in the supplementary materials) reveals that the asymptotic variance of the former is $1 + (c-1)\bar{\sigma}^2/\sigma^2(x)$ times that of the latter, indicating that the former tends to be more variable than the latter given a fixed N and c. These patterns of comparisons between the two winning estimators (based on different pooling designs) are also observed in the numerical examples (e.g. Figures 4– 7). Based on our discussions in Section 4.2, we believe that there is still room for improvement by more carefully/selectively incorporating individual covariate information within a pool to relate to the pooled response of that pool, as opposed to either using all covariate information (as in $\hat{m}_1(x)$ and $\hat{m}_2(x)$) or using one individual's covariate information (as in $\hat{m}_3(x)$). Following this more selective incorporation of covariate information for each pool, an alternative construction of the weight function in the objective function may be needed accordingly to exploit a more sensible measure of distance between selected individuals' covariate information and x, the value at which the mean function is of interest. We are hopeful that this more refined strategy for constructing the objective function can lead to a local polynomial estimator that outperforms all three estimators proposed in the current study despite the pooling design.

Another follow-up research is motivated by the fact that, in many applications, covariates of interest cannot be measured precisely or observed directly. It is of interest then to carry out local polynomial regression to infer $m(x)$ using pooled response data and individual-level covariate data that are prone to measurement error.

Acknowledgements

The authors are grateful to the Associate Editor and the two anonymous referees for their insightful comments and helpful suggestions that greatly improved our work. Wang's work was supported by the National Institutes of Health (NIH) under the Award Number R03 AI135614.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by National Institutes of Health [R03 AI135614].

ORCID

Dewei Wang http://orcid.org/0000-0003-0822-8563

References

Bilder, C.R., and Tebbs, J.M. (2009), 'Bias, Efficiency, and Agreement for Group-Testing Regression Models', *Journal of Statistical Computation and Simulation*, 79(1), 67–80.

Deckert, A., Bärnighausen, T., and Kyei, N. (2020), 'Pooled-Sample Analysis Strategies for Covid-19 Mass Testing: A Simulation Study'. Bulletin of the World Health Organization. E-pub: 2 April 2020.

Delaigle, A., Fan, J., and Carroll, R. (2009), 'A Design-Adaptive Local Polynomial Estimator for the Errors-in-Variables Problem', *Journal of the American Statistical Association*, 104(485), 348–359.

Delaigle, A., and Hall, P. (2012), 'Nonparametric Regression with Homogeneous Group Testing Data', *The Annals of Statistics*, 40(1), 131–158.

Dhawan, P., and Richmond, A. (2002), 'Role of CXCL1 in Tumorigenesis of Melanoma', *Journal of Leukocyte Biology*, 72(1), 9–18.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66* (Vol. 66), New York: Chapman & Hall/CRC.

Fan, J., Heckman, N.E., and Wand, M.P. (1995), 'Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions', *Journal of the American Statistical Association*, 90(429), 141–150.

Fukuda, K. (2006), 'Age–Period–Cohort Decomposition of Aggregate Data: An Application to Us and Japanese Household Saving Rates', *Journal of Applied Econometrics*, 21(7), 981–998.

Genuis, S.J., Curtis, L., and Birkholz, D. (2013), 'Gastrointestinal Elimination of Perfluorinated Compounds Using Cholestyramine and Chlorella Pyrenoidosa', *ISRN Toxicology*, 2013, 657849.

Gu, J., Li, Q., and Yang, J.-C. (2015), 'Multivariate Local Polynomial Kernel Estimators: Leading Bias and Asymptotic Distribution', *Econometric Reviews*, 34(6-10), 979–1010.

Heffernan, A., English, K., Toms, L., Calafat, A., Valentin-Blasini, L., Hobson, P., Broomhall, S., Ware, R., Jagals, P., Sly, PD, and Mueller, JF (2016), 'Cross-Sectional Biomonitoring Study of Pesticide Exposures in Queensland, Australia, Using Pooled Urine Samples', *Environmental Science and Pollution Research*, 23(23), 23436–23448.

Jiang, R., Manchanda, P., and Rossi, P.E. (2009), 'Bayesian Analysis of Random Coefficient Logit Models Using Aggregate Data', *Journal of Econometrics*, 149(2), 136–148.

Kärrman, A., Mueller, J.F., Van Bavel, B., Harden, F., Toms, L.-M.L., and Lindström, G. (2006), 'Levels of 12 Perfluorinated Chemicals in Pooled Australian Serum, Collected 2002–2003, in Relation to Age, Gender, and Region', *Environmental Science & Technology*, 40(12), 3742–3748.

Kato, K., Calafat, A.M., Wong, L.-Y., Wanigatunga, A.A., Caudill, S.P., and Needham, L.L. (2009), 'Polyfluoroalkyl Compounds in Pooled Sera From Children Participating in the National Health and Nutrition Examination Survey 2001–2002', *Environmental Science & Technology*, 43(7), 2641–2647.

Kendziorski, C., Zhang, Y., Lan, H., and Attie, A. (2003), 'The Efficiency of Pooling MRNA in Microarray Experiments', *Biostatistics (Oxford, England)*, 4(3), 465–477.

Lin, J., and Wang, D. (2018), 'Single-Index Regression for Pooled Biomarker Data', *Journal of Nonparametric Statistics*, 30(4), 813–833.

Linton, O., and Whang, Y.-J. (2002), 'Nonparametric Estimation with Aggregated Data', *Econometric Theory*, 18(2), 420–468.

Liu, Y., McMahan, C., and Gallagher, C. (2017), 'A General Framework for the Regression Analysis of Pooled Biomarker Assessments', *Statistics in Medicine*, 36(15), 2363–2377.

Ma, C.-X., Vexler, A., Schisterman, E.F., and Tian, L. (2011), 'Cost-Efficient Designs Based on Linearly Associated Biomarkers', *Journal of Applied Statistics*, 38(12), 2739–2750.

Malinovsky, Y., Albert, P.S., and Schisterman, E.F. (2012), 'Pooling Designs for Outcomes Under a Gaussian Random Effects Model', *Biometrics*, 68(1), 45–52.

Martinez-Espineira, R. (2003), 'Estimating Water Demand Under Increasing-Block Tariffs Using Aggregate Data and Proportions of Users Per Block', *Environmental and Resource Economics*, 26(1), 5–23.

Masry, E. (1996), 'Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates', *Journal of Time Series Analysis*, 17(6), 571–599.

McMahan, C.S., McLain, A.C., Gallagher, C.M., and Schisterman, E.F. (2016), 'Estimating Covariate-Adjusted Measures of Diagnostic Accuracy Based on Pooled Biomarker Assessments', *Biometrical Journal*, 58(4), 944–961.

Mitchell, E.M., Lyles, R.H., Manatunga, A.K., Danaher, M., Perkins, N.J., and Schisterman, E.F. (2014), 'Regression for Skewed Biomarker Outcomes Subject to Pooling', *Biometrics*, 70(1), 202–211.

Mitchell, E.M., Lyles, R.H., Manatunga, A.K., and Schisterman, E.F. (2015), 'Semiparametric Regression Models for a Right-Skewed Outcome Subject to Pooling', *American Journal of Epidemiology*, 181(7), 541–548.

Mosites, E., Rodriguez, E., Caudill, S.P., Hennessy, T.W., and Berner, J. (2020), 'A Comparison of Individual-Level Vs. Hypothetically Pooled Mercury Biomonitoring Data From the Maternal Organics Monitoring Study (moms), Alaska, 1999–2012', *International Journal of Circumpolar Health*, 79(1), 1726256.

Shih, J.H., Michalowska, A.M., Dobbin, K., Ye, Y., Qiu, T.H., and Green, J.E. (2004), 'Effects of Pooling MRNA in Microarray Class Comparisons', *Bioinformatics (Oxford, England)*, 20(18), 3318–3325.

Shu, C., and Burn, D.H. (2003), 'Spatial Patterns of Homogeneous Pooling Groups for Flood Frequency Analysis', *Hydrological Sciences Journal*, 48(4), 601–618.

Tsou, C.-L., Peters, W., Si, Y., Slaymaker, S., Aslanian, A.M., Weisberg, S.P., Mack, M., and Charo, I.F. (2007), 'Critical Roles for Ccr2 and Mcp-3 in Monocyte Mobilization From Bone Marrow and Recruitment to Inflammatory Sites', *The Journal of Clinical Investigation*, 117(4), 902–909.

Vansteelandt, S., Goetghebeur, E., and Verstraeten, T. (2000), 'Regression Models for Disease Prevalence with Diagnostic Tests on Pools of Serum Samples', *Biometrics*, 56(4), 1126–1133.