Journal of Multivariate Analysis 102 (2011) 1126–1140

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Nonparametric additive model-assisted estimation for survey data

Li Wang [a,*], Suojin Wang [b]

[a] Department of Statistics, University of Georgia, Athens, GA 30602, United States

[b] Department of Statistics, Texas A&M University, College Station, TX 77843, United States

ARTICLE INFO

Article history:
Received 9 June 2010
Available online 28 March 2011

AMS 2000 subject classifications:
primary 62D05
secondary 62G08

Keywords:
Calibration
Horvitz–Thompson estimator
Local linear regression
Model-assisted estimation
Spline
Superpopulation

ABSTRACT

An additive model-assisted nonparametric method is investigated to estimate the finite population totals of massive survey data with the aid of auxiliary information. A class of estimators is proposed to improve the precision of the well known Horvitz–Thompson estimators by combining the spline and local polynomial smoothing methods. These estimators are calibrated, asymptotically design-unbiased, consistent, normal and robust in the sense of asymptotically attaining the Godambe-Joshi lower bound to the anticipated variance. A consistent model selection procedure is further developed to select the significant auxiliary variables. The proposed method is sufficiently fast to analyze large survey data of high dimension within seconds. The performance of the proposed method is assessed empirically via simulation studies.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

Auxiliary information is often available in many surveys for all elements of the population of interest. For instance, in many countries, administrative registers provide extensive sources of auxiliary information. Complete registers can give access to variables such as sex, age, income and country of birth. Studies of labor force characteristics or household expenditure patterns, for example, might benefit from these auxiliary data. Another example is the satellite images or GPS data used in spatial sampling. These data are often collected at the population level, which are often available at little or no extra cost, especially compared to the cost of collecting the survey data.

If no information other than the inclusion probabilities is used to estimate the population total, a well-known design unbiased estimator is the Horvitz–Thompson (HT) estimator. Nowadays, "cheap" auxiliary information can be regularly used to obtain higher precision estimates for the unknown finite population quantities. For instance, post-stratification, calibration and regression estimation are different design-based approaches used to improve the precision of estimators. Auxiliary information can also be used to increase the accuracy of the finite population distribution function; see, for example, [28]. Model-assisted estimation [22] provides a convenient way to incorporate auxiliary variables to develop more efficient survey estimators. By model-assisted, it is meant that a superpopulation model is adopted (for example, model (1) below), in which the finite population is modeled conditionally on the auxiliary information; see, for instance, [4,6,8,9].

The traditional parametric model-assisted approach assumes that the superpopulation model is fully described by a finite set of parameters, e.g., the regression estimator introduced in [22]. However, survey data now being collected by many government, health and social science organizations have more complex design features. It is difficult to obtain any prior model information to address various hypotheses. In this sense, a preselected parametric model is too restricted to fit

* Corresponding author.
E-mail address: lilywang@uga.edu (L. Wang).

0047-259X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2011.03.006

unexpected features. In contrast, nonparametric regression provides a useful tool for studying the dependence of variables of interest on auxiliary information without constraining the dependence to a fixed form with few parameters. The flexibility of nonparametric regression is extremely helpful to capture the complicated relationship between variables as well as in obtaining robust predictions; see [10,12] for details.

Breidt and Opsomer [2] first proposed a nonparametric model-assisted estimator based on local polynomial regression, which generalizes the parametric framework in survey sampling and improves the precision of the survey estimators immensely. Their investigation is only based on one auxiliary variable. Most surveys, however, involve more than one study variable, perhaps, many; see [20]. For example, the remote sensing data which provide a wide and growing range of variables to be employed. In this context, when the dimension of the auxiliary information vector is high, one unavoidable issue is the "curse of dimensionality", which refers to the poor convergence rate of nonparametric estimation of general multivariate functions. One solution is regression in the form of an additive model; see [13].

Estimation and inference for additive models have been well-studied in the literature; see, for example, the classic backfitting estimators of [13], the marginal integration estimators of [16], the smoothing backfitting estimators of [17], the spline estimators of Stone [25,26] and the spline-backfitted kernel estimators of [30]. In a survey sampling context, [1] discussed a semiparametric possible extension to multiple auxiliary variables via using the penalized splines; [19] applied the generalized additive models (GAMs) in an interaction model for the estimation of variables from forest inventory and analysis surveys; and [3] proposed a special case of the GAMs with an identity link function. For large and high dimensional survey data, it is important that estimation and inference methods are efficient and computationally easily implemented. However, few methods are theoretically justified and computational efficient when there are multiple nonparametric terms. The kernel based backfitting and marginal integration approaches are computationally expensive, limiting their use for high dimensional data; see [18] for some numerical comparisons of these methods. Spline methods, on the other hand, provide only convergence rates but no asymptotic distributions, so no measures of confidence can be assigned to the estimators.

Challenged by these demands, we propose approximating the nonparametric components by using the spline-backfitted local polynomial: spline does a quick initial estimation of all additive components and removes them all except the ones of interest; kernel smoothing is then applied to the cleaned univariate data to estimate with the asymptotic distribution. This two-step estimator is both computationally expedient for analyzing large and high dimensional survey data, and theoretically reliable as the estimator is uniformly oracle with asymptotic confidence intervals. The resulting estimator of population total can therefore be easily calculated, and more importantly allow for formal derivation of the asymptotic properties of the estimator.

In practice, a large number of variables may be collected and some of the insignificant ones should be excluded from the final model in order to enhance the predictability. The selection of auxiliary variables is a fundamental issue for model-assisted survey sampling methods. In this paper, we propose a consistent variable selection method for the additive model-assisted survey sampling based on the Bayes information criterion (BIC). A comprehensive Monte Carlo study demonstrates superior performance of the proposed methods.

The rest of the paper is organized as follows. Section 2 gives details of the superpopulation model and proposed method of estimation. Section 3 describes the weighting, calibration and asymptotic properties of the proposed estimator. Section 4 describes the auxiliary variable selection procedure for the superpopulation model under simple random sampling design (SRS). Section 5 reports the findings in an extensive simulation study. Lengthy technical arguments are given in the Appendix.

2. Superpopulation model and proposed estimator

In what follows, let $U_N = \{1, \ldots, i, \ldots, N\}$ be the finite population of N elements, called the target population, and i represents the ith element of the population. Let $\mathbf{x}_i = \{x_{i1}, \ldots, x_{\mathrm{id}}\}$ be a d-dimensional auxiliary variable vector, $i \in U_N$. We are interested in the estimation of the population total $t_y = \sum_{i \in U_N} y_i$, where y_i is the value of the study variable, y, for the ith element. To this end, a sample s of size n_N is drawn from U_N according to a fixed sampling design $p_N(\cdot)$, where $p_N(s)$ is the probability of drawing the sample s. The inclusion probabilities, known for all $i \in U_N$, are $\pi_{iN} \equiv \pi_i = \Pr\{i \in s\} = \sum_{s \ni i} p_N(s)$. In addition to the π_i, denote $\pi_{ijN} \equiv \pi_{ij} = \Pr\{i, j \in s\} = \sum_{s \ni i,j} p_N(s)$ the inclusion probability for both elements $i, j \in U_N$.

Let $\{(\mathbf{x}_i, y_i)\}_{i \in U_N}$ be a realization of $(\mathbf{X}, Y)$ from an infinite superpopulation, ξ, satisfying

$$Y = m(\mathbf{X}) + \sigma(\mathbf{X})\varepsilon, \tag{1}$$

in which the unknown d-variate function m has a simpler form of

$$m(\mathbf{X}) = c + \sum_{\alpha=1}^{d} m_\alpha(X_\alpha), \qquad E_\xi[m_\alpha(X_\alpha)] \equiv 0, \quad 1 \le \alpha \le d, \tag{2}$$

the function $\sigma(\cdot)$ is the unknown standard deviation function and the standard error ε satisfies that $E_\xi(\varepsilon|\mathbf{X}) = 0$ and $E_\xi(\varepsilon^2|\mathbf{X}) = 1$. In the following, we assume that the auxiliary variable X_α is distributed on a compact interval $[a_\alpha, b_\alpha]$, $\alpha = 1, \ldots, d$. Without loss of generality, we take all intervals $[a_\alpha, b_\alpha] = [0, 1]$. To estimate the additive components in (2), we employ a two-stage procedure based on the spline-backfitted local polynomial smoothing.

For any $\alpha = 1, \ldots, d$, we introduce a knot sequence with J interior knots $k_{0\alpha} = 0 < k_{1\alpha} < \cdots < k_{J\alpha} < 1 = k_{(J+1)\alpha}$, where $J \equiv J_N$ increases when n_N increases, and the precise order is given in assumption (A5). Denote the piecewise linear truncated power spline basis

$$\mathbf{\Gamma}(\mathbf{x}) \equiv \{1, x_\alpha, (x_\alpha - k_{1\alpha})_+, \ldots, (x_\alpha - k_{J\alpha})_+, \alpha = 1, \ldots, d\}^{\mathrm{T}}, \tag{3}$$

where $(a)_+ = a$ if $a > 0$ and 0 otherwise. For the local linear smoothing, let $K_h(x) = h^{-1}K(x/h)$, where K denotes a kernel function and $h = h_N$ is the bandwidth; see assumption (A6).

We now describe our two-stage estimator for the population total t_y. At the first stage, we apply the spline smoothing to obtain a quick initial estimator of $m(\mathbf{x}_i)$,

$$\hat{m}(\mathbf{x}_i) = \hat{b}_0 + \sum_{\alpha=1}^{d} \hat{b}_{0,\alpha} x_{i\alpha} + \sum_{\alpha=1}^{d} \sum_{j=1}^{J} \hat{b}_{j,\alpha} (x_{i\alpha} - k_{j\alpha})_+,$$

where $\hat{b}_0$ and $\hat{b}_{j,\alpha}, j = 0, 1, \ldots, J, \alpha = 1, \ldots, d$ are the minimizers of the following

$$\sum_{i\in s} \pi_i^{-1} \left\{ y_i - b_0 - \sum_{\alpha=1}^{d} b_{0,\alpha} x_\alpha - \sum_{\alpha=1}^{d} \sum_{j=1}^{J} b_{j,\alpha} (x_{i\alpha} - k_{j\alpha})_+ \right\}^2 \tag{4}$$

over a $G_d \equiv 1 + (J+1)d$ dimensional vector. Because the components $m_\alpha(x_\alpha)$ each can only be identified up to an additive constant, we center the estimator of $m_\alpha(x_\alpha)$ and define the centered pilot estimator of the αth component as

$$\hat{m}_\alpha(x_\alpha) = \hat{b}_{0,\alpha} x_\alpha + \sum_{j=1}^{J} \hat{b}_{j,\alpha} (x_\alpha - k_{j\alpha})_+ - \hat{c}_\alpha, \tag{5}$$

where $\hat{c}_\alpha = N^{-1} \sum_{i\in s} \pi_i^{-1} \left\{ \hat{b}_{0,\alpha} x_{i\alpha} + \sum_{j=1}^{J} \hat{b}_{j,\alpha} (x_{i\alpha} - k_{j\alpha})_+ \right\}$. The pilot estimators in (5) are then used to construct the new pseudo-responses

$$\hat{y}_{i\alpha} = y_i - N^{-1}\hat{t}_y - \sum_{\beta\neq\alpha} \hat{m}_\beta(x_{i\alpha}), \quad i \in s,\ \alpha = 1, \ldots, d, \tag{6}$$

where $\hat{t}_y$ is the well-known HT estimator.

At the second stage, a local polynomial smoothing is applied to the cleaned univariate data $\{x_{i\alpha}, \hat{y}_{i\alpha}\}_{i\in s}$ to achieve the "oracle" property in [30]. To be specific, considering the local linear smoothing, for any $\alpha = 1, \ldots, d$, we minimize

$$\sum_{i\in s} \pi_i^{-1} \{\hat{y}_{i\alpha} - a_{0,\alpha} - a_{1,\alpha}(x_{i\alpha} - x) K_h(x_{i\alpha} - x)\}^2, \tag{7}$$

with respect to $a_{0,\alpha}$ and $a_{1,\alpha}$. The spline-backfitted local linear (SBLL) estimator of the α-th component m_α is $\hat{m}^*_\alpha = \hat{a}_{0,\alpha}$ in (7). The final sample design-based SBLL estimator of $m(\mathbf{x})$ is defined as

$$\hat{m}^*(\mathbf{x}) = \frac{1}{N}\hat{t}_y + \sum_{\alpha=1}^{d} \hat{m}^*_\alpha(x_\alpha). \tag{8}$$

Substituting $\hat{m}^*_i \equiv \hat{m}^*(\mathbf{x}_i)$ into the existing generalized difference estimator (see page 221 of [22]), the SBLL estimator for t_y is defined by

$$\hat{t}_{y,\mathrm{SBLL}} = \sum_{i\in U_N} \hat{m}^*_i + \sum_{i\in s} \frac{y_i - \hat{m}^*_i}{\pi_i} = \sum_{i\in s} \frac{y_i}{\pi_i} + \sum_{i\in U_N} \left(1 - \frac{I_i}{\pi_i}\right) \hat{m}^*_i, \tag{9}$$

where $I_i = 1$ if $i \in s$ and $I_i = 0$ otherwise.

Remark 1. In the first step spline smoothing, the number of knots J_N can be determined by n_N and a tuning constant c:

$$J_N = \min([c n_N^{1/4} \log(n_N)] + 1, [(n_N/2 - 1)/d - 1]). \tag{10}$$

As discussed in [30], the choice of c makes little difference. In the second step local polynomial smoothing, one can use the quartic kernel and the rule-of-thumb bandwidth.

3. Properties of the estimator

3.1. Weighting and calibration

In the last decade, calibration estimation has developed into an important field of research in survey sampling. As discussed in [7,15], calibration is a highly desirable property for survey weights, which allows the survey practitioner to simply adjust the original design weights to incorporate the information of the auxiliary variables. Several national statistical agencies have developed software to compute calibrated weights based on auxiliary information available in population registers and other sources. The proposed SBLL estimator in this paper also shares this property in certain sense.

Let $\mathbf{y}_s$ be the column vector of the response values y_i for $i \in s$ and define the diagonal matrix of inverse inclusion probabilities $\mathbf{\Pi}_s = \text{diag}\{1/\pi_i\}_{i\in s}$. For $\mathbf{\Gamma}(\mathbf{x})$ in (3), denote $\mathbf{\Gamma}_s = \{\mathbf{\Gamma}(\mathbf{x}_i)^{\mathrm{T}}\}_{i\in s}$ the sample truncated power spline matrix. Let $\mathbf{B}_s$ be the collection of the estimated spline coefficient in (4). Then $\mathbf{B}_s = (\mathbf{\Gamma}_s^{\mathrm{T}}\mathbf{\Pi}_s\mathbf{\Gamma}_s)^{-1}\mathbf{\Gamma}_s^{\mathrm{T}}\mathbf{\Pi}_s\mathbf{y}_s$. Thus the pilot spline estimator of $m_\alpha(x_\alpha)$ in (5) can be written as

$$\hat{m}_\alpha(x_\alpha) = \{\mathbf{\Gamma}(\mathbf{x})^{\mathrm{T}}\mathbf{D}_\alpha\mathbf{B}_s - N^{-1}\mathbf{1}_n^{\mathrm{T}}\mathbf{\Pi}_s\mathbf{\Gamma}_s\mathbf{D}_\alpha\mathbf{B}_s\}\mathbf{y}_s, \tag{11}$$

where $\mathbf{1}_n$ is a vector of length n_N with all "1"s, and

$$\mathbf{D}_\alpha = \text{diag}\{0,\ldots,0,\underbrace{1,\ldots,1}_{\text{from } (J+1)(\alpha-1)+2 \text{ to } (J+1)\alpha+1},0,\ldots,0\} \tag{12}$$

is a $G_d \times G_d$ diagonal matrix. Denoting the spline smoothing matrix and its centered version by

$$\mathbf{\Psi}_{s\alpha} = \mathbf{\Gamma}_s\mathbf{D}_\alpha(\mathbf{\Gamma}_s^{\mathrm{T}}\mathbf{\Pi}_s\mathbf{\Gamma}_s)^{-1}\mathbf{\Gamma}_s^{\mathrm{T}}\mathbf{\Pi}_s, \qquad \mathbf{\Psi}^*_{s\alpha} = (\mathbf{I} - N^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\mathbf{\Pi}_s)\mathbf{\Psi}_{s\alpha},$$

we have $\hat{\mathbf{m}}_\alpha \equiv \{\hat{m}_\alpha(x_{i\alpha})\}_{i\in s} = \mathbf{\Psi}^*_{s\alpha}\mathbf{y}_s$, for $\alpha = 1,\ldots,d$. Further for $\hat{y}_{i\alpha}$ in (6), let $\hat{\mathbf{y}}_\alpha \equiv \{\hat{y}_{i\alpha}\}_{i\in s} = \mathbf{y}_s - \frac{1}{N}\hat{t}_y\mathbf{1}_n - \sum_{\beta\neq\alpha}\hat{\mathbf{m}}_\beta$, and define the matrices

$$\mathbf{X}_{si\alpha} = \{(1 \quad x_{k\alpha} - x_{i\alpha})\}_{k\in s}, \qquad \mathbf{W}_{si\alpha} = \text{diag}\left\{\frac{1}{\pi_k}K_h(x_{k\alpha} - x_{i\alpha})\right\}_{k\in s}.$$

Then the SBLL estimator of m_α at $x_{i\alpha}$ can be written as

$$\hat{m}^*_{i\alpha} \equiv \hat{m}^*_\alpha(x_{i\alpha}) = \mathbf{e}_1^{\mathrm{T}}(\mathbf{X}_{si\alpha}^{\mathrm{T}}\mathbf{W}_{si\alpha}\mathbf{X}_{si\alpha})^{-1}\mathbf{X}_{si\alpha}^{\mathrm{T}}\mathbf{W}_{si\alpha}\hat{\mathbf{y}}_\alpha, \tag{13}$$

where $\mathbf{e}_1 = (1,0)^{\mathrm{T}}$. Therefore, the SBLL estimator in (8) of $m(\mathbf{x})$ at $\mathbf{x}_i$ is

$$\begin{aligned}\hat{m}^*_i &= \frac{1}{N}\hat{t}_y + \sum_{\alpha=1}^{d}\mathbf{e}_1^{\mathrm{T}}(\mathbf{X}_{si\alpha}^{\mathrm{T}}\mathbf{W}_{si\alpha}\mathbf{X}_{si\alpha})^{-1}\mathbf{X}_{si\alpha}^{\mathrm{T}}\mathbf{W}_{si\alpha}\left(\mathbf{y}_s - \frac{\hat{t}_y}{N}\mathbf{1}_n - \sum_{\beta\neq\alpha}\mathbf{\Psi}^*_{s\beta}\mathbf{y}_s\right)\\ &\equiv \boldsymbol{\rho}_{si}^{\mathrm{T}}\mathbf{y}_s,\end{aligned}$$

where

$$\boldsymbol{\rho}_{si}^{\mathrm{T}} = \mathbf{e}_1^{\mathrm{T}}\left\{\sum_{\alpha=1}^{d}(\mathbf{X}_{si\alpha}^{\mathrm{T}}\mathbf{W}_{si\alpha}\mathbf{X}_{si\alpha})^{-1}\mathbf{X}_{si\alpha}^{\mathrm{T}}\mathbf{W}_{si\alpha}\left(\mathbf{I} + \frac{1-\mathrm{d}}{\mathrm{d}N}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\mathbf{\Pi}_s - \sum_{\beta\neq\alpha}\mathbf{\Psi}^*_{s\beta}\right)\right\}.$$

Similar to [21], we define the "g-weight"

$$g_{is} = 1 + \pi_i\sum_{j\in U_N}\left(1 - \frac{I_j}{\pi_j}\right)\boldsymbol{\rho}_{sj}^{T}\mathbf{a}_i, \tag{14}$$

where $\mathbf{a}_i$ is a n_N-dimensional vector with a "1" in the ith position and a "0" elsewhere. Thus the proposed estimator $\hat{t}_{y,\text{SBLL}}$ in (8) can be written as

$$\hat{t}_{y,\text{SBLL}} = \sum_{i\in s}\frac{y_i}{\pi_i} + \sum_{j\in U_N}\left(1 - \frac{I_j}{\pi_j}\right)\boldsymbol{\rho}_{sj}^{T}\mathbf{y}_s \equiv \sum_{i\in s}g_{is}y_i/\pi_i,$$

which is a linear combination of the sample y_i's with a sampling weight, π_i^{-1}, and the "g-weight". Because the weights are independent of y_i, they can be applied to any study variable of interest.

As we show below, the weight system gives our estimator of the known total $\sum_{i\in U_N}x_{i\alpha}$ to be itself.

Theorem 1. *For any $\alpha = 1,\ldots,d$ and the "g-weight" defined in (14),*

$$\hat{t}_{x_\alpha,\text{SBLL}} \equiv \sum_{i\in s}g_{is}x_{i\alpha}/\pi_i = \sum_{i\in U_N}x_{i\alpha}.$$

Proof. Let $\mathbf{x}_\alpha = \{x_{i\alpha}\}_{i\in s}$. We have

$$\hat{t}_{x_\alpha,\mathrm{SBLL}} = \sum_{i\in s}\pi_i^{-1}x_{i\alpha} + \sum_{j\in U_N}(1-I_j\pi_j^{-1})\mathbf{e}_1^T\left\{\sum_{\gamma=1}^{d}(\mathbf{X}_{sj\gamma}^T\mathbf{W}_{sj\gamma}\mathbf{X}_{sj\gamma})^{-1}\mathbf{X}_{sj\gamma}^T\mathbf{W}_{sj\gamma}\left(\mathbf{I}+\frac{1-\mathrm{d}}{\mathrm{d}N}\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{\Pi}_s-\sum_{\beta\neq\gamma}\boldsymbol{\Psi}_{s\beta}^*\right)\right\}\mathbf{x}_\alpha.$$

Observe that

$$(\boldsymbol{\Gamma}_s^T\boldsymbol{\Pi}_s\boldsymbol{\Gamma}_s)^{-1}\boldsymbol{\Gamma}_s^T\boldsymbol{\Pi}_s\mathbf{x}_\alpha = \boldsymbol{\tau}_\alpha, \qquad \boldsymbol{\Psi}_{s\beta}\mathbf{x}_\alpha = \boldsymbol{\Gamma}_s\mathbf{D}_\beta\boldsymbol{\tau}_\alpha = \begin{cases}\mathbf{x}_\alpha, & \text{for } \beta=\alpha\\ \mathbf{0}, & \text{for } \beta\neq\alpha\end{cases},$$

where $\boldsymbol{\tau}_\alpha$ is the vector of dimension G_d with a "1" in the $\{2+(J+1)(\alpha-1)\}$th position and "0" elsewhere. Then we have,

$$\left(\mathbf{I}+\frac{1-\mathrm{d}}{\mathrm{d}N}\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{\Pi}_s-\sum_{\beta\neq\gamma}\boldsymbol{\Psi}_{s\beta}^*\right)\mathbf{x}_\alpha = \begin{cases}\left(\mathbf{I}+\dfrac{1-\mathrm{d}}{\mathrm{d}N}\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{\Pi}_s\right)\mathbf{x}_\alpha, & \text{for } \gamma=\alpha\\ \dfrac{1}{\mathrm{d}N}\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{\Pi}_s\mathbf{x}_\alpha, & \text{for } \gamma\neq\alpha\end{cases}.$$

Note that for any $i\in U_N$,

$$\mathbf{e}_1^T(\mathbf{X}_{si\alpha}^T\mathbf{W}_{si\alpha}\mathbf{X}_{si\alpha})^{-1}\mathbf{X}_{si\alpha}^T\mathbf{W}_{si\alpha}\mathbf{x}_\alpha = x_{i\alpha}, \qquad \mathbf{e}_1^T(\mathbf{X}_{si\alpha}^T\mathbf{W}_{si\alpha}\mathbf{X}_{si\alpha})^{-1}\mathbf{X}_{si\alpha}^T\mathbf{W}_{si\alpha}\mathbf{1}_n = 1.$$

Thus

$$\begin{aligned}&\mathbf{e}_1^T\left\{\sum_{\gamma=1}^{d}(\mathbf{X}_{sj\gamma}^T\mathbf{W}_{sj\gamma}\mathbf{X}_{sj\gamma})^{-1}\mathbf{X}_{sj\gamma}^T\mathbf{W}_{sj\gamma}\left(\mathbf{I}+\frac{1-\mathrm{d}}{\mathrm{d}N}\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{\Pi}_s-\sum_{\beta\neq\gamma}\boldsymbol{\Psi}_{s\beta}^*\right)\right\}\mathbf{x}_\alpha\\ &= \mathbf{e}_1^T(\mathbf{X}_{si\alpha}^T\mathbf{W}_{si\alpha}\mathbf{X}_{si\alpha})^{-1}\mathbf{X}_{si\alpha}^T\mathbf{W}_{si\alpha}\{\mathbf{I}+(\mathrm{d}N)^{-1}(1-d)\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{\Pi}_s\}\mathbf{x}_\alpha\\ &\quad + d^{-1}\mathbf{e}_1^T\sum_{\gamma\neq\alpha}(\mathbf{X}_{sj\gamma}^T\mathbf{W}_{sj\gamma}\mathbf{X}_{sj\gamma})^{-1}\mathbf{X}_{sj\gamma}^T\mathbf{W}_{sj\gamma}\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{\Pi}_s\mathbf{x}_\alpha = x_{j\alpha}.\end{aligned}$$

Hence the proposed SBLL estimator defined in (9) preserves the calibration property. □

3.2. Assumptions

For the asymptotic properties of the estimators, we adopt the traditional asymptotic framework in [2] where both the population and sample sizes increase as $N\to\infty$. There are two sources of "variation" to be considered here. The first is introduced by the random sample design and the corresponding measure is denoted by p. The "O_p", "o_p" and "$E_p(\cdot)$" notation below is with respect to this measure. The second is associated with the superpopulation from which the finite population is viewed as a sample. The corresponding measure and notation are "ξ". For simplicity, let $\pi_{ij}-\pi_i\pi_j=\Delta_{ij}$.

(A1) *The density $f(\mathbf{x})$ of $\mathbf{X}$ is continuous and bounded away from 0 and ∞. The marginal densities $f_\alpha(x_\alpha)$ of x_α have continuous derivatives and are bounded away from 0 and ∞.*

(A2) *The second order derivative of $m_\alpha(x_\alpha)$ is continuous, $\forall\alpha=1,\ldots,d$.*

(A3) *There exists a positive constant M such that $E_\xi(|\varepsilon|^{2+\delta}|\mathbf{X})<M$ for some $\delta>1/2$; $\sigma(\mathbf{x})$ is continuous on $[0,1]^d$ and bounded away from 0 and ∞.*

(A4) *As $N\to\infty$, $n_N\to\infty$ and $n_NN^{-1}\to\pi<1$.*

(A5) *The number of knots $J_N\sim n_N^{1/4}\log(n_N)$.*

(A6) *The kernel function K is Lipschitz continuous, bounded, nonnegative, symmetric, and supported on $[-1,1]$. The bandwidth $h_N\sim n_N^{-1/5}$, i.e., $c_hn_N^{-1/5}\le h_N\le C_hn_N^{-1/5}$ for some positive constants c_h, C_h.*

(A7) *For all N, $\min_{i\in U_N}\pi_i\ge\lambda>0$, $\min_{i,j\in U_N}\pi_{ij}\ge\lambda^*>0$ and*

$$\limsup_{N\to\infty} n_N\max_{i,j\in U_N,i\neq j}|\Delta_{ij}|<\infty.$$

(A8) *Let $D_{k,N}$ be the set of all distinct k-tuples $(i_1,i_2,\ldots,i_k)$ from U_N.* Then

$$\limsup_{N\to\infty} n_N^2\max_{(i_1,i_2,i_3,i_4)\in D_{4,N}}|E_p[(I_{i_1}-\pi_{i_1})(I_{i_2}-\pi_{i_2})(I_{i_3}-\pi_{i_3})(I_{i_4}-\pi_{i_4})]|<\infty,$$

$$\limsup_{N\to\infty} n_N^2\max_{(i_1,i_2,i_3,i_4)\in D_{4,N}}|E_p[(I_{i_1}I_{i_2}-\pi_{i_1i_2})(I_{i_3}I_{i_4}-\pi_{i_3i_4})]|<\infty,$$

$$\limsup_{N\to\infty} n_N^2\max_{(i_1,i_2,i_3)\in D_{3,N}}|E_p[(I_{i_1}-\pi_{i_1})^2(I_{i_2}-\pi_{i_2})(I_{i_3}-\pi_{i_3})]|<\infty.$$

Remark 2. Assumptions (A1)–(A3) are typical in the smoothing literature; see, for instance, [10,12,30]. Assumption (A5) is about how to choose the number of interior knots J_N for the spline estimation in the first stage. In practice, J_N can be determined by (10). Assumption (A6) is how to select the kernel function and the corresponding bandwidth. Such assumptions were used in [30] in the additive autoregressive model fitting. Assumptions (A7) and (A8) involve the inclusion probabilities of the design, which were also assumed in [2].

3.3. Asymptotic properties of the estimator

Like the local polynomial estimators in [2], the following theorem shows that the estimator $\hat{t}_{y,\mathrm{SBLL}}$ in (9) is asymptotically design unbiased and design consistent.

Theorem 2. *Under assumptions* (A1)–(A7), *the estimator* $\hat{t}_{y,\mathrm{SBLL}}$ *in* (9) *is asymptotically design unbiased in the sense that*

$$\lim_{N\to\infty} E_p\left[\frac{\hat{t}_{y,\mathrm{SBLL}} - t_y}{N}\right] = 0 \quad \textit{with } \xi\textit{-probability } 1,$$

and is design consistent in the sense that for all $\eta > 0$,

$$\lim_{N\to\infty} E_p[I_{\{|\hat{t}_{y,\mathrm{SBLL}} - t_y| > N\eta\}}] = 0 \quad \textit{with } \xi\textit{-probability } 1.$$

Let $\tilde{t}_{y,\mathrm{SBLL}}$ be the population-based generalized difference estimator of t_y when the entire realization were known; see (A.4) in Appendix A.1 for the formal definition. Like the estimators in the local polynomial estimators in [2], the penalized spline estimators in [1], and the backfitting estimators in [3], the following theorem shows that the proposed estimator $\hat{t}_{y,\mathrm{SBLL}}$ also inherits the limiting distribution of the "oracle" estimator $\tilde{t}_{y,\mathrm{SBLL}}$.

Theorem 3. *Under assumptions* (A1)–(A8),

$$\frac{N^{-1}(\tilde{t}_{y,\mathrm{SBLL}} - t_y)}{\mathrm{Var}_p^{1/2}(N^{-1}\tilde{t}_{y,\mathrm{SBLL}})} \xrightarrow{d} N(0,1)$$

as $N \to \infty$ *implies*

$$\frac{N^{-1}(\hat{t}_{y,\mathrm{SBLL}} - t_y)}{\widehat{V}^{1/2}(N^{-1}\hat{t}_{y,\mathrm{SBLL}})} \xrightarrow{d} N(0,1),$$

where

$$\widehat{V}(N^{-1}\hat{t}_{y,\mathrm{SBLL}}) = \frac{1}{N^2}\sum_{i,j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\frac{y_i - \hat{m}_i^*}{\pi_i}\frac{y_j - \hat{m}_j^*}{\pi_j}. \tag{15}$$

The next theorem proves that $\hat{t}_{y,\mathrm{SBLL}}$ is robust as in [2] and it also asymptotically attains the Godambe–Joshi lower bound to the anticipated variance

$$\mathrm{Var}[N^{-1}(\hat{t}_y - t_y)] = E[N^{-1}(\hat{t}_y - t_y)]^2 - E^2[N^{-1}(\hat{t}_y - t_y)],$$

where the expectation is taken over both design, p_N, and population ξ in (1).

Theorem 4. *Under assumptions* (A1)–(A8), $\hat{t}_{y,\mathrm{SBLL}}$ *asymptotically attains the Godambe–Joshi lower bound, in the sense that*

$$n_N E\left(\frac{\hat{t}_{y,\mathrm{SBLL}} - t_y}{N}\right)^2 = \frac{n_N}{N^2}\sum_{i\in U_N}\sigma^2(\mathbf{x}_i)\frac{1-\pi_i}{\pi_i} + o(1).$$

The proofs of Theorems 2–4 are given in the Appendix.

4. Auxiliary variable selection

In this section, we propose a BIC-based method to select the auxiliary variables for use in the superpopulation model (2).

The BIC was first proposed in [23] for the selection of parametric models. Recently, [14] proposed a fast and consistent model selection method based on spline estimation with the BIC to select significant lags in non-linear additive autoregression. Analogous to the approach in [14], if the entire realization were known by an "oracle", one can select significant auxiliary variables based on the BIC. For an index set of variables $r \subset \{1,\ldots,d\}$, the BIC is defined as

$$\mathrm{BIC}^{(r)} = \log\{\mathrm{AMSE}^{(r)}(N^{-1}\hat{t}_{y,\mathrm{SBLL}})\} + \frac{\mathcal{J}_r}{n_N}\log(n_N), \tag{16}$$

where $\mathcal{J}_r = 1 + \sum_{\alpha\in r}(J_N + 1)$, and $\mathrm{AMSE}(N^{-1}\hat{t}_{y,\mathrm{SBLL}})$ is the asymptotic mean squared error (AMSE) of $N^{-1}\hat{t}_{y,\mathrm{SBLL}}$ in (A.11), i.e. the asymptotic expectation of $\{N^{-1}(\hat{t}_{y,\mathrm{SBLL}} - t_y)\}^2$.

Next let $f = n_N/N$ be the fixed sampling fraction. Under a simple random sampling (SRS) design, if $\sigma^2(\mathbf{x}) = \mathbf{c}^{\mathrm{T}}\mathbf{x}$,

$$\mathrm{AMSE}(N^{-1}\hat{t}_{y,\mathrm{SBLL}}) = \frac{1-f}{n_N(N-1)}\sum_{i\in U_N}(y_i - \tilde{m}_i^*)^2.$$

Thus, using similar arguments in Section 5 of [14], we can show that the above BIC in (16) is consistent under SRS.

By Theorem A.2, $\mathrm{AMSE}(N^{-1}\hat{t}_{y,\mathrm{SBLL}})$ can be estimated consistently by

$$\widehat{V}_g \equiv \widehat{V}_g(N^{-1}\hat{t}_{y,\mathrm{SBLL}}) = \frac{1}{N^2}\sum_{i,j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\frac{g_{is}(y_i - \hat{m}_i^*)}{\pi_i}\frac{g_{js}(y_j - \hat{m}_j^*)}{\pi_j}, \tag{17}$$

a modified version of (15) proposed by [21] with the "g-weight" in (14). So the sample-based BIC is defined as

$$\mathrm{BIC}^{(r)} = \log\{\widehat{V}_g^{(r)}\} + \frac{\mathcal{J}_r}{n_N}\log(n_N), \tag{18}$$

and we select the subsect $\hat{r} \subset \{1, \ldots, d\}$ that gives the smallest BIC value.

Remark 3. Under SRS design, the variance estimator given in (17) can be simplified as

$$\widehat{V}_g = \frac{1-f}{n_N(n_N-1)}\sum_{i\in s} g_{is}^2(y_i - \hat{m}_i^*)^2.$$

In practice, we first decide on a set of candidate variables to be selected. Since a full search through all possible subsets of variables is in general computationally too costly in actual implementation of the BIC method, we consider a *forward selection* procedure and a *backward selection* procedure. Let d denote the total number of candidate variables to be selected from. In the *forward selection* procedure, we pre-specify the maximal number of variables $d_{\max} = \min\left\{d, \left[\frac{n_N}{2(J_N+1)}\right]\right\}$ that are allowed in the model, in which $[a]$ denotes the integer part of a. We start from the empty set of auxiliary variables, add one variable at a time to the current model, choosing between the various candidate variables that have not yet been selected by minimizing the BIC in (18). The process stops when the number of variables selected reaches $d_{\max}$. In the *backward selection* procedure, we start with a set of variables of the maximal size $d_{\max}$, delete one variable at a time by minimizing the BIC and stop when no variable remains in the model. If $d_{\max} < d$, we first apply the *forward selection* procedure, then we start with the maximal set of variables selected in the last step of the forward stage.

5. Simulation study

In this section, simulations are carried out to investigate the finite-sample performance of $\hat{t}_{y,\mathrm{SBLL}}$. For comparison we also obtained the results of four other estimators: the HT estimator which does not make use of the auxiliary population, the linear regression (LREG) estimator in [22], the one-step linear spline (LS) estimator defined by

$$\hat{t}_{y,\mathrm{LS}} = \sum_{i\in s}(y_i - \hat{m}_i)/\pi_i + \sum_{i\in U_N}\hat{m}_i, \qquad \hat{m}_i = N^{-1}\hat{t}_y + \sum_{\alpha=1}^{d}\hat{m}_{i\alpha}$$

with $\hat{m}_{i\alpha} \equiv \hat{m}_\alpha(x_{i\alpha})$ given in (5), and the single-index model-assisted (SIM) estimator in [27]. The number of knots J_N for the LS and SBLL is determined by (10).

For the superpopulation model (1), the following four additive models (no interactions) were considered:

$$\begin{aligned}
&\text{2-dim linear:} Y = -1 + 2X_3 + 4X_6 + \sigma_0\varepsilon,\\
&\text{2-dim quadratic:} Y = 5.5 - 6X_2 + 8(X_2 - 0.5)^2 - 3X_{10} + 32(X_{10} - 0.5)^3 + \sigma_0\varepsilon,\\
&\text{3-dim mixed:} Y = 8(X_2 - 0.5)^2 + \exp(2X_5 - 1) + \sin\{2\pi(X_8 - 0.5)\} + \sigma_0\varepsilon,\\
&\text{5-dim sinusoid:} Y = 2 + \sum_{\alpha=1}^{d}\sin\{2\pi(X_\alpha - 0.5)\} + \frac{\sigma_0}{2}\left(\sum_{\alpha=1}^{d}X_\alpha\right)^{1/2}\varepsilon, \quad d = 5.
\end{aligned}$$

The auxiliary variable vectors $\mathbf{x}_i$, $i \in U_N$, were generated from i.i.d. uniform (0, 1) random vectors. The errors ε were generated from i.i.d. $N(0, 1)$ with noise level $\sigma_0 = 0.1, 0.4$. The population size was $N = 1000$. SRS Samples were generated of size $n_N = 50$, 100 and 200. For each combination of noise level and sample size, 1000 replicated SRS samples were selected from the same population, the estimators were calculated, and the design bias and the design mean squared errors were computed empirically.

Table 1 shows the ratios of the mean squared error (MSE) for the various estimators to the proposed SBLL estimators. From the table, one sees that the model-assisted estimators, LREG, LS, SIM and SBLL, perform much better than the simple HT regardless the type of mean function, standard error and sample size. For Model 1, LREG is expected to be the preferred

Table 1
Ratio of MSE of the HT, LREG, LS and SIM estimators to that of the SBLL estimator and the average computing time of the SBLL estimator based on 1000 replications of SRS samples from four fixed populations of size $N = 1000$.

Model	Error	Sample size	MSE Ratio				SBLL (s)
	σ	n_N	HT	LREG	LS	SIM	
1	0.1	50	140.36	0.89	1.12	1.60	0.07
		100	148.03	0.91	1.07	1.33	0.07
		200	147.03	0.92	1.10	1.02	0.09
	0.4	50	9.78	0.92	1.16	1.24	0.07
		100	10.50	0.95	1.10	1.02	0.07
		200	10.47	0.98	1.05	1.04	0.09
2	0.1	50	134.05	28.38	2.11	19.77	0.07
		100	282.47	58.10	1.03	36.58	0.07
		200	313.93	66.63	0.98	41.15	0.09
	0.4	50	18.45	4.25	2.36	3.44	0.07
		100	23.67	5.34	1.04	3.69	0.07
		200	23.36	5.63	1.02	3.92	0.09
3	0.1	50	63.14	30.83	1.10	37.12	0.07
		100	103.33	49.62	1.01	50.76	0.07
		200	115.13	56.57	1.02	57.04	0.09
	0.4	50	6.80	3.46	1.11	3.93	0.07
		100	8.18	4.20	1.14	4.40	0.07
		200	18.39	4.52	1.09	4.57	0.09
4	0.1	50	55.81	25.26	1.01	27.61	0.07
		100	151.59	62.63	1.03	65.78	0.07
		200	230.44	97.91	0.97	99.45	0.09
	0.4	50	9.97	4.75	1.03	5.22	0.07
		100	16.35	7.10	1.01	7.44	0.07
		200	19.95	8.60	1.05	8.74	0.09

Table 2
Monte Carlo bias, standard error and the square root of the average estimated variances (15) of the population total based on 1000 simulations.

Model	σ	n	Bias	SE	Est. SE
1	0.1	50	−0.10	14.69	13.18
		100	−0.36	9.85	9.32
		200	−0.13	6.55	6.29
	0.4	50	−1.62	57.73	51.81
		100	−1.55	38.51	36.77
		200	−0.42	25.71	24.86
2	0.1	50	1.27	24.49	14.06
		100	0.62	11.52	9.13
		200	0.37	7.06	6.10
	0.4	50	2.41	67.66	52.45
		100	−0.47	40.94	36.15
		200	−0.13	26.54	24.33
3	0.1	50	2.29	20.40	13.38
		100	0.90	10.91	8.74
		200	0.48	6.82	5.88
	0.4	50	2.17	64.89	50.86
		100	−0.04	40.17	35.44
		200	0.32	26.30	23.99
4	0.1	50	−1.98	29.04	18.04
		100	−0.51	12.28	8.22
		200	−0.10	6.38	4.82
	0.4	50	−4.38	69.69	43.31
		100	−1.18	37.92	27.72
		200	−0.37	22.58	18.56

estimator, since the assumed model is correctly specified. However, not much efficiency is lost by using SBLL instead of LREG and the MSE ratios of LREG to SBLL are at least 0.89 for all cases. For all other scenarios, SBLL performs consistently better than LREG. The SBLL estimators also improve upon the LS estimators across almost every combination of noise level and sample size, which implies that our second local linear smoothing step is not redundant.

To see how fast the computation is, Table 1 also provides the average time of generating one sample of size n_N and obtaining the SBLL estimator on an ordinary PC with Intel Pentium IV 1.86 GHz processor and 1.0 GB RAM. It shows that the proposed SBLL estimation is extremely fast. For instance, for Model 4, the SBLL estimation of a 5-dimensional of size 200 takes on average merely 0.2 s. We also carried out simulations for high dimensional data with sample size $n_N = 1000$

Table 3
Simulation results for auxiliary variable selection based on 100 replications of SRS samples from four fixed populations of size $N = 1000$. (Here the MSE Ratio is the ratio of MSE of the SBLL estimator calculated by using the selected model to the MSE of the oracle SBLL estimates computed by using the true model.)

Model	σ_0	n	Forward				Backward			
			C	U	O	MSE Ratio	C	U	O	MSE Ratio
1	0.1	50	72	0	28	1.150	73	0	27	1.124
		100	97	0	3	1.001	97	0	3	1.001
		200	99	0	1	0.999	99	0	1	0.999
	0.4	50	76	0	24	1.147	77	0	23	1.145
		100	98	0	2	1.002	98	0	2	1.002
		200	100	0	0	1.000	100	0	0	1.000
2	0.1	50	87	0	13	1.255	87	0	13	1.255
		100	96	0	4	1.012	96	0	4	1.012
		200	100	0	0	1.000	100	0	0	1.000
	0.4	50	79	0	21	1.019	80	0	20	1.022
		100	98	0	2	1.000	98	0	2	1.000
		200	100	0	0	1.000	100	0	0	1.000
3	0.1	50	87	0	13	1.082	86	0	14	1.082
		100	91	0	9	1.000	91	0	9	1.001
		200	100	0	0	1.000	100	0	0	1.000
	0.4	50	83	0	17	1.020	83	0	17	1.020
		100	99	0	1	1.000	99	0	1	1.000
		200	100	0	0	1.000	100	0	0	1.000
4	0.1	50	68	0	32	1.277	69	0	31	1.277
		100	88	0	12	1.029	88	0	12	1.029
		200	100	0	0	1.000	100	0	0	1.000
	0.4	50	69	0	31	1.063	69	0	31	1.063
		100	97	0	3	1.000	97	0	3	1.031
		200	100	0	0	1.000	100	0	0	1.000

generated from the population of size 10 000. Remarkably, it takes on average less than 60 s to get the SBLL estimator even when the dimension reaches 50.

In Table 2 we give the Monte Carlo bias and standard error of the SBLL estimator based on its sampling distribution over 1000 replications. Table 2 also shows the square root of the average estimated variance of the population total (15). We see that the biases of the SBLL estimator are very small and the variance estimator appears to perform well for medium sample size.

Next we conducted simulations to evaluate the performance of the variable selection method. We generated 100 replications for each of the above models. The variables were searched from $\{1, 2, \ldots, 10\}$ for all methods and we set the maximum number of variables allowed in the model to be 10. Table 3 shows the number of correct fit (C), underfit (U) and overfit (O) based on the BIC in (18) over 100 simulation runs. Here underfitting means that the method misses at least one of the significant variables. From Table 3, we can see that both the forward and the backward selection procedures perform very well for a moderately large sample size. We also obtained the ratio of MSE of the SBLL estimates calculated by using the selected model to the MSE of the oracle SBLL estimates computed by using the true model. In all the cases, the ratios are very close to 1 or exactly 1 for moderately large sample size.

6. Discussion

Nonparametric additive methods enhance the flexibility of the models that survey practitioners use. However, due to the limitations in either interpretability, computational complexity or theoretical reliability, these models have not been widely used as general tools in survey data analysis. In this paper, we have advanced additive models as flexible, computationally efficient and theoretically attractive tools for studying survey data. We also developed a consistent procedure to select the significant auxiliary variables under simple random sampling design.

The proposed method in this paper is appropriate only for survey data that follow a simple additive model. The limitation of the basic additive model is that the interactions between the input features are not considered. There are other models, for instance, single-index model [27], additive model with second-order interaction terms [24], which reduce dimensionality but also incorporate interactions. The additive partially linear model [11] is another parsimonious candidate when one believes that the relationship between the study variable and some of the auxiliary variables has a parametric form, while the relationship between the study variable and the remaining auxiliary covariates may not be linear. These alternative models are supposed to be more efficient in certain cases, but obtaining the asymptotics is likely to be very complicated, thus we leave it as future research work.

Finally, in our methodology development, we have assumed that the auxiliary variables are available for all population elements. It would be interesting to consider the limited auxiliary information case [5] where only some summary quantities such as means are available at the population level. This is also a challenging problem for future research.

Acknowledgments

The research work of the first author was supported by NSF grant DMS-0905730. The authors thank Professor Lijian Yang for helpful discussions. The authors would also like to thank two anonymous referees for their insightful comments.

Appendix

To show the asymptotic properties of the proposed estimator $\hat{t}_{y,\mathrm{SBLL}}$, we first introduce an "oracle" SBLL estimator of t_y if the entire realization were known.

A.1. The population-based estimator

If the entire realization were known, let $\mathbf{\Gamma}_U = \{\mathbf{\Gamma}(\mathbf{x}_i)^{\mathrm{T}}\}_{i\in U_N}$ be the population-based truncated power spline matrix, where $\mathbf{\Gamma}(\mathbf{x})$ is given in (3). Let $\mathbf{y}$ be the vector of the response values y_i for $i \in U_N$. Further let $\mathbf{B}_U = (\mathbf{\Gamma}_U^{\mathrm{T}}\mathbf{\Gamma}_U)^{-1}\mathbf{\Gamma}_U^{\mathrm{T}}\mathbf{y}$. The centered pilot estimators of $m_\alpha(x_\alpha)$ at the first stage are

$$\tilde{m}_\alpha(x_\alpha) = \mathbf{\Gamma}(\mathbf{x})^{\mathrm{T}}\mathbf{D}_\alpha\mathbf{B}_U - N^{-1}\mathbf{1}_N^{\mathrm{T}}\mathbf{\Gamma}_U\mathbf{D}_\alpha\mathbf{B}_U, \tag{A.1}$$

where vector $\mathbf{1}_N^{\mathrm{T}} = \{1, 1, \ldots, 1\}$ of length N. The pilot estimators for all elements in the population are denoted by

$$\tilde{\mathbf{m}}_\alpha \equiv \{\tilde{m}_\alpha(x_{i\alpha})\}_{i\in U_N} = (\mathbf{I} - N^{-1}\mathbf{1}_N\mathbf{1}_N^{\mathrm{T}})\mathbf{\Gamma}_U\mathbf{D}_\alpha\mathbf{B}_U, \quad \alpha = 1, \ldots, d.$$

For the second stage kernel smoothing, define the matrices

$$\mathbf{X}_{Ui\alpha} = \left\{\begin{pmatrix}1 & x_{k\alpha} - x_{i\alpha}\end{pmatrix}\right\}_{k\in U_N}, \qquad \mathbf{W}_{Ui\alpha} = \mathrm{diag}\{K_h(x_{k\alpha} - x_{i\alpha})\}_{k\in U_N}.$$

Then the SBLL estimator of each component at $\mathbf{x}_i$ is given by

$$\tilde{m}_{i\alpha}^* \equiv \mathbf{e}_1^{\mathrm{T}}(\mathbf{X}_{Ui\alpha}^{\mathrm{T}}\mathbf{W}_{Ui\alpha}\mathbf{X}_{Ui\alpha})^{-1}\mathbf{X}_{Ui\alpha}^{\mathrm{T}}\mathbf{W}_{Ui\alpha}\tilde{\mathbf{y}}_\alpha, \tag{A.2}$$

where $\tilde{\mathbf{y}}_\alpha = \mathbf{y} - \frac{1}{N}t_y\mathbf{1}_N - \sum_{\beta\neq\alpha}\tilde{\mathbf{m}}_\beta$ is collection of the pseudo-responses. The SBLL estimator of $m(\mathbf{x}_i)$ based on the entire population is given by

$$\tilde{m}_i^* = \frac{1}{N}t_y + \sum_{\alpha=1}^{d}\tilde{m}_{i\alpha}^*, \quad i \in U_N. \tag{A.3}$$

Clearly, $\tilde{m}_i^*$ is the prediction at $\mathbf{x}_i$ based on the entire finite population. If these $\tilde{m}_i^*$ were known, a design-unbiased estimator of t_y would be

$$\tilde{t}_{y,\mathrm{SBLL}} = \sum_{i\in U_N}\tilde{m}_i^* + \sum_{i\in s}\frac{y_i - \tilde{m}_i^*}{\pi_i}. \tag{A.4}$$

The proof of the asymptotic properties of $\hat{t}_{y,\mathrm{SBLL}}$ uses reasoning similar to that in [2], in which a key step is the Taylor linearization. Recall that our proposed estimator involves two smoothing stages: spline smoothing in the first stage and kernel smoothing in the second stage. In the following, we establish the Taylor linearization for these two smoothing stages one by one.

A.2. Taylor linearization at the first stage

Lemma A.1. *Under assumptions* (A1)–(A7), *for any* $\alpha = 1, \ldots, d$,

$$\lim_{N\to\infty}\sup_{x_\alpha\in[0,1]}|\hat{m}_\alpha(x_\alpha) - \tilde{m}_\alpha(x_\alpha)| = O_p\{J_N(N^{-1}\log N)^{1/2}\},$$

where $\hat{m}_\alpha(x_\alpha)$ *and* $\tilde{m}_\alpha(x_\alpha)$ *are the pilot estimators given in* (11) *and* (A.1).

Proof. Let $\mathbf{S} = N^{-1}\mathbf{\Gamma}_U^T\mathbf{\Gamma}_U$ and $\mathbf{V} = N^{-1}\mathbf{\Gamma}_U^T\mathbf{y}$ be matrices with components $s_{jj'} = N^{-1}\sum_{k\in U_N}\Gamma_{U,kj}\Gamma_{U,kj'}$ and $v_j = N^{-1}\sum_{k\in U_N}\Gamma_{U,kj}y_k$, respectively. Denote $\mathbf{S}_\pi = N^{-1}\mathbf{\Gamma}_s^T\mathbf{\Pi}_s\mathbf{\Gamma}_s$ and $\mathbf{V}_\pi = N^{-1}\mathbf{\Gamma}_s^T\mathbf{\Pi}_s\mathbf{y}_s$ the sample versions of the matrices $\mathbf{S}$ and $\mathbf{V}$ with components $s_{\pi,jj'} = N^{-1}\sum_{k\in s}\Gamma_{s,kj}\Gamma_{s,kj'}/\pi_k$ and $v_{\pi,j} = N^{-1}\sum_{k\in s}\Gamma_{s,kj}y_k/\pi_k$. For each $\alpha = 1,\ldots,d$ and the spline basis $\mathbf{\Gamma}(\mathbf{x})$ in (3), let

$$\zeta(\mathbf{S}_\pi, \mathbf{V}_\pi; x_\alpha) = \mathbf{\Gamma}(\mathbf{x})^T\mathbf{D}_\alpha(\mathbf{S}_\pi^{-1}\mathbf{V}_\pi - \mathbf{S}^{-1}\mathbf{V}) \tag{A.5}$$

be a nonlinear function of $\{s_{\pi,jj'}\}_{1\le j,j'\le G_d}$ and $\{v_{\pi,j}\}_{j=1}^{G_d}$ with respect to x_α. The difference $\hat{m}_\alpha(x_\alpha) - \tilde{m}_\alpha(x_\alpha) = \zeta(\mathbf{S}_\pi, \mathbf{V}_\pi; x_\alpha) + O_p(N^{-1/2})$. Simple calculation shows that the first order derivatives of ζ in (A.5) of $s_{\pi,jj'}$ and $v_{\pi,j}$ are

$$\frac{\partial\zeta}{\partial s_{\pi,jj'}} = \mathbf{\Gamma}(\mathbf{x})^T\mathbf{D}_\alpha(-\mathbf{S}_\pi^{-1}\mathbf{\Lambda}_{jj'}\mathbf{S}_\pi^{-1})\mathbf{V}_\pi, \quad 1\le j,j'\le G_d,$$
$$\frac{\partial\zeta}{\partial v_{\pi,j}} = \mathbf{\Gamma}(\mathbf{x})^T\mathbf{D}_\alpha\mathbf{S}_\pi^{-1}\boldsymbol{\lambda}_j, \quad 1\le j\le G_d,$$

where $\boldsymbol{\lambda}_j$ is a G_d-vector with "1" in the jth component and "0" elsewhere; and $\mathbf{\Lambda}_{jj'}$ is a $G_d\times G_d$ matrix with "1" in positions (j,j') and (j',j) and "0" everywhere else.

Using the Taylor linearization, one can approximate ζ in (A.5) by a linear one so that the difference between $\hat{m}_\alpha(x_\alpha)$ and $\tilde{m}_\alpha(x_\alpha)$ can be decomposed as $\sum_{j=1}^{G_d}\varphi_{\alpha j}(x_\alpha)(v_{\pi,j}-v_j) - \hat{c}_\alpha + \tilde{c}_\alpha - \sum_{1\le j,j'\le G_d}\psi_{\alpha jj'}(x_\alpha)(s_{\pi,jj'}-s_{jj'}) + Q_{\alpha N}(x_\alpha)$, where for any $1\le j,j'\le G_d$,

$$\varphi_{\alpha j}(x_\alpha) = \left.\frac{\partial\zeta}{\partial v_{\pi,j}}\right|_{v_{\pi,j}=v_j} = \mathbf{\Gamma}(\mathbf{x})^T\mathbf{D}_\alpha\mathbf{S}^{-1}\boldsymbol{\lambda}_j,$$
$$\psi_{\alpha jj'}(x_\alpha) = \left.\frac{\partial\zeta}{\partial s_{\pi,jj'}}\right|_{s_{\pi,jj'}=s_{jj'}} = \mathbf{\Gamma}(\mathbf{x})^T\mathbf{D}_\alpha(\mathbf{S}^{-1}\mathbf{\Lambda}_{jj'}\mathbf{S}^{-1})\mathbf{V},$$

and $Q_{\alpha N}(x_\alpha)$ is the remainder. Note that

$$\begin{aligned}\sum_{j=1}^{G_d}\varphi_{\alpha j}(x_\alpha)(v_{\pi,j}-v_j) &= N^{-1}\sum_{k\in U_N}\sum_{j=1}^{G_d}\varphi_{\alpha j}(x_\alpha)\Gamma_{U,kj}y_k\left(1-\frac{I_k}{\pi_k}\right)\\ &\quad - N^{-1}\sum_{k\in U_N}\sum_{j=1}^{G_d}\varphi_{\alpha j}(x_\alpha)(\Gamma_{U,kj}-\Gamma_{s,kj})\left(1-\frac{I_k}{\pi_k}\right)y_k\\ &\quad + N^{-1}\sum_{k\in U_N}\sum_{j=1}^{G_d}\varphi_{\alpha j}(x_\alpha)y_k(\Gamma_{U,kj}-\Gamma_{s,kj}).\end{aligned}$$

By the discretization method given in Lemma A.4 of [30], the Borel–Cantelli lemma entails that each single term in the right-hand side of the above is of the order $O_p\{J_N(N^{-1}\log N)^{1/2}\}$. Therefore, we have

$$\sup_{x_\alpha\in[0,1]}\left|\sum_{j=1}^{G_d}\varphi_{\alpha j}(x_\alpha)(v_{\pi,j}-v_j)\right| = O_p\{J_N(N^{-1}\log N)^{1/2}\}.$$

Similar arguments lead to $\sup_{x_\alpha\in[0,1]}\left|\sum_{1\le j,j'\le G_d}\psi_{\alpha jj'}(x_\alpha)(s_{\pi,jj'}-s_{jj'})\right|$ to be of the order $O_p\{J_N(N^{-1}\log N)^{1/2}\}$, and $\sup_{x_\alpha\in[0,1]}|Q_{\alpha N}(x_\alpha)| = o_p\{J_N(N^{-1}\log N)^{1/2}\}$. Thus $\sup_{x_\alpha\in[0,1]}\zeta(\mathbf{S}_\pi, \mathbf{V}_\pi; x_\alpha) = O_p\{J_N(N^{-1}\log N)^{1/2}\}$. The desired result is established. □

A.3. Taylor linearization at the second stage

Let

$$t_{i\alpha q} = \sum_{k\in U_N}K_h(x_{k\alpha}-x_{i\alpha})(x_{k\alpha}-x_{i\alpha})^{q-1},$$
$$\hat{t}_{i\alpha q} = \sum_{k\in s}\frac{1}{\pi_k}K_h(x_{k\alpha}-x_{i\alpha})(x_{k\alpha}-x_{i\alpha})^{q-1},$$

for $q = 1, 2, 3$ and

$$t_{i\alpha q} = \sum_{k\in U_N} K_h(x_{k\alpha} - x_{i\alpha})(x_{k\alpha} - x_{i\alpha})^{q-4}\tilde{y}_{k\alpha},$$

$$\hat{t}_{i\alpha q} = \sum_{k\in s} \frac{1}{\pi_k} K_h(x_{k\alpha} - x_{i\alpha})(x_{k\alpha} - x_{i\alpha})^{q-4}\hat{y}_{k\alpha},$$

for $q = 4, 5$. We rewrite $\tilde{m}^*_{i\alpha}$ in (A.2) and $\hat{m}^*_{i\alpha}$ in (13) by

$$\tilde{m}^*_{i\alpha} = \frac{t_{i\alpha 3}t_{i\alpha 4} - t_{i\alpha 2}t_{i\alpha 5}}{t_{i\alpha 1}t_{i\alpha 3} - t^2_{i\alpha 2}}, \qquad \hat{m}^*_{i\alpha} = \frac{\hat{t}_{i\alpha 3}\hat{t}_{i\alpha 4} - \hat{t}_{i\alpha 2}\hat{t}_{i\alpha 5}}{\hat{t}_{i\alpha 1}\hat{t}_{i\alpha 3} - \hat{t}^2_{i\alpha 2}}.$$

Let $z_{i\alpha k} = \sum_{q=1}^5 \frac{\partial \hat{m}^*_{i\alpha}}{\partial(N^{-1}\hat{t}_{i\alpha q})}\Big|_{\hat{\mathbf{t}}_{i\alpha}=\mathbf{t}_{i\alpha}} z_{i\alpha kq}$, where $\mathbf{t}_{i\alpha} = \{t_{i\alpha q}\}_{q=1}^5$ and

$$z_{i\alpha kq} = \begin{cases} K_h(x_{k\alpha} - x_{i\alpha})(x_{k\alpha} - x_{i\alpha})^{q-1}, & \text{for } q = 1, 2, 3, \\ K_h(x_{k\alpha} - x_{i\alpha})(x_{k\alpha} - x_{i\alpha})^{q-4}y_{k\alpha}, & \text{for } q = 4, 5. \end{cases}$$

Then one can approximate $\hat{m}^*_{i\alpha} - \tilde{m}^*_{i\alpha}$ by a linear sum, i.e.,

$$\hat{m}^*_{i\alpha} - \tilde{m}^*_{i\alpha} = \frac{1}{N}\sum_{k\in U_N} z_{i\alpha k}\left(\frac{I_k}{\pi_k} - 1\right) - L_{i\alpha N} + R_{i\alpha N}, \tag{A.6}$$

where $L_{i\alpha N} = \sum_{q=1}^4 L_{i\alpha Nq}$ with

$$L_{i\alpha N1} = \frac{1}{N^2}(\hat{t}_y - t_y)\frac{\partial \hat{m}^*_{i\alpha}}{\partial(N^{-1}\hat{t}_{i\alpha 4})}\Big|_{\hat{\mathbf{t}}_{i\alpha}=\mathbf{t}_{i\alpha}} \sum_{k\in U_N} z_{i\alpha k1}\left(\frac{I_k}{\pi_k} - 1\right),$$

$$L_{i\alpha N2} = \frac{1}{N}\frac{\partial \hat{m}^*_{i\alpha}}{\partial(N^{-1}\hat{t}_{i\alpha 4})}\Big|_{\hat{\mathbf{t}}_{i\alpha}=\mathbf{t}_{i\alpha}} \sum_{k\in U_N} z_{i\alpha k1}\left(\frac{I_k}{\pi_k} - 1\right)\sum_{\beta\neq\alpha}\{\hat{m}_\beta(x_{k\beta}) - \tilde{m}_\beta(x_{k\beta})\},$$

$$L_{i\alpha N3} = \frac{1}{N^2}(\hat{t}_y - t_y)\frac{\partial \hat{m}^*_{i\alpha}}{\partial(N^{-1}\hat{t}_{i\alpha 5})}\Big|_{\hat{\mathbf{t}}_{i\alpha}=\mathbf{t}_{i\alpha}} \sum_{k\in U_N} z_{i\alpha k2}\left(\frac{I_k}{\pi_k} - 1\right),$$

$$L_{i\alpha N4} = \frac{1}{N}\frac{\partial \hat{m}^*_{i\alpha}}{\partial(N^{-1}\hat{t}_{i\alpha 5})}\Big|_{\hat{\mathbf{t}}_{i\alpha}=\mathbf{t}_{i\alpha}} \sum_{k\in U_N} z_{i\alpha k2}\left(\frac{I_k}{\pi_k} - 1\right)\sum_{\beta\neq\alpha}\{\hat{m}_\beta(x_{k\beta}) - \tilde{m}_\beta(x_{k\beta})\},$$

and $R_{i\alpha N}$ is the remainder. Similar to the proof of Lemma 3 in [2],

$$\frac{n_N}{N}\sum_{i\in U_N} E_p[R^2_{i\alpha N}] = O(n_N^{-1}h_N^{-2}). \tag{A.7}$$

Lemma A.2. *Under assumptions* (A1)–(A8), $N^{-1}\sum_{i\in U_N} E_p(L^2_{i\alpha N}) \to 0$.

Proof. By the Cauchy–Schwartz inequality, it suffices to show that for $q = 1, \ldots, 4$, $N^{-1}\sum_{i\in U_N} E_p(L^2_{i\alpha Nq}) \to 0$. Without loss of generality, we only show the cases for $q = 1$ and 2. Similarly to the proof of Lemma 2 (v) in [2], the first order derivatives of $\hat{m}^*_{i\alpha}$ with respect to $N^{-1}\hat{t}_{i\alpha q}$ evaluated at $\hat{\mathbf{t}}_i = \mathbf{t}_i$ are uniformly bounded in i. So by assumption (A7)

$$\begin{aligned}\frac{1}{N}\sum_{i\in U_N} E_p(L^2_{i\alpha N1}) &= \frac{1}{N^5}E_p\left[\left\{(\hat{t}_y - t_y)\frac{\partial \hat{m}^*_{i\alpha}}{\partial(N^{-1}\hat{t}_{i\alpha 4})}\Big|_{\hat{\mathbf{t}}_i=\mathbf{t}_i}\sum_{k\in U_N} z_{i\alpha k1}\left(\frac{I_k}{\pi_k} - 1\right)\right\}^2\right] \\ &\le \frac{C}{N^5}\sum_{j,k,l,p\in U_N}\left|z_{i\alpha j1}z_{i\alpha l1}y_k y_p\frac{\pi_{jl} - \pi_j\pi_l}{\pi_j\pi_l}\frac{\pi_{kp} - \pi_k\pi_p}{\pi_k\pi_p}\right| \le \frac{C}{N^3}\sum_{k,p\in U_N}|y_k y_p| \to 0.\end{aligned}$$

Next

$$\begin{aligned}\frac{1}{N}\sum_{i\in U_N} E_p(L_{i\alpha N2})^2 &= \frac{1}{N^3}E_p\left[\frac{\partial \hat{m}^*_{i\alpha}}{\partial(N^{-1}\hat{t}_{i\alpha 4})}\Big|_{\hat{\mathbf{t}}_{i\alpha}=\mathbf{t}_{i\alpha}}\sum_{k\in U_N} z_{i\alpha k1}\left(\frac{I_k}{\pi_k} - 1\right)\sum_{\beta\neq\alpha}\{\hat{m}_\beta(x_{k\beta}) - \tilde{m}_\beta(x_{k\beta})\}\right]^2 \\ &\le CN^{-3}\sum_{k\in U_N}\sum_{l\in U_N} E_p\left|\left(\frac{I_k}{\pi_k} - 1\right)\left(\frac{I_l}{\pi_l} - 1\right)\sum_{\beta\neq\alpha}\sum_{\gamma\neq\alpha}\{\hat{m}_\beta(x_{k\beta}) - \tilde{m}_\beta(x_{k\beta})\}\{\hat{m}_\gamma(x_{l\gamma}) - \tilde{m}_\gamma(x_{l\gamma})\}\right|.\end{aligned}$$

By Lemma A.1, $N^{-1}\sum_{i\in U_N} E_p(L^2_{i\alpha N2}) \to 0$, and the lemma follows immediately. □

A.4. Proofs of Theorems 2–4

Lemma A.3. *Under assumptions* (A1)–(A8), *for the population and sample based SBLL estimators of* $m(x_{i\alpha})$ *given in* (8) *and* (A.3),

$$\lim_{N\to\infty}\frac{1}{N}E_p\left[\sum_{i\in U_N}(\hat{m}_i^*-\tilde{m}_i^*)^2\right]=0.$$

Proof. According to (A.6), one has

$$\begin{aligned}\frac{1}{N}E_p\left[\sum_{i\in U_N}(\hat{m}_{i\alpha}^*-\tilde{m}_{i\alpha}^*)^2\right]&=\frac{1}{N^3}\sum_{i\in U_N}\sum_{k,l\in U_N}\Delta_{kl}\frac{z_{i\alpha k}}{\pi_k}\frac{z_{i\alpha l}}{\pi_l}\\&\quad-\frac{2}{N^2}\sum_{i,k\in U_N}z_{i\alpha k}E_p\left[\left(\frac{I_k}{\pi_k}-1\right)(L_{i\alpha N}-R_{i\alpha N})\right]+\frac{1}{N}\sum_{i\in U_N}E_p(L_{i\alpha N}-R_{i\alpha N})^2.\end{aligned}$$

Following from Lemmma 4 in [2] and assumption (A7), the first term converges to zero as $N\to\infty$. The third term also converges to zero by (A.7) and Lemma A.2. By the Cauchy–Schwartz inequality, $\lim_{N\to\infty}\frac{1}{N}E_p\left[\sum_{i\in U_N}(\hat{m}_{i\alpha}^*-\tilde{m}_{i\alpha}^*)^2\right]=0$, $\alpha=1,\ldots,d$. Note that

$$\begin{aligned}\sum_{i\in U_N}(\hat{m}_i^*-\tilde{m}_i^*)^2&=\sum_{i\in U_N}\left\{\frac{1}{N}(\hat{t}_y-t_y)+\sum_{\alpha=1}^{d}(\hat{m}_{i\alpha}^*-\tilde{m}_{i\alpha}^*)\right\}^2\\&=\frac{1}{N}(\hat{t}_y-t_y)^2+\frac{2}{N}(\hat{t}_y-t_y)\sum_{i\in U_N}\sum_{\alpha=1}^{d}(\hat{m}_{i\alpha}^*-\tilde{m}_{i\alpha}^*)+\sum_{i\in U_N}\left\{\sum_{\alpha=1}^{d}(\hat{m}_{i\alpha}^*-\tilde{m}_{i\alpha}^*)\right\}^2.\end{aligned}$$

By assumption (A.7),

$$\frac{1}{N^2}E_p(\hat{t}_y-t_y)^2\le\left(\frac{1}{\lambda}+\frac{n_N\max_{i,j\in U_N,i\neq j}|\Delta_{ij}|}{\lambda^2}\right)\frac{1}{N^2}\sum_{i\in U_N}y_i^2\to 0.$$

Thus the desired result is obtained from the Cauchy–Schwartz inequality. □

Proof of Theorem 2. Note that $E_p[I_i]=\pi_i$ and

$$\frac{\hat{t}_{y,\mathrm{SBLL}}-t_y}{N}=\frac{1}{N}\sum_{i\in U_N}(y_i-\tilde{m}_i^*)(I_i/\pi_i-1)+\frac{1}{N}\sum_{i\in U_N}(\hat{m}_i^*-\tilde{m}_i^*)(1-I_i/\pi_i).\tag{A.8}$$

Then

$$\begin{aligned}E_p\left|\frac{\hat{t}_{y,\mathrm{SBLL}}-t_y}{N}\right|&\le\frac{1}{N}E_p\left|\sum_{i\in U_N}(y_i-\tilde{m}_i^*)(I_i/\pi_i-1)\right|\\&\quad+\frac{1}{N^2}\left\{E_p\left[\sum_{i\in U_N}(\hat{m}_i^*-\tilde{m}_i^*)^2\right]E_p\left[\sum_{i\in U_N}(1-I_i/\pi_i)^2\right]\right\}^{1/2}.\end{aligned}\tag{A.9}$$

According to assumptions (A1)–(A6), $\limsup_{N\to\infty}\frac{1}{N}\sum_{i\in U_N}(y_i-\tilde{m}_i^*)^2<\infty$. Following the same arguments of Theorem 1 in [2], the first term on the right of (A.9) converges to zero as $N\to\infty$. For the second term, (A7) implies that

$$E_p\left[\frac{1}{N}\sum_{i\in U_N}(1-I_i/\pi_i)^2\right]=\sum_{i\in U_N}\frac{\pi_i(1-\pi_i)}{N\pi_i^2}\le\frac{1}{\lambda}.$$

According to Lemma A.3, $\lim_{N\to\infty}\frac{1}{N}\sum_{i\in U_N}E_p[(\hat{m}_i^*-\tilde{m}_i^*)^2]\to 0$ and the result follows from the Markov's inequality. □

The next theorem is to derive the asymptotic mean squared error of the proposed spline estimator in (9).

Theorem A.1. *Under assumptions* (A1)–(A8),

$$n_NE_p\left(\frac{\hat{t}_{y,\mathrm{SBLL}}-t_y}{N}\right)^2=\frac{n_N}{N^2}\sum_{i,j\in U_N}\Delta_{ij}\frac{y_i-\tilde{m}_i^*}{\pi_i}\frac{y_j-\tilde{m}_j^*}{\pi_j}+o(1).\tag{A.10}$$

Denote

$$\mathrm{AMSE}(N^{-1}\hat{t}_{y,\mathrm{SBLL}}) = \frac{1}{N^2}\sum_{i,j\in U_N}\Delta_{ij}\frac{y_i-\tilde{m}_i^*}{\pi_i}\frac{y_j-\tilde{m}_j^*}{\pi_j} \tag{A.11}$$

the asymptotic mean squared error in (A.10). The next result shows that it can be estimated consistently by $\widehat{V}(N^{-1}\hat{t}_{y,\mathrm{SBLL}})$ in (15).

Theorem A.2. *Under* (A1)–(A8),

$$\lim_{N\to\infty} n_N E_p|\widehat{V}(N^{-1}\hat{t}_{y,\mathrm{SBLL}}) - \mathrm{AMSE}(N^{-1}\hat{t}_{y,\mathrm{SBLL}})| = 0.$$

The proofs of Theorems A.1 and A.2 are somewhat trivial and we refer the readers to [29].

Proof of Theorem 3. According to (A.8),

$$\frac{\hat{t}_{y,\mathrm{SBLL}}-t_y}{N} = \frac{\tilde{t}_{y,\mathrm{SBLL}}-t_y}{N} + \sum_{i\in U_N}\frac{\tilde{m}_i^*-\hat{m}_i^*}{N}\left(\frac{I_i}{\pi_i}-1\right).$$

From the proof of Theorem A.1, $\sum_{i\in U_N}\frac{\tilde{m}_i^*-\hat{m}_i^*}{N}\left(\frac{I_i}{\pi_i}-1\right) = o_p(n_N^{-1/2})$. Theorem A.2 implies that $\widehat{V}(N^{-1}\hat{t}_{y,\mathrm{SBLL}})/\mathrm{AMSE}(N^{-1}\hat{t}_{y,\mathrm{SBLL}}) \to 1$ in probability. The desired result follows. □

Proof of Theorem 4. Let

$$T_1 = \frac{n_N^{1/2}}{N}\sum_{i\in U_N}\{\tilde{m}_i^* - m(\mathbf{x}_i)\}\left(1-\frac{I_i}{\pi_i}\right), \qquad T_2 = \frac{n_N^{1/2}}{N}\sum_{i\in U_N}(\hat{m}_i^*-\tilde{m}_i^*)\left(1-\frac{I_i}{\pi_i}\right),$$

$$T_3 = \frac{n_N^{1/2}}{N}\sum_{i\in U_N}\sigma(\mathbf{x}_i)\varepsilon_i\left(\frac{I_i}{\pi_i}-1\right).$$

Then $n_N^{1/2}N^{-1}(\hat{t}_{y,\mathrm{SBLL}}-t_y)$ can be represented as the sum of T_1, T_2 and T_3. For the first term,

$$\begin{aligned} ET_1^2 &= \frac{n_N}{N^2}\sum_{i,j\in U_N}[E(m(\mathbf{x}_i)-\tilde{m}_i^*)(m(\mathbf{x}_j)-\tilde{m}_j^*)]\frac{\Delta_{ij}}{\pi_i\pi_j} \\ &\le \frac{n_N}{N}\left(\frac{1}{\lambda}+\frac{N\max_{i,j\in U_N, i\neq j}|\Delta_{ij}|}{\lambda^2}\right)\frac{1}{N}\sum_{i\in U_N}E\{m(\mathbf{x}_i)-\tilde{m}_i^*\}^2. \end{aligned}$$

By Theorem 2.1 in [30], $|m(\mathbf{x}_i)-\tilde{m}_i^*| = o_p(n^{-2/5}\log n)$, for any $i\in U_N$, which implies that $ET_1^2\to 0$. Now for T_2

$$ET_2^2 \le \frac{n_N}{N}\left(\frac{1}{\lambda}+\frac{N\max_{i,j\in U_N, i\neq j}|\Delta_{ij}|}{\lambda^2}\right)\frac{1}{N}\sum_{i\in U_N}E(\hat{m}_i^*-\tilde{m}_i^*)^2.$$

By Lemma A.1, $ET_2^2\to 0$. Finally,

$$ET_3^2 = \frac{n_N}{N^2}\sum_{i\in U_N}\sigma^2(\mathbf{x}_i)\frac{1-\pi_i}{\pi_i} \le \frac{n_N}{N\lambda}\frac{1}{N}\sum_{i\in U_N}\sigma^2(\mathbf{x}_i),$$

$$\limsup_{N\to\infty} ET_3^2 \le \frac{1}{\lambda}\limsup_{N\to\infty}\frac{1}{N}\sum_{i\in U_N}\sigma^2(\mathbf{x}_i) < \infty.$$

By the Cauchy–Schwartz inequality the cross product terms go to zero as $N\to\infty$. The desired result follows. □

References

[1] F.J. Breidt, G. Claeskens, J.D. Opsomer, Model-assisted estimation for complex surveys using penalised splines, Biometrika 92 (2005) 831–846.
[2] F.J. Breidt, J.D. Opsomer, Local polynomial regression estimators in survey sampling, Ann. Statist. 28 (2000) 1026–1053.
[3] F.J. Breidt, J.D. Opsomer, A.A. Johnson, M.G. Ranalli, Semiparametric model-assisted estimation for natural resource surveys, Surv. Methodol. 33 (2007) 35–44.
[4] R.L. Chambers, Robust case-weighting for multipurpose establishment surveys, J. Official Statist. 12 (1996) 3–32.
[5] R.L. Chambers, A.H. Dorfman, S. Wang, Limited information likelihood analysis of survey data, J. R. Stat. Soc. Ser. B 60 (1998) 397–411.

[6] R.L. Chambers, A.H. Dorfman, T.E. Wehrly, Bias robust estimation in finite populations using nonparametric calibration, J. Amer. Statist. Assoc. 88 (1993) 268–277.
[7] J.C. Deville, C.E. Särndal, Calibration estimators in survey sampling, J. Amer. Statist. Assoc. 87 (1992) 376–382.
[8] A.H. Dorfman, Nonparametric regression for estimating totals in finite populations, in: Proceedings of the Section on Survey Research Methods, Amer. Statist. Assoc., Alexandria, VA, 1992, pp. 622–625.
[9] A.H. Dorfman, P. Hall, Estimators of the finite population distribution function using nonparametric regression, Ann. Statist. 21 (1993) 1452–1475.
[10] J. Fan, I. Gijbels, Local Polynomial Modelling, its Applications, Chapman, Hall, London, 1996.
[11] Y. Fan, Q. Li, A kernel-based method for estimating additive partially linear models, Statist. Sinica 13 (2003) 739–762.
[12] W. Härdle, Applied Nonparametric Regression, Cambridge University Press, Cambridge, 1990.
[13] T.J. Hastie, R.J. Tibshirani, Generalized Additive Models, Chapman, Hall, London, 1990.
[14] J.Z. Huang, L. Yang, Identification of nonlinear additive autoregression models, J. R. Stat. Soc. Ser. B 66 (2004) 463–477.
[15] J.K. Kim, Calibration estimation using empirical likelihood in survey sampling, Statist. Sinica 19 (2009) 145–158.
[16] O.B. Linton, J.P. Nielsen, A kernel method of estimating structured nonparametric regression based on marginal integration, Biometrika 82 (1995) 93–101.
[17] E. Mammen, O. Linton, J. Nielsen, The existence, asymptotic properties of a backfitting projection algorithm under weak conditions, Ann. Statist. 27 (1999) 1443–1490.
[18] C. Martins-Filho, K. Yang, Finite sample performance of kernel-based regression methods for non-parametric additive models under common bandwidth selection criterion, J. Nonparametr. Stat. 19 (2007) 23–62.
[19] J.D. Opsomer, F.J. Breidt, G.G. Moisen, G. Kauermann, Model-assisted estimation of forest resources with generalized additive models (with discussion), J. Amer. Statist. Assoc. 102 (2007) 400–416.
[20] C.E. Särndal, S. Lundström, Estimation in Surveys with Nonresponse, Wiley, New York, 2005.
[21] C.E. Särndal, B. Swensson, J. Wretman, The weighted residual technique for estimating the variance of the general regression estimator of the finite population total, Biometrika 76 (1989) 527–537.
[22] C.E. Särndal, B. Swensson, J. Wretman, Model Assisted Survey Sampling, Springer-Verlag, New York, 1992.
[23] G.E. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[24] S. Sperlich, D. Tjøstheim, L. Yang, Nonparametric estimation and testing of interaction in additive models, Econom. Theory 18 (2002) 197–251.
[25] C.J. Stone, Additive regression, other nonparametric models, Ann. Statist. 13 (1985) 689–705.
[26] C.J. Stone, The use of polynomial splines, their tensor products in multivariate function estimation, Ann. Statist. 22 (1994) 118–184.
[27] L. Wang, Single-index model-assisted estimation in survey sampling, J. Nonparametr. Stat. 21 (2009) 487–504.
[28] S. Wang, A.H. Dorfman, A new estimator for the finite population distribution function, Biometrika 83 (1997) 639–652.
[29] L. Wang, S. Wang, Nonparametric additive model-assisted estimation for survey data, http://arxiv.org/abs/1101.0831.
[30] L. Wang, L. Yang, Spline-backfitted kernel smoothing of nonlinear additive autoregression model, Ann. Statist. 35 (2007) 2474–2503.