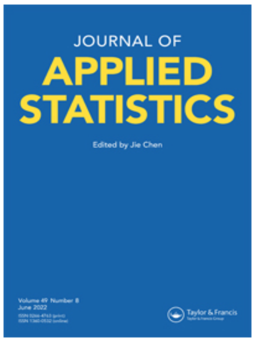

Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

Model-assisted estimation in high-dimensional settings for survey data

Mehdi Dagdoug, Camelia Goga & David Haziza

To cite this article: Mehdi Dagdoug, Camelia Goga & David Haziza (2022): Model-assisted estimation in high-dimensional settings for survey data, Journal of Applied Statistics, DOI: 10.1080/02664763.2022.2047905

To link to this article: https://doi.org/10.1080/02664763.2022.2047905

View supplementary material

Published online: 17 Mar 2022.

Submit your article to this journal

Article views: 49

View related articles

View Crossmark data

Citing articles: 1 View citing articles

JOURNAL OF APPLIED STATISTICS
https://doi.org/10.1080/02664763.2022.2047905

Taylor & Francis
Taylor & Francis Group

Model-assisted estimation in high-dimensional settings for survey data

Mehdi Dagdoug[a], Camelia Goga[a] and David Haziza[b]

[a]Laboratoire de Mathématiques de Besançon, Université de Bourgogne Franche-Comté, Besançon, France; [b]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada

ABSTRACT

Model-assisted estimators have attracted a lot of attention in the last three decades. These estimators attempt to make an efficient use of auxiliary information available at the estimation stage. A working model linking the survey variable to the auxiliary variables is specified and fitted on the sample data to obtain a set of predictions, which are then incorporated in the estimation procedures. A nice feature of model-assisted procedures is that they maintain important design properties such as consistency and asymptotic unbiasedness irrespective of whether or not the working model is correctly specified. In this article, we examine several model-assisted estimators from a design-based point of view and in a high-dimensional setting, including linear regression and penalized estimators. We conduct an extensive simulation study using data from the Irish Commission for Energy Regulation Smart Metering Project, to assess the performance of several model-assisted estimators in terms of bias and efficiency in this high-dimensional data set.

ARTICLE HISTORY
Received 30 September 2020
Accepted 20 February 2022

KEYWORDS
Design consistency; elastic net; Lasso; random forest; ridge regression; XGBoost

1. Introduction

Surveys conducted by national statistical offices (NSO) aim at estimating finite population parameters, which are those describing some aspects of the finite population under study. In this article, the interest lies in estimating the population total of a survey variable Y. Population totals can be estimated unbiasedly using the well-known Horvitz–Thompson estimator [23]. In the absence of nonsampling errors, the Horvitz–Thompson estimator is unbiased with respect to the customary design-based inferential approach, whereby the properties of estimators are evaluated with respect to the sampling design; e.g., see [36]. However, Horvitz–Thompson type estimators may exhibit a large variance in some situations. The efficiency of the Horvitz–Thompson estimator can be improved by incorporating some auxiliary information, capitalizing on the relationship between the survey variable Y and a set of auxiliary variables $\mathbf{x}$. The resulting estimation procedures, referred to as model-assisted estimation procedures, use a working model as a vehicle for constructing point estimators. Model-assisted estimators remain design-consistent even if the working

CONTACT David Haziza ✉ dhaziza@uottawa.ca .

Supplemental data for this article can be accessed here. https://doi.org/10.1080/02664763.2022.2047905

model is misspecified, which is a desirable feature. When the working model provides an adequate description of the relationship between Y and $\mathbf{x}$, model-assisted estimators are expected to be more efficient than the Horvitz–Thompson estimator.

The class of model-assisted estimators includes a wide variety of procedures, some of which have been extensively studied in the literature both theoretically and empirically. When the working model is the customary linear regression model, the resulting estimator is the well-known generalized regression estimator (GREG), e.g. Särndal [35], Särndal and Wright [37] and Särndal *et al.* [36]. Other works include model-assisted procedures based on generalized linear models [15,26], local polynomial regression [4], splines [3,16,17,27], neural nets [30], generalized additive models [31], nonparametric additive models [42], regression trees [29,41] and random forests [12].

Due to the recent advances of information technology, NSOs have now access to a variety of data sources, some of which may exhibit a large number of observations on a large number of variables. So far, the properties of model-assisted estimator have been established under the customary asymptotic framework in finite population sampling [24] for which both the population size N and the sample size n increase to infinity, while assuming that the number of auxiliary variables p is fixed. In other words, existing results require n to be large relative to p. This framework is generally not adequate in the context of high-dimensional data sets as p may be of the same order as n, or even larger, i.e. $p > n$. A more appropriate asymptotic framework would let p increase to infinity in addition to N and n. Cardot *et al.* [8] studied dimension reduction through principal component analysis and established the design consistency of the resulting calibration estimator. More recently, Ta *et al.* [38] investigated the properties of the GREG estimator from a model point of view and when p is allowed to diverge and [10] studied the asymptotic variance of the calibration estimator when the number p of calibration variables is going to infinity.

The aim of this paper is to give a general consistency result for a class of model-assisted estimators when the number p of auxiliary variables is allowed to grow to infinity. This class of model-assisted estimators includes the GREG estimator as well as model-assisted estimators based on penalization methods such as ridge, lasso and elastic net. The latter methods were proposed to cope with multicollinearity between predictors in a high-dimension setting. Under mild regularity assumptions, we show that these model-assisted estimators are design consistent provided that p^3/n goes to zero. As we argue in Section 3, this rate can be improved if one is willing to make additional assumptions about the rate of convergence of the estimated regression coefficient. In particular, we lay out a set of additional conditions under which the model-assisted ridge estimator is consistent if p/n goes to zero and moreover, is $\sqrt{n}$-consistent if $p = \mathcal{O}(n^a)$ with $a \in [0, 1/2)$. Also, provided that the predictors are orthogonal, we show that both the model-assisted lasso and elastic net estimators are consistent provided that p/n goes to zero.

To the best of our knowledge, an empirical comparison of penalized or nonparametric model-assisted estimators in terms of bias and efficiency in a high-dimensional setting is currently lacking. We aim to fill this gap in the article. To assess the performance of several model-assisted estimators in a high-dimensional setting, we conduct a large simulation study using data from the Irish Commission for Energy Regulation Smart Metering Project. The data set consists of electricity consumption recorded every half an hour for a 2-year period and for more than 6000 households and businesses, leading to highly correlated data. Due to the high-dimensional feature, model-assisted estimators based on a linear model tend to breakdown and penalized and reduction dimension based estimators may provide good alternatives.

The paper is organized as follows. In Section 2, we introduce the theoretical setup. In Section 3, we investigate the asymptotic properties of several model-assisted estimators: the GREG estimator as well as estimators based on ridge regression, lasso and elastic net. Section 4 contains an empirical comparison to assess the performance of several model-assisted estimators in terms of bias and efficiency. In our empirical experiments, we included model-assisted estimators based on ridge regression, lasso and elastic net, principal component regression as well as model-assisted estimators based on CART, random forests, XGBoost and CUBIST. We considered three sampling designs: simple random sampling without replacement, stratified simple random sampling without replacement and stratified fixed-size without replacement proportional to size sampling. We make some final remarks in Section 5. The technical details, including the proofs of some results, are relegated to the Supplementary Material.

2. The setup

Consider a finite population $U = \{1, 2, \ldots, N\}$ of size N. We are interested in estimating $t_y = \sum_{i \in U} y_i$, the population total of the survey variable Y. We select a sample S from U according to a sampling design $\mathcal{P}(S)$ with first-order and second-order inclusion probabilities $\{\pi_i\}_{i \in U}$ and $\{\pi_{i\ell}\}_{i,\ell \in U}$, respectively. In the absence of nonsampling errors, the Horvitz–Thompson estimator

$$\widehat{t}_\pi = \sum_{i \in S} \frac{y_i}{\pi_i} \tag{1}$$

is design-unbiased for t_y provided that $\pi_i > 0$ for all $i \in U$; that is, $\mathbb{E}_{\mathrm{p}}(\widehat{t}_\pi) = t_y$, where $\mathbb{E}_p(\cdot)$ denotes the expectation operator with respect to the sampling design $\mathcal{P}(S)$. In the sequel, unless stated otherwise, the properties of estimators are evaluated with respect to the design-based approach. Under mild conditions [4,34], it can be shown that the Horvitz–Thompson estimator $\widehat{t}_\pi$ is design consistent for t_y.

At the estimation stage, we assume that a collection of auxiliary variables, $X_1, X_2, \ldots, X_p$, is recorded for all $i \in S$. Moreover, we assume that the corresponding population totals are available from an external source (e.g. a census or an administrative file). Let $\mathbf{x}_i = [\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{ip}]^\top$ be the $\mathbf{x}$-vector associated with unit i. Also, we denote by $\boldsymbol{X}_U = (\mathbf{x}_i^\top)_{i \in U}$ the $N \times p$ design matrix and $\boldsymbol{X}_S = (\mathbf{x}_i^\top)_{i \in S}$ its sample counterpart.

Model-assisted estimation starts with postulating the following working model:

$$\xi : \quad y_i = f(\mathbf{x}_i) + \epsilon_i, \qquad i \in U, \tag{2}$$

where $f(\cdot)$ is an unknown function and the errors ϵ_i are independent random variables such that $\mathbb{E}_\xi[\epsilon_i | \mathbf{x}_i] = 0$ and $\mathbb{V}_\xi(\epsilon_i | \mathbf{x}_i) = \sigma^2$, where σ^2 is an unknown parameter. Although we assume a homoscedastic variance structure, our results can be easily extended to the case of unequal variances of the form $\mathbb{V}_\xi(\epsilon_i | \mathbf{x}_i) = \sigma^2 \nu(\mathbf{x}_i)$ for some known function $\nu(\cdot)$.

The unknown function $f(\cdot)$ is estimated by $\widehat{f}(\cdot)$ from the sample data $(\mathbf{x}_i, y_i)_{i\in S}$. The fitted model is then used to construct the model-assisted estimator

$$\widehat{t}_{ma} = \sum_{i\in U} \widehat{f}(\mathbf{x}_i) + \sum_{i\in S} \frac{y_i - \widehat{f}(\mathbf{x}_i)}{\pi_i}, \tag{3}$$

where $\widehat{f}(\mathbf{x})$ denotes the prediction at $\mathbf{x}$ under the working model (2). Whenever the predictor $\widehat{f}(\cdot)$ is sample dependent, the estimator $\widehat{t}_{ma}$ is design biased, but can be shown to be asymptotically design unbiased and design consistent for a wide class of working models, as the population size N and the sample size n increase.

3. Least squares and penalized model-assisted estimators

3.1. The GREG estimator

Suppose that the regression function $f(\cdot)$ is approximated by a linear combination of $X_j, j = 1, \ldots, p$. The working model (2) reduces to

$$\xi : \ y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i \in U, \tag{4}$$

where $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^\top \in \mathbb{R}^p$ is a vector of unknown coefficients. Under a hypothetical census, where we observe y_i and $\mathbf{x}_i$ for all $i \in U$, the vector $\boldsymbol{\beta}$ would be estimated by $\widetilde{\boldsymbol{\beta}}$ through the ordinary least square criterion at the population level:

$$\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} ||\boldsymbol{y}_U - \boldsymbol{X}_U\boldsymbol{\beta}||_2^2 = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i\in U} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2, \tag{5}$$

where $\boldsymbol{y}_U = (y_i)_{i\in U}$. Provided that the matrix $\boldsymbol{X}_U$ is of full rank, the solution to (5) is unique and given by

$$\widetilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}_U^\top \boldsymbol{X}_U\right)^{-1} \boldsymbol{X}_U^\top \boldsymbol{y}_U = \left(\sum_{i\in U} \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} \sum_{i\in U} \mathbf{x}_i y_i. \tag{6}$$

In practice, the vector $\widetilde{\boldsymbol{\beta}}$ in (6) cannot be computed as the y-values are recorded for the sample units only. An estimator of $\widetilde{\boldsymbol{\beta}}$, denoted by $\widehat{\boldsymbol{\beta}}$, is obtained from (6) by estimating each total separately using the corresponding Horvitz–Thompson estimator. Alternatively, the estimator $\widehat{\boldsymbol{\beta}}$ can be obtained using the following weighted least square criterion at the sample level:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left(\boldsymbol{y}_S - \boldsymbol{X}_S\boldsymbol{\beta}\right)^\top \boldsymbol{\Pi}_S^{-1} \left(\boldsymbol{y}_S - \boldsymbol{X}_S\boldsymbol{\beta}\right)^\top = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i\in S} \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{\pi_i}, \tag{7}$$

where $\boldsymbol{\Pi}_S = \operatorname{diag}(\pi_i)_{i\in S}$ and $\boldsymbol{y}_S = (y_i)_{i\in S}$. Again, the solution to (7) is unique provided that $\boldsymbol{X}_S$ is of full rank and it is given by

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}_S^\top \boldsymbol{\Pi}_S^{-1} \boldsymbol{X}_S\right)^{-1} \boldsymbol{X}_S^\top \boldsymbol{\Pi}_S^{-1} \boldsymbol{y}_S = \left(\sum_{i\in S} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\pi_i}\right)^{-1} \sum_{i\in S} \frac{\mathbf{x}_i y_i}{\pi_i}. \tag{8}$$

The prediction of $f(\cdot)$ at $\mathbf{x}$ under the working model (4) is $\widehat{f}_{\mathrm{lr}}(\mathbf{x}) = \mathbf{x}^\top\widehat{\boldsymbol{\beta}}$. Plugging $\widehat{f}_{\mathrm{lr}}(\cdot)$ in (3) leads to the well-known GREG estimator [36]:

$$\begin{aligned}\widehat{t}_{\mathrm{greg}} &= \sum_{i\in U}\widehat{f}_{\mathrm{lr}}(\mathbf{x}_i) + \sum_{i\in S}\frac{y_i - \widehat{f}_{\mathrm{lr}}(\mathbf{x}_i)}{\pi_i}\\ &= \sum_{i\in U}\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}} + \sum_{i\in S}\frac{y_i - \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}}{\pi_i}. \end{aligned} \tag{9}$$

If the intercept is included in the working model, the GREG estimator reduces to the population total of the fitted values $\widehat{f}_{lr}(\mathbf{x}_i) = \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}$; that is, $\widehat{t}_{\mathrm{greg}} = \sum_{i\in U}\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}$. Also, the GREG estimator can be written as a weighted sum of the sample y-values:

$$\widehat{t}_{\mathrm{greg}} = \sum_{i\in S} w_{iS}y_i, \tag{10}$$

where

$$w_{iS} = \frac{1}{\pi_i}\left\{1 - \mathbf{x}_i^\top\left(\sum_{i\in S}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}\right)^{-1}\left(\sum_{i\in S}\frac{\mathbf{x}_i}{\pi_i} - \sum_{i\in U}\mathbf{x}_i\right)\right\}, \quad i\in S.$$

These weights can be also obtained as the solution of a calibration problem [14]. More specifically, the weights w_{iS} are such that the generalized chi-square distance $\sum_{i\in S}(w_{iS} - \pi_i^{-1})^2/\pi_i^{-1}$ is minimized subject to the calibration constraints $\sum_{i\in S} w_{iS}\mathbf{x}_i = \sum_{i\in U}\mathbf{x}_i$. This attractive feature may not be shared by model-assisted estimators derived under more general working models.

3.2. Penalized least square estimators

While model-assisted estimators based on linear regression working models are easy to implement, they tend to breakdown when the number of auxiliary variables p is growing large. Also, when some of the predictors are highly related to each other, a problem known as multicollinearity, the ordinary least square estimator $\widetilde{\boldsymbol{\beta}}$ given by (6) may be highly unstable. As noted by Hoerl and Kennard [22], 'the worse the conditioning of $\boldsymbol{X}_U^\top\boldsymbol{X}_U$, the more $\widetilde{\boldsymbol{\beta}}$ can be expected to be too long and the distance from $\widetilde{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}$ will tend to be large'. In survey sampling, the effect of multicollinearity on the stability of point estimators has first been studied by Bardsley and Chambers [1] under the model-based approach. Chambers [9] and Rao and Singh [33] studied this problem in the context of calibration. These authors noted that the use of a large number of calibration constraints may lead to highly dispersed calibration weights, potentially resulting in unstable estimators.

In a classical *iid* linear regression setting, penalization procedures such as ridge, lasso or elastic-net can be used to help circumvent some of the difficulties associated with the usual least squares estimator $\widetilde{\boldsymbol{\beta}}$. Let $\widetilde{\boldsymbol{\beta}}_{\mathrm{pen}}$ be an estimator of $\boldsymbol{\beta}$ obtained through the penalized least square criterion at the population level:

$$\widetilde{\boldsymbol{\beta}}_{\text{pen}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i\in U} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \sum_{\ell=1}^{t} \lambda_\ell ||\boldsymbol{\beta}||_{\nu_\ell}^{\gamma_\ell}, \tag{11}$$

where λ_ℓ, ν_ℓ and γ_ℓ are positive real numbers, $\|\cdot\|_\nu$ is a given norm and t is a fixed positive integer representing the number of different norm constraints. The values of ν_ℓ, γ_ℓ and t are typically predetermined. The tuning parameter λ_ℓ controls the strength of the penalty that one wants to impose on the norm of $\boldsymbol{\beta}$. Most often, the value of λ_ℓ is selected through a cross-validation procedure. The coefficients γ_ℓ and ν_ℓ are specific to the penalization method. Hence, they affect the properties of the resulting estimator $\widetilde{\boldsymbol{\beta}}_{\text{pen}}$. Three special cases are considered below.

When $t = 1, \gamma_1 = 2$ and $\nu_1 = 2, \lambda_1 = \lambda$, the estimator is known as the ridge regression estimator [21]:

$$\widetilde{\boldsymbol{\beta}}_{\text{ridge}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i\in U} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda ||\boldsymbol{\beta}||_2^2,$$

where $||\boldsymbol{\beta}||_2^2 = \sum_{j=1}^p \beta_j^2$ is the usual Euclidean norm of $\boldsymbol{\beta}$. The solution is given explicitly by

$$\widetilde{\boldsymbol{\beta}}_{\text{ridge}} = \left(\boldsymbol{X}_U^\top \boldsymbol{X}_U + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}_U^\top \boldsymbol{y}_U = \left(\sum_{i\in U} \mathbf{x}_i \mathbf{x}_i^\top + \lambda \boldsymbol{I}_p\right)^{-1} \sum_{i\in U} \mathbf{x}_i y_i, \tag{12}$$

where $\boldsymbol{I}_p$ denotes the $p \times p$ identity matrix.

When $t = 1, \nu_1 = 1$ and $\lambda_1 = \lambda$, the estimator $\widetilde{\boldsymbol{\beta}}_{\text{pen}}$ is known as the lasso estimator [39]:

$$\widetilde{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i\in U} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda ||\boldsymbol{\beta}||_1, \tag{13}$$

where $||\boldsymbol{\beta}||_1 = \sum_{j=1}^p |\beta_j|$ is the L^1-norm of $\boldsymbol{\beta}$. As for the ridge, the lasso has the effect of shrinking the coefficients but, unlike the ridge, it can set some coefficients β_j to zero. Except when the auxiliary variables are orthogonal, there is no closed-form formula for the lasso estimator $\widetilde{\boldsymbol{\beta}}_{\text{lasso}}$ [20]. In survey sampling, McConville *et al.* [28] investigated the design-based properties of the lasso model-assisted estimator for fixed p.

The elastic-net estimator, that was suggested by Zou and Hastie [43], combines two norms: the Euclidean norm $\|\cdot\|_2$ and the L^1 norm, $\|\cdot\|_1$. If, in (11), we set $t = 2, \gamma_1 = 1, \nu_1 = 1, \gamma_2 = 2, \nu_2 = 2, \lambda_1 = \lambda\alpha$ and $\lambda_1 = \lambda(1-\alpha)$, the resulting estimator is the elastic-net estimator, which can be viewed as a trade-off between the ridge estimator and the lasso estimator, realizing variable selection and regularization simultaneously:

$$\widetilde{\boldsymbol{\beta}}_{\text{en}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i\in U} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \left[\alpha ||\boldsymbol{\beta}||_1 + (1-\alpha) ||\boldsymbol{\beta}||_2^2\right],$$

for $\lambda > 0$ and $\alpha \in [0, 1]$ a parameter that is usually chosen using a grid of multiple values of α. The penalized regression estimator $\widetilde{\boldsymbol{\beta}}_{\text{pen}}$ in (11) is unknown as the y-values are not observed for the non-sample units. To overcome this issue, we use the following weighted penalized least square criterion at the sample level:

$$\widehat{\boldsymbol{\beta}}_{\text{pen}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i\in S} \frac{1}{\pi_i}\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \sum_{\ell=1}^{t} \lambda_\ell ||\boldsymbol{\beta}||_{\nu_\ell}^{\gamma_\ell}. \tag{14}$$

A model-assisted estimator based on a penalized regression procedure is obtained from (3) by replacing $\widehat{f}(\mathbf{x})$ with $\widehat{f}_{\text{pen}}(\mathbf{x}) = \mathbf{x}^\top \widehat{\boldsymbol{\beta}}_{\text{pen}}$, leading to

$$\begin{aligned}\widehat{t}_{\text{pen}} &= \sum_{i\in U} \widehat{f}_{\text{pen}}(\mathbf{x}_i) + \sum_{i\in S} \frac{y_i - \widehat{f}_{\text{pen}}(\mathbf{x}_i)}{\pi_i}\\ &= \left(\sum_{i\in U} \mathbf{x}_i^\top\right) \widehat{\boldsymbol{\beta}}_{\text{pen}} + \sum_{i\in S} \frac{y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{\text{pen}}}{\pi_i},\end{aligned} \tag{15}$$

where $\widehat{\boldsymbol{\beta}}_{\text{pen}}$ is a generic notation used to denote the estimated regression coefficient obtained through either lasso, ridge or elastic net. Unlike the GREG estimator, $\widehat{t}_{\text{greg}}$, the penalized model-assisted estimator is sensitive to unit change of the X-variables because $\widehat{\boldsymbol{\beta}}_{\text{pen}}$ is sensitive to unit change. This is why, as in the classical regression setting, standardization of the X-variables is recommended before computing $\widehat{\boldsymbol{\beta}}_{\text{pen}}$. If the intercept is included in the model, then it is usually left un-penalized.

Remark 3.1: In the case of ridge regression, the estimator $\widehat{\boldsymbol{\beta}}_{\text{ridge}}$ is given by

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = \left(\boldsymbol{X}_S^\top \boldsymbol{\Pi}_S^{-1} \boldsymbol{X}_S + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}_S^\top \boldsymbol{\Pi}_S^{-1} \boldsymbol{y}_S = \left(\sum_{i\in S} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\pi_i} + \lambda \boldsymbol{I}_p\right)^{-1} \sum_{i\in S} \frac{\mathbf{x}_i y_i}{\pi_i}. \tag{16}$$

Using (16) in (15) leads to the ridge model-assisted estimator $\hat{t}_{\text{ridge}}$ that can be expressed as a weighted sum of sampled y-values, $\hat{t}_{\text{ridge}} = \sum_{i\in S} w_{iS}(\lambda) y_i$ with weights given by

$$w_{iS}(\lambda) = \frac{1}{\pi_i}\left\{1 - \mathbf{x}_i^\top \left(\sum_{i\in S} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\pi_i} + \lambda \boldsymbol{I}_p\right)^{-1} \left(\sum_{i\in S} \frac{\mathbf{x}_i}{\pi_i} - \sum_{i\in U} \mathbf{x}_i\right)\right\}, \quad i \in S.$$

These weights can also be obtained through a penalized calibration problem. It can be shown that they minimize the penalized generalized chi-square distance, $\sum_{i\in S}(w_{iS} - \pi_i^{-1})^2/\pi_i^{-1} + \lambda^{-1} \| \sum_{i\in S} w_{iS} \mathbf{x}_i - \sum_{i\in U} \mathbf{x}_i \|_2^2$ [2,9]. If some X-variables are left un-penalized in (11), the resulting weights ensure consistency between the survey estimates and their corresponding population totals associated with these variables.

We end this section by noting that the penalized model-assisted estimator $\widehat{t}_{\text{pen}}$ is sensitive to the choice of the penalty parameter λ_ℓ. In the case of ridge regression, Bardsley and Chambers [1] suggested the ridge trace method for selecting the penalty parameter λ. This method consists of plotting the weights $w_{iS}(\lambda), i \in S$ for values of λ from a pre-determined grid values and to choose the value of λ for which the weights $w_{iS}(\lambda)$ are positive for all

$i \in S$ and $\sum_{i\in S} w_{iS}(\lambda)\mathbf{x}_i - \sum_{i\in U} \mathbf{x}_i$ is the smallest difference among all the differences considered for the grid values of λ. Using the fact that the modified penalty $\lambda^* = \lambda/(1+\lambda)$ lies between 0 and 1 and is an increasing function of λ, Beaumont and Bocci [2] proposed a method based on the bisection algorithm to first determine λ^* and then, λ. Guggemos and Tillé [19] implemented a Fisher scoring algorithm in order to find the value of λ which maximizes a design-based estimated log-likelihood criterion. In case of the lasso model-assisted estimator, McConville *et al.* [28] used a cross-validation procedure to choose the best value of λ. More research is needed to suggest a unified criterion in order to find the best penalty in a sample-based framework. This is beyond the scope of the article. Most of the computer software use a cross-validation criterion to choose the best penalty parameter.

3.3. Consistency of the GREG and penalized GREG estimators in a high-dimensional setting

We adopt the asymptotic framework of [24] and consider an increasing sequence of embedded finite populations $\{U_v\}_{v\in\mathbb{N}}$ of size $\{N_v\}_{v\in\mathbb{N}}$. In each finite population U_v, a sample, of size n_v, is selected according to a sampling design $\mathcal{P}_v(S_v)$ with first-order inclusion probabilities $\pi_{i,v}$ and second-order inclusion probabilities $\pi_{i\ell,v}$. While the finite populations are considered to be embedded, we do not require this property to hold for the samples $\{S_v\}_{v\in\mathbb{N}}$. This asymptotic framework assumes that v goes to infinity, so that both the finite population sizes $\{N_v\}_{v\in\mathbb{N}}$, the samples sizes $\{n_v\}_{v\in\mathbb{N}}$ and the number of auxiliary variables $\{p_v\}_{v\in\mathbb{N}}$ go to infinity. To improve readability, we shall use the subscript v only in the quantities U_v, N_v, n_v and p_v; for instance, quantities such as $\pi_{i,v}$ shall be simply denoted by π_i.

The following assumptions are required to establish the consistency of the GREG and penalized GREG estimators in a high-dimensional setting.

(H1) We assume that there exists a positive constant C_1 such that $N_v^{-1}\sum_{i\in U_v} y_i^2 < C_1$.

(H2) We assume that $\lim_{v\to\infty} \dfrac{n_v}{N_v} = \pi \in (0,1)$.

(H3) There exist a positive constant c such that $\min_{i\in U_v} \pi_i \geqslant c > 0$; also, we assume that $\lim\sup_{v\to\infty} n_v \max_{i\neq\ell\in U_v} |\pi_{i\ell} - \pi_i\pi_\ell| < \infty$.

(H4) We assume that there exists a positive constant C_2 such that, for all $i \in U_v$, $||\mathbf{x}_i||_2^2 \leq C_2 p_v$, where $\|\cdot\|_2$ denotes the usual Euclidean norm.

(H5) We assume that $\|\widehat{\boldsymbol{\beta}}\|_1 = \mathcal{O}_{\mathrm{p}}(p_v)$, where $\widehat{\boldsymbol{\beta}}$ is the least square estimator given in (8) and $\|\cdot\|_1$ denotes the L^1 norm.

The assumptions (H1), (H2) and (H3) were used by Breidt and Opsomer [4] in a nonparametric setting and similar assumptions were used by Robinson and Särndal [34] to establish the consistency of the GREG estimator in a fixed dimensional setting. These assumptions hold for many usual sampling designs such as simple random sampling without replacement, stratified designs [4], or high-entropy sampling designs. Assumptions (H4) and (H5) can be viewed, respectively, as extensions of Assumption A.1 and Assumption A.3 in [34] to p_v-dimensional vectors with p_v growing to infinity. Assumption (H5) is not very restrictive in this high-dimensional setting as it requires that components of $\widehat{\boldsymbol{\beta}}$ are all bounded. When p_v is fixed, then our assumptions essentially reduce to those of [34].

Result 3.1: *Assume (H1)–(H5). Consider a sequence of GREG estimators $\{\widehat{t}_{greg}\}_{v\in\mathbb{N}}$ of t_y. Then,*

$$\frac{1}{N_v}(\widehat{t}_{\text{greg}} - t_y) = \mathcal{O}_{\text{p}}\left(\sqrt{\frac{p_v^3}{n_v}}\right).$$

If the numbers of auxiliary variables $\{p_v\}_{v\in\mathbb{N}}$ and the sample sizes $\{n_v\}_{v\in\mathbb{N}}$ satisfy $p_v^3/n_v = o(1)$, then $N_v^{-1}(\widehat{t}_{greg} - t_y) = o_{\text{p}}(1)$.

The $\sqrt{n}$-consistency obtained by Robinson and Särndal [34] is a special case of Result 3.1 with $p_v = \mathcal{O}(1)$. Result 3.1 highlights the fact that the rate of convergence decreases as the number of auxiliary variables p_v increases. Yet, this result guarantees the existence of a consistent GREG estimator, even when the number of auxiliary variables is allowed to diverge. An improved consistency rate for $\widehat{t}_{\text{greg}}$ may be obtained if, in (H5), the usual Euclidean norm is used instead of L^1-norm. Establishing the rate of convergence of the sampling error $\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$ may also be utilized to obtain a lower consistency rate for $\widehat{t}_{\text{greg}}$; e.g. see [10].

The next result establishes the design-consistency of model-assisted penalized regression estimators. The proof is similar to that of Result 3.1 and is given in the Supplementary Material.

Result 3.2: *Assume (H1)–(H4). Consider a sequence of penalized model-assisted estimators $\{\widehat{t}_{\text{pen}}\}_{v\in\mathbb{N}}$ of t_y obtained by either ridge, lasso or elastic-net. Then,*

$$\frac{1}{N_v}(\widehat{t}_{\text{pen}} - t_y) = \mathcal{O}_{\text{p}}\left(\sqrt{\frac{p_v^3}{n_v}}\right).$$

If the numbers of auxiliary variables $\{p_v\}_{v\in\mathbb{N}}$ and the sample sizes $\{n_v\}_{v\in\mathbb{N}}$ satisfy $p_v^3/n_v = o(1)$, then $N_v^{-1}(\widehat{t}_{pen} - t_y) = o_{\text{p}}(1)$.

The above result makes no use of the asymptotic convergence rate of $\widehat{\boldsymbol{\beta}}_{\text{pen}}$ which depends on the penalization method. For example, if one can establish that $||\widehat{\boldsymbol{\beta}}_{\text{pen}}||_1 = \mathcal{O}_{\text{p}}(\gamma_v)$, then $N_v^{-1}(\widehat{t}_{\text{pen}} - t_y) = \mathcal{O}_{\text{p}}(\gamma_v\sqrt{p_v/n_v})$. Alternatively, improved consistency rates of $\widehat{t}_{\text{pen}}$ may be obtained if one can establish the magnitude of the sampling error $\widehat{\boldsymbol{\beta}}_{\text{pen}} - \widetilde{\boldsymbol{\beta}}_{\text{pen}}$ in a high-dimension setting. In other words, obtaining these improved rates requires additional assumptions, unlike Result 3.2 which is obtained under relatively mild assumptions.

Next, we show that, under additional assumptions on the auxiliary variables, the model-assisted ridge estimator is L^1 design-consistent for t_y if p_v/n goes to zero and that it has the usual $\sqrt{n}$-consistency rate if $p_v = O(n_v^a)$ with $0 \leq a < 1/2$, which constitutes a significant improvement over Result 3.2.

Result 3.3: *Assume (H1)–(H4). Also, assume that there exists a positive constant $\tilde{C}$ such that $\lambda_{max}(\boldsymbol{X}_{U_v}^\top \boldsymbol{X}_{U_v}) \leqslant \tilde{C}N_v$, where $\lambda_{max}(\boldsymbol{X}_{U_v}^\top \boldsymbol{X}_{U_v})$ is the largest eigenvalue of $\boldsymbol{X}_{U_v}^\top \boldsymbol{X}_{U_v}$. Assume also that $N_v/\lambda_v = \mathcal{O}(1)$.*

(1) *Then, there exists a positive constant C such that $\mathbb{E}_p[||\widehat{\boldsymbol{\beta}}_{\text{ridge}}||_2^2] \leqslant C$ and*

$$\frac{1}{N_v}\mathbb{E}_p\left|\widehat{t}_{\text{ridge}} - t_y\right| = \mathcal{O}\left(\sqrt{\frac{p_v}{n_v}}\right).$$

If the numbers of auxiliary variables $\{p_v\}_{v\in\mathbb{N}}$ and the sample sizes $\{n_v\}_{v\in\mathbb{N}}$ satisfy $p_v/n_v = o(1)$, then $N_v^{-1}\mathbb{E}_p|\widehat{t}_{\text{ridge}} - t_y| = o(1)$.

(2) $\mathbb{E}_{\text{p}}(||\widehat{\boldsymbol{\beta}}_{\text{ridge}} - \tilde{\boldsymbol{\beta}}_{\text{ridge}}||_2^2) = \mathcal{O}(p_v/n_v)$. *Thus, if $p_v/n_v = o(1)$, then* $\mathbb{E}_{\text{p}}(||\widehat{\boldsymbol{\beta}}_{\text{ridge}} - \tilde{\boldsymbol{\beta}}_{\text{ridge}}||_2^2) = o(1)$.

(3) *We have the following asymptotic equivalence:*

$$\frac{1}{N_v}\left(\widehat{t}_{\text{ridge}} - t_y\right) = \frac{1}{N_v}\left(\widehat{t}_{\text{diff},\lambda} - t_y\right) + \mathcal{O}_{\text{p}}\left(\frac{p_v}{n_v}\right),$$

where

$$\widehat{t}_{\text{diff},\lambda} = \sum_{i\in S_v} y_i/\pi_i - \left(\sum_{i\in S_v}\mathbf{x}_i/\pi_i - \sum_{i\in U_v}\mathbf{x}_i\right)^\top \tilde{\boldsymbol{\beta}}_{\text{ridge}}$$

and

$$\frac{1}{N_v}\mathbb{E}_p\left|\widehat{t}_{\text{ridge}} - t_y\right| = \mathcal{O}\left(\frac{1}{\sqrt{n_v}}\right) + \mathcal{O}\left(\frac{p_v}{n_v}\right).$$

If $p_v = \mathcal{O}(n_v^a)$ with $0 \leq a < 1/2$, then

$$\frac{1}{N_v}\left(\widehat{t}_{\text{ridge}} - t_y\right) = \frac{1}{N_v}\left(\widehat{t}_{\text{diff},\lambda} - t_y\right) + o_{\text{p}}(1)$$

and

$$\frac{1}{N_v}\mathbb{E}_p\left|\widehat{t}_{\text{ridge}} - t_y\right| = \mathcal{O}\left(\frac{1}{\sqrt{n_v}}\right).$$

It follows from Result 3.3 that, for $p_v = \mathcal{O}(n_v^a)$ with $0 \leq a < 1/2$, the asymptotic variance of the model-assisted ridge estimator $\widehat{t}_{\text{ridge}}$ is equal to the variance of the generalized difference estimator $\widehat{t}_{\text{diff},\lambda}$. For $a = 1/2$, we note that the model-assisted estimator is still $\sqrt{n}$-design consistent but the remainder term is no longer negligible with respect to $\widehat{t}_{\text{diff},\lambda}$ and the variability of this term should be consider to compute the asymptotic variance of $\widehat{t}_{\text{ridge}}$. The case of model-assisted estimators based on lasso and elastic-net is more intricate. This is due to the fact that both estimators involve the L^1-norm. As a result, a closed-form expression of these estimators cannot be obtained. However, if the predictors are orthogonal, a closed-form expression exists for the lasso and elastic-net estimators and improved consistency rates can be obtained, see Proposition 3.1 below. The case of non-orthogonal predictors is more challenging and is beyond the scope of this article.

Proposition 3.1: *Suppose assumptions (H1)–(H3) and that the sampling design and the X-variables are such that the columns of $\mathbf{\Pi}_{S_v}^{-1/2}\mathbf{X}_{S_v}$ are orthogonal. Suppose also that there exist positive quantities C_3 and C_4 such that $max_{j=1,\dots,p_v} N_v^{-1}\sum_{i\in U_v} x_{ij}^4 \leq C_3 <$*

∞ and $\min_{j=1,\ldots,p_v} N_v^{-1} \sum_{i\in U_v} x_{ij}^2 \geq C_4 > 0$. Then, $N_v^{-1}(\widehat{t}_{\text{greg}} - t_y) = \mathcal{O}_{\text{p}}(\sqrt{p_v/n_v})$ and $N_v^{-1}(\widehat{t}_{\text{pen}} - t_y) = \mathcal{O}_{\text{p}}(\sqrt{p_v/n_v})$, where $\widehat{t}_{\text{pen}}$ denotes either the lasso or the elastic-net estimator.

4. Simulation study

In this section, we provide an empirical comparison of several model-assisted estimators, in addition to the estimators discussed in Section 3. In addition, we considered model-assisted estimators based on principal component regression [8], regression trees [5], random forests [6], k-nearest neighbours, XGBoost [11] and Cubist [32]. For a description of these methods, see [13,20] and the references therein.

We used data from the Irish Commission for Energy Regulation (CER) Smart Metering Project that was conducted in 2009–2010 (CER, 2011)[1] [8]. This project focused on energy consumption and energy regulation. About 6000 smart meters were installed to collect the electricity consumption of Irish residential and business customers every half an hour over a period of about 2 years.

We considered a period of 14 consecutive days and a population of $N = 6291$ smart meters (households and companies). Each day consisted of 48 measurements, leading to 672 measurements for each household. We denote by $X_j = X(t_j), j = 1, \ldots, 672$, the electricity consumption (in kW) at instant t_j and by x_{ij} the value of X_j recorded by the ith smart meter for $i = 1, \ldots, 6,291$. It should be noted that the matrix $N^{-1}\mathbf{X}^\top\mathbf{X}$ was ill-conditioned with a condition number equal to 254 753. This suggests that some of the X-variables were highly correlated with each other.

We generated four survey variables based on these auxiliary variables according to the following models:

$$\begin{aligned}
Y_1 &= 400 + 2X_1 + X_2 + 2X_3 + \mathcal{N}(0, 1500);\\
Y_2 &= 500 + 2X_4 + 400\mathbb{1}\,(X_5 > 156) - 400\mathbb{1}\,(X_5 \leqslant 156) + 1000\mathbb{1}\,(X_2 > 190)\\
&\quad + 300\mathbb{1}\,(X_5 > 200) + \mathcal{N}(0, 1500);\\
Y_3 &= 1 + \cos(2X_1 + X_2 + 2X_3)^2 + \epsilon_1;\\
Y_4 &= 4 + 3 \cdot \mathbb{V}\left(\{X_1 + X_2\}^2\right)^{-1/2} \times \{X_1 + X_2\}^2 + \mathcal{N}(0, 0.01),
\end{aligned}$$

where $\mathbb{V}(\cdot)$ denotes the empirical variance and the errors ϵ_1 in the model for Y_3 were generated from an $\mathcal{E}xp(10)$ and these errors were centred so as to obtain a mean equal to zero.

Our goal was to estimate the population totals $t_{y_j} = \sum_{i\in U} y_{ij}, j = 1, \ldots, 4$. From the population, we selected $R = 2500$ samples, of size $n = 600$, which corresponds to a sampling fraction n/N of about 10%. We considered three sampling schemes: simple random sampling without replacement, stratified simple random sampling without replacement with optimal allocation and stratified without replacement proportional to size sampling with proportional allocation.

In each sample, we computed 12 model-assisted estimators of the form

$$\widehat{t}_{ma}^{(j)} = \sum_{i\in U} \widehat{f}^{(j)}(\mathbf{x}_i) + \sum_{i\in S} \frac{y_i - \widehat{f}^{(j)}(\mathbf{x}_i)}{\pi_i}, \quad j = 1, 2, \ldots, 12,$$

where the predictors $\widehat{f}^{(j)}(\mathbf{x}_i)$, $j = 1, 2, \ldots, 12$, were obtained using the following procedures:

Procedure 1: 'LR' : Deterministic linear regression, leading to the GREG estimator.

Procedure 2: 'CART': Classification and regression tree algorithm [5], leading to an estimator closely related to that of [29] and implemented with the *R*-package `rpart`.

Procedure 3: 'RF': Random forests with the algorithm of [6] with $B = 1000$ trees, a minimal number of elements in each terminal node $n_0 = 5$ and $p_0 = \lfloor\sqrt{p}\rfloor$ variables selected randomly at each split, where $\lfloor\cdot\rfloor$ denotes the customary floor function. The algorithm leads to the estimator described in [12]. Simulations were implemented with the *R*-package `ranger`.

Procedure 4: 'Ridge': Ridge regression with a regularization parameter determined by cross-validation and implemented with the *R*-package `glmnet`. The estimator was studied by [18].

Procedure 5: 'Lasso': Lasso regression with a regularization parameter determined by cross-validation and implemented with the *R*-package `glmnet` [28].

Procedure 6: 'EN': Elastic net regression with penalization coefficients determined by cross-validation with the *R*-package `glmnet`.

Procedure 7: 'XGB': XGBoost algorithm [20] with 50 trees in the additive model, each tree being with a depth of at most 6 and a learning rate $\lambda = 0.01$. Simulations were implemented with the *R*-package `XGBoost`.

Procedure 8: '5NN': 5-nearest neighbours predictor with the Euclidean distance and implemented with the *R*-package `caret`.

Procedure 9: 'Cubist': A cubist algorithm [25] with 5 models in each predictor, implemented with the *R*-package `cubist`; the algorithm and its adaptation for survey data are described in [13].

Procedure 10: 'PCR1': Principal component regression based on the first $\lfloor p^{1/4}\rfloor$ components kept and implemented with the *R*-package `pls` [8].

Procedure 11: 'PCR2': Principal component regression based on the first $\lfloor p^{2/4}\rfloor$ components kept.

Procedure 12: 'PCR3': Principal component regression based on the first $\lfloor p^{3/4}\rfloor$ components kept.

As a measure of bias of the model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, 2, \ldots, 12$, we computed the Monte Carlo percent relative bias defined as

$$RB_{MC}\left(\widehat{t}_{ma}^{(j)}\right) = 100 \times \frac{1}{R}\sum_{r=1}^{R}\frac{(\widehat{t}_{ma}^{(j,r)} - t_y)}{t_y}, \quad j = 1, 2, \ldots, 12,$$

where $\widehat{t}_{ma}^{(j,r)}$ denotes the estimator $\widehat{t}_{ma}^{(j)}$ at the rth iteration, $r = 1, \ldots, R$. As a measure of efficiency, we computed the relative of efficiency, using the Horvitz–Thompson estimator $\widehat{t}_\pi$ given by (1), as the reference. That is,

$$RE_{MC}\left(\widehat{t}_{ma}^{(j)}\right) = 100 \times \frac{MSE_{MC}(\widehat{t}_{ma}^{(j)})}{MSE_{MC}(\widehat{t}_\pi)}, \quad j = 1, 2, \ldots, 12,$$

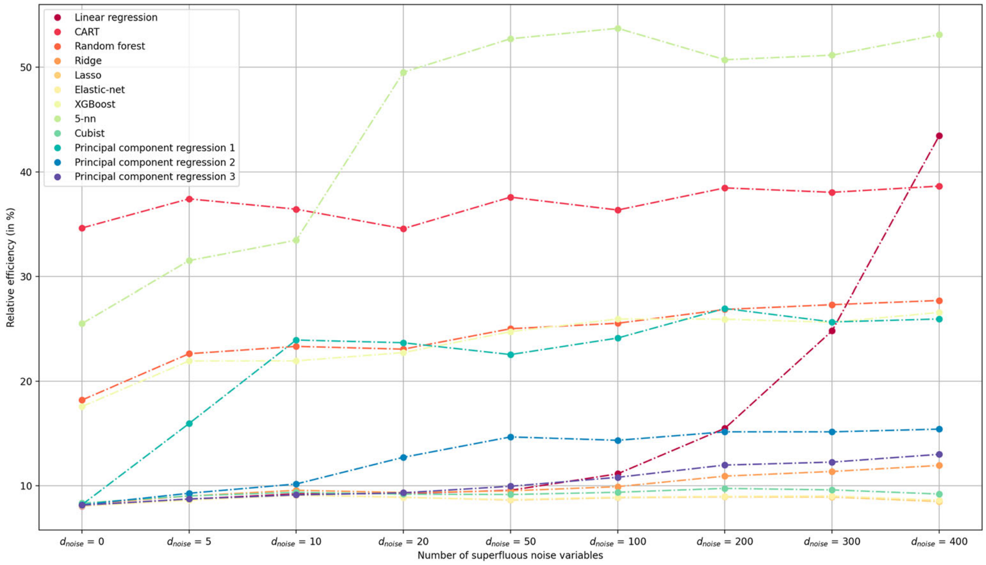

Figure 1. Relative efficiency of model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ for the estimation of the total of Y_1 with SRSWOR ($n = 600$) and increasing number of auxiliary variables.

where $MSE_{MC}(\widehat{t}_{ma}^{(j)}) = R^{-1}\sum_{r=1}^{R}(\widehat{t}_{ma}^{(j,r)} - t_y)^2$ and $MSE_{MC}(\widehat{t}_\pi)$ is defined similarly.

We were also interested in investigating to which extent the model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ were affected by the inclusion of a large number of predictors in the working models. To that end, in addition to the variables $X_1, \ldots, X_5$, we included d_{noise} predictors in the working models. These predictors were available in the Irish data set. We used the following values for d_{noise}: 5, 10, 20, 50, 100, 200, 300 and 400.

4.1. Simple random sampling without replacement

In this section, we present the results obtained under simple random sampling without replacement (SRSWOR) of size $n = 600$. All the point estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$, exhibited a negligible or small percent RB with a maximum value of about 3.1% (obtained in the case of the GREG estimator). For this reason, results pertaining to relative bias are not reported here.

Figures 1–4 display the relative efficiency of the model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ as a function of the number of auxiliary variables incorporated in the working models. To improve readability, we have truncated some large values of RE, when applicable.

We begin by discussing the results on relative efficiency pertaining to the estimation of the total of the survey variable Y_1. For low-dimensional settings, the GREG estimator was very efficient with values of RE below 10%. These results can be explained by the fact that Y_1 was linearly related to the $\mathbf{x}$-variables. However, as the number of variables d_{noise} increased, the efficiency of the GREG estimator rapidly deteriorated, suggesting that the performance of the GREG estimator is sensitive to the dimension of the $\mathbf{x}$-vector. As

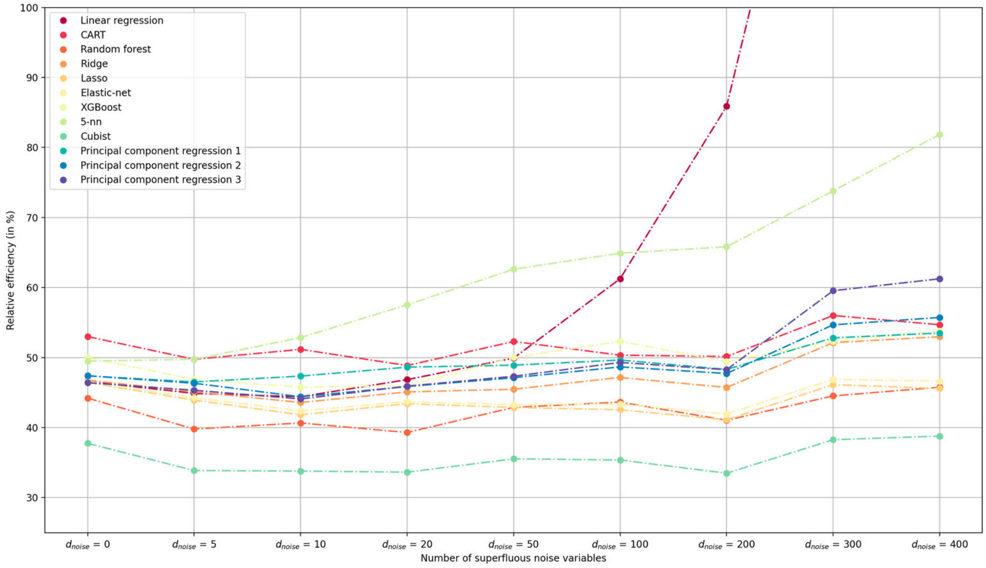

Figure 2. Relative efficiency of model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ for the estimation of the total of Y_2 with SRSWOR, $n = 600$ and increasing number of auxiliary variables.

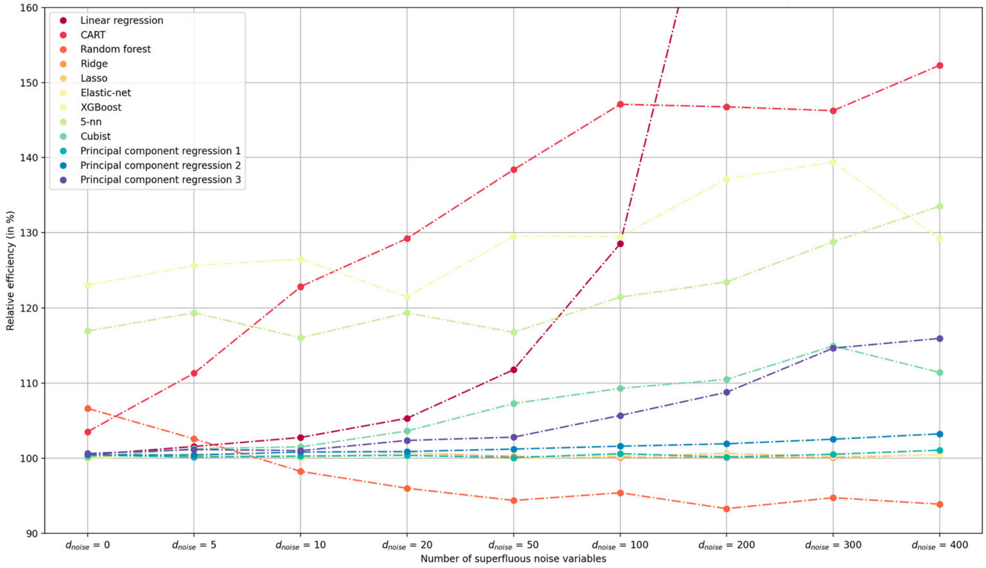

Figure 3. Relative efficiency of model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ for the estimation of the total of Y_3 with SRSWOR, $n = 600$ and increasing number of auxiliary variables.

expected, model-assisted estimators based on regularization methods such as ridge, lasso, elastic-net or dimension reduction methods such as principal components regression, performed generally very well. Unlike the GREG, these estimators were not much affected by the number of auxiliary variables incorporated in the model. Turning to the model-assisted estimator based on a 5-nn, we note that it was less efficient than most competitors and that

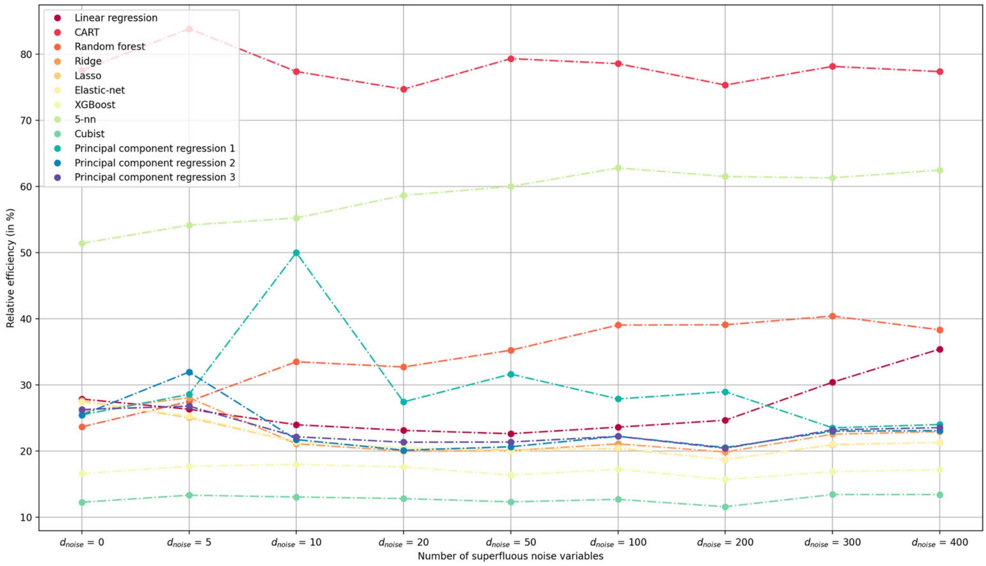

Figure 4. Relative efficiency of model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ for the estimation of the total of Y_4 with SRSWOR, $n = 600$ and increasing number of auxiliary variables.

its efficiency got worse as d_{noise} increased, a phenomenon referred to as the curse of dimensionality. The model estimators based on XGBoost, Cubist and random forests performed quite well and did not seem to be affected by the number of auxiliary variables incorporated in the model. Finally, the estimators based on CART were less efficient than those obtained through the other machine learning methods.

The results pertaining to the survey variable Y_2 and displayed in Figure 2 were fairly consistent with those obtained for the survey variable Y_1 with one exception: the Cubist algorithm was significantly more efficient than the other procedures in all the scenarios.

Turning to the survey variable Y_3 (see Figure 3), the model-assisted estimator based on random forests was significantly more efficient than the Horvitz–Thompson estimator, especially for large values of d_{noise}. The other procedures led to estimators less efficient than the Horvitz–Thompson estimator with values of RE above 100. In particular, the GREG estimator broke down as the number of auxiliary variable increased. The performance of model-assisted estimators based on CART and XGBoost algorithms deteriorated as the dimension increased. In a high-dimension setting with highly correlated predictors, random forests improved over CART due to the random subsampling of p_0 variables among the p variables, generating then decorrelated trees [20].

The results in Figure 4 about the survey variable Y_4 were similar to the ones in previous figures. Most estimators remained mostly unaffected by the number of auxiliary variables d_{noise}. Again, the model-assisted estimator based on the Cubist algorithm was the best in all the scenarios.

4.2. Stratified simple random sampling with optimal allocation

In the second simulation study, we partitioned the Irish residential and business customer population into four strata $U_1, \ldots, U_4$, using an equal quantile method with respect to the

Table 1. First-order inclusion probabilities and sampling weights within strata.

Stratum	1	2	3	4
π_i	0.012	0.022	0.028	0.316
$w_i = \pi_i^{-1}$	77.85	43.83	35.11	3.16

variable, X_1, the electricity consumption at instant t_1. From the population, we selected $R = 2500$ stratified simple random samples, of size $n = 600$. The stratum sample sizes n_h were determined using an X_2-optimal allocation, where X_2 denotes the electricity consumption recorded at instant t_2. This led to $n_1 = 20, n_2 = 36, n_3 = 45$ and $n_4 = 499$. The first-order inclusion probabilities, $\pi_i = n_h/N_h, i \in U_h$ and the sampling weights $w_i = \pi_i^{-1}$ are shown in Table 1.

We confined to the survey variables Y_1 and Y_3 only and we aimed at estimating t_{y_1} and t_{y_3}. It is worth pointing out that the resulting sampling design was informative as the variables used at the design stage (X_1 and X_2) were also related to the survey variables Y_1 and Y_3. In fact, the Monte Carlo coefficient of correlation between the sampling weights and Y_1 was approximately equal to 0.402. We do not report the coefficient of correlation between the sampling weights and Y_3 as the relationship between Y_3 and the set of predictors X_1, X_3 is not linear.

Again, in each sample we computed twelve model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ for each of t_{y_1} and t_{y_3}. Since most machine learning software packages do not take the sampling weights into account, we have included the design variables X_1 and X_2 in the set of predictors.

We begin by discussing the results pertaining to the estimation of the total of the survey variable Y_1. Figures 5 and 6 display the Monte Carlo percent relative bias and the Monte Carlo relative efficiency as a function of the number of variables d_{noise}. Except for the model-assisted estimators based on 5-nn and random forest, the other estimators exhibit a small value of RB for all values of d_{noise}. Again, the 5-nn model-assisted estimator suffered from the curse of dimensionality. Turning to the estimator based on random forests, we note from Figure 5 that the bias increased as the number of predictors d_{noise} increased. For instance, for $d_{noise} = 400$, the value of RB was just above 10%. This significant bias may be explained by the fact that random forests is the only procedure among the ones considered in our simulation that randomly selects $p_0 = \sqrt{p}$ variables among the initial p predictors at each split. For instance, for $d_{noise} = 400$, only 20 variables are randomly selected at each split. As a result, most predictions obtained through a random forests algorithm were based on misspecified working models, leading to potentially bad fits and large residuals. Also, each prediction corresponds to a weighted mean computed within each node with $n_0 = 5$ observations only. Therefore, each predictions corresponds to a ratio-type estimate based on five observations only. This, together with the fact that the sampling weights are highly variable, constitutes a conducive ground for the occurrence of small sample bias. In terms of efficiency, except for the GREG, the 5-nn and the random forest estimators, the other procedures performed well with values of RE ranging from 60% to 80%. The best procedures were Cubist and Lasso.

We now turn to the survey variable Y_3. First, the Monte Carlo relative bias was negligible for all the estimation procedures and are not reported here. Results about relative efficiency

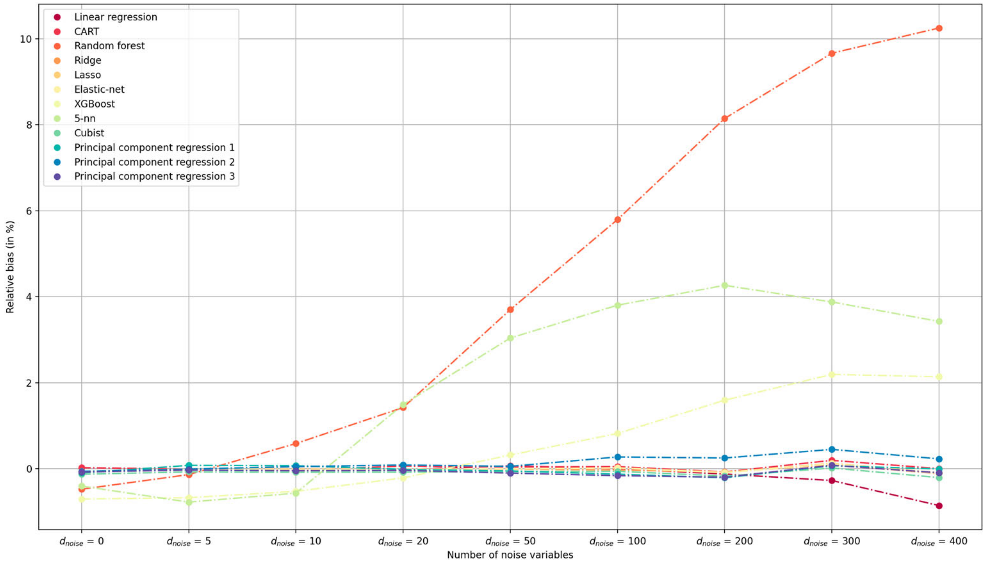

Figure 5. Relative bias of model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ for the estimation of the total of Y_1 with stratified simple random sampling with X_2-optimal allocation, $n = 600$ with increasing number of auxiliary variables.

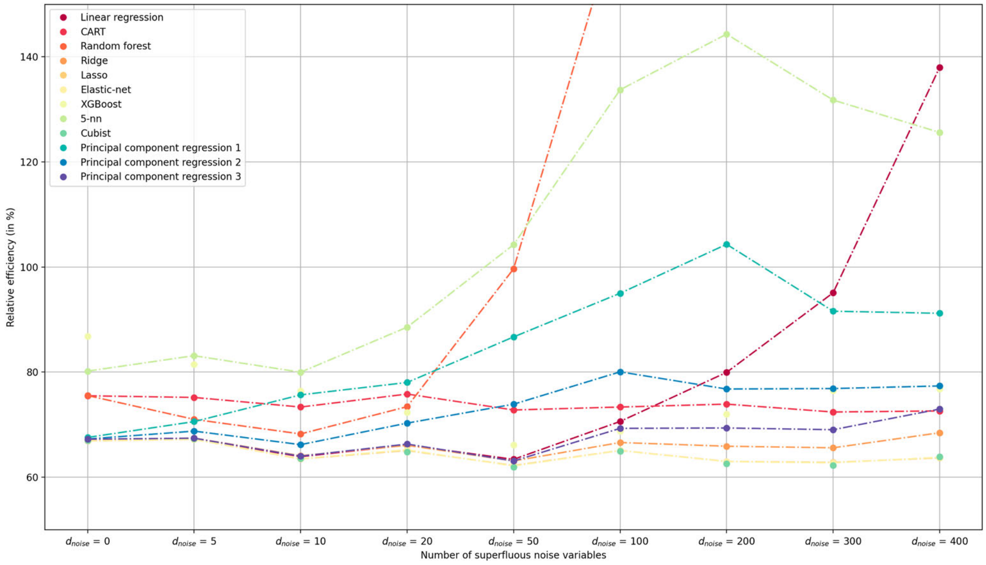

Figure 6. Relative efficiency of model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ for the estimation of the total of Y_1 with stratified simple random sampling with X_2-optimal allocation, $n = 600$ and increasing number of auxiliary variables.

are plotted in Figure 7. Random forests performed extremely well and their performance improved as d_{noise} increased. This suggests that the method was able to extract the information contained in the predictors. This was also true for Cubist and XGBoost, although to a lesser extent.

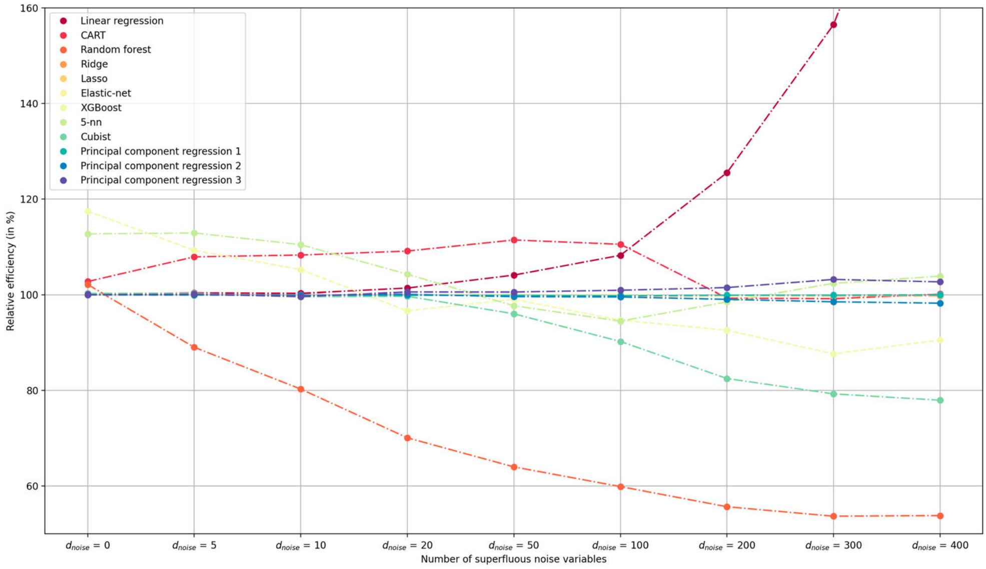

Figure 7. Relative efficiency of model-assisted estimators $\widehat{t}^{(j)}_{ma}, j = 1, \ldots, 12$ for the estimation of the total of Y_3 with stratified simple random sampling with X_2-optimal allocation, $n = 600$ and increasing number of auxiliary variables.

To get a better understanding of the performance of random forests for the estimation of the total of the survey variable Y_1, we conducted additional scenarios based on different values of the hyper parameters n_0, the number of observations within each terminal nodes, and p_0, the number of variables randomly selected at each split among the initial p model variables. We used the following values for n_0 and p_0:

- $n_0 = 5$ observations and $p_0 = \sqrt{p}$ variables which are the default choices in the *R*-package `ranger`;
- $n_0 = 5$ observations and $p_0 = p$ variables;
- $n_0 = 5$ observations and $p_0 = \sqrt{p}$ variables, with, in addition, the design variables X_1, X_2, as well as the vector of inclusion probabilities and the vector of strata that were selected with probability 1, at each split, besides the p_0 variables;
- $n_0 = n^{13/20}$ observations and $p_0 = \sqrt{p}$ variables.

The Monte Carlo percent relative bias is displayed in Figure 8. We note that relative bias was much smaller (always less than 1%) when the design variables were considered besides p_0 variables at each split. To a lesser extent, the bias decreased when more observations were allowed in each terminal node. These results suggest, that, when the sampling design is informative, to avoid significant small sample bias, we recommend to force the design variables to be selected at each split. This option is available in the *R* package `ranger`.

4.3. *Stratified inclusion probability proportional-to-size sampling without replacement*

We consider the stratified population described in Section 4.2. In each stratum, we selected units according to a fixed-size inclusion probability proportional-to-size sampling without

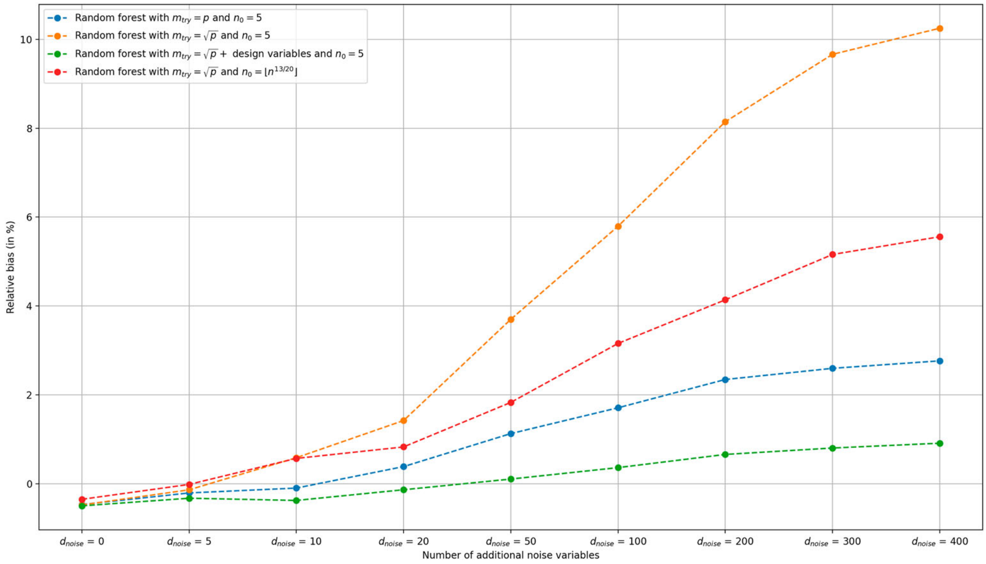

Figure 8. Comparison of different configurations of hyper-parameters for $\widehat{t}_{rf}$ for the estimation of the total of Y_1 with stratified simple random sampling and X_2-optimal allocation, $n = 600$.

replacement using X_2, the electricity consumption at instant $t = 2$, as the size variable. In each stratum, we used the sample size n_h were determined according to proportional allocation; i.e. $n_h = n \cdot N_h/N$. The first-order inclusion probabilities were then given by

$$\pi_i = \frac{n_h x_{i2}}{\sum_{j \in U_h} x_{j2}}, \quad i \in U_h, \quad \text{and} \quad h = 1, 2, 3, 4.$$

As in Section 4.2, we focused on estimating t_{y_1} and t_{y_3} and we computed the same twelve model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$. The inclusion probabilities were highly correlated with the survey variable Y_1, with a correlation coefficient of about 0.62; we do not report the coefficient of correlation in the case of Y_3 as the underlying relationship was non-linear. Based on findings from Section 4.2, we adopted the following configuration for the random forest algorithm: we considered $n_0 = 5$ observations in each terminal node and, at each split, we randomly selected $p_0 = \sqrt{p}$ variables. Note that the design variables X_1 and X_2 as well as the vector of inclusion probabilities and the vector of stratum indicators were selected with probability 1 at each split in addition to the p_0 variables.

All the estimators exhibited a negligible relative bias (less than 1%). Figures 9 and 10 show the relative efficiency corresponding to t_{y1} t_{y3}, respectively.

From Figure 9, we note that most estimators exhibited a behaviour similar to that obtained in the case the stratified simple random sampling based on an X_2-optimal allocation (see Section 4.2). However, we note that the estimators PCR1 and PCR2 did poorly unlike in the case stratified simple random sampling based on an X_2-optimal allocation. This poor behaviour may be due to the fact that the sampling design was now much more informative and keeping a few principal components only may have led to a loss of information. The estimator PCR3 based on more principal components did better than PCR1

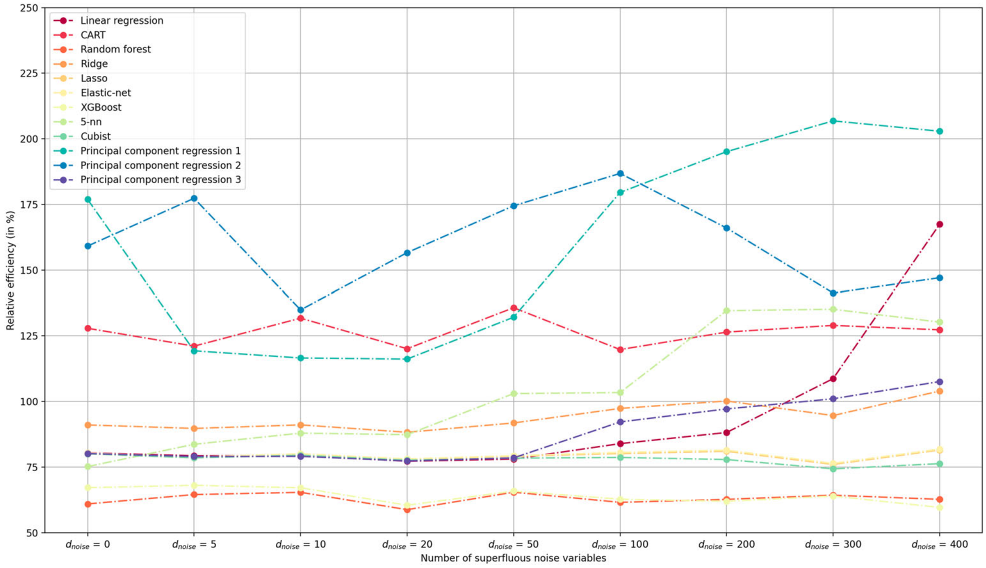

Figure 9. Relative efficiency of model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ for the estimation of the total of Y_1 with stratified without replacement X_2-proportional to size sampling, $n = 600$ and increasing number of auxiliary variables.

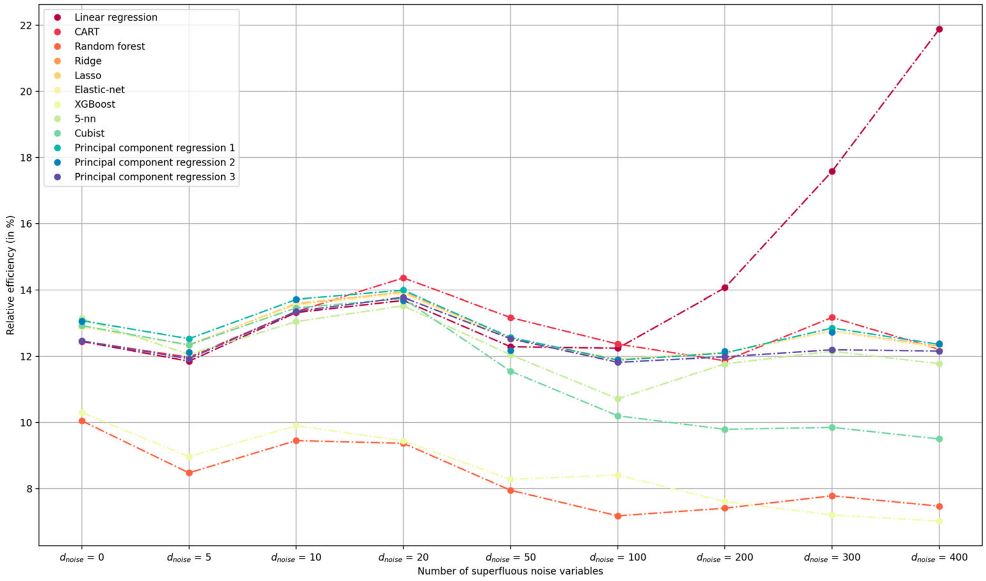

Figure 10. Relative efficiency of model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ for the estimation of the total of Y_3 with stratified without replacement X_2-proportional to size sampling, $n = 600$ and increasing number of auxiliary variables.

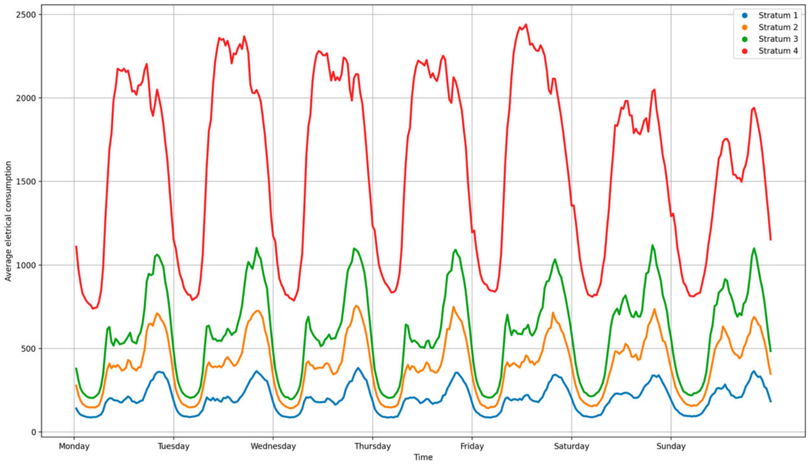

Figure 11. Average electricity consumption on each stratum during first week.

and PCR2. From Figure 10, we note that the use of model-assisted estimators led to significant improvement over the Horvitz–Thompson estimator, with value of relative efficiency ranging from 6% to 22%.

4.4. Stratified simple random sampling with proportional allocation

In this section, we consider a more realistic scenario based again on the Irish residential and business customer data. As a stratification variable, we used the mean electricity consumption recorded during the first week. Again, we constructed four strata using an equal-quantile method based, this time, on the mean electricity consumption; see also [7] who used a similar design. The mean trajectories during the first week within each stratum are plotted in Figure 11. From Figure 11, we note that Stratum 1 corresponds to consumers with low global levels of electricity consumption, whereas Stratum 4 consists of consumers who have high levels of electricity consumption.

Our aim was to estimate the total electricity consumption recorded on the Monday of the second week and given by $t_y = \sum_{i=1}^{6291} \sum_{j=336}^{384} y_{ij}$, where y_{ij} is the electricity consumption recorded for the i th unit at the jth instant. Within each stratum, we selected a sample, of size n_h, according to simple random sampling without replacement. The n_h's were determined according to proportional allocation; i.e. $n_h = n \times (N_h/N)$ with $n = 600$. In each of the 2500 samples, we computed the same 12 model-assisted estimators as in the previous sections. Again, we computed the Monte Carlo percent relative bias and the relative efficiency for each the 12 estimators. The results are presented in Table 2.

From Table 2, we note that the 5-nn model-assisted estimator was the only estimator to exhibit a non-negligible bias. Although it was less efficient than its competitors, it was more efficient than the Horvitz–Thompson estimator with a value of RE of about 65.6%. The ridge estimator was the most efficient with a value of RE equal to 4% and was closely followed by lasso, elastic-net, Cubist and principal components model-assisted estimators. The GREG estimator performed very well with a value of RE of about 9.3%. Random forests led to considerable improvement over the CART model-assisted estimator with values of

Table 2. Monte Carlo percent relative bias and relative efficiency of several model-assisted estimators under stratified simple random sampling with proportional allocation.

Estimator	Relative bias	Relative efficiency
LR	0.2	9.3
CART	−0.1	41.0
RF	−1.1	17.0
Ridge	0.1	4.0
Lasso	0.2	4.1
EN	0.2	4.1
XGB	−1.7	24.9
NN5	−4.0	65.6
Cubist	−0.0	4.3
PCR1	0.1	4.9
PCR2	0.1	4.2
PCR3	0.1	4.2

RE of 17% and 41%, respectively. Still, random forests were less efficient than the GREG estimator, which is not surprising as the relationship between the survey variable and the auxiliary variables was linear.

5. Final remarks

In this paper, we have examined a number of model-assisted estimation procedures in a high-dimensional setting both theoretically and empirically. If the relationship between the survey variable and the auxiliary information can be well described by a linear model, our results suggest that penalized estimators such as ridge, lasso and elastic net perform very well in terms of bias and efficiency, even in the case $p = n$. Model-assisted estimators based on random forests, Cubist and XGBoost methods were mostly unaffected by the number of predictors incorporated in the working model, even in the case of complex relationships between the study and the auxiliary variables. As expected, the GREG estimator suffered from poor performances in the case of a large number of auxiliary variables.

The procedure Cubist stood out from the other machine learning procedure with very good performances in virtually all the scenarios. Further work is needed to establish the theoretical properties of model-assisted estimators based on Cubist in both low-dimensional and high-dimensional settings.

Variance estimation is an important stage of the estimation process. Further research includes identifying the regularity conditions under which the variance estimators are design consistent in a high-dimensional setting.

We end this article by mentioning that virtually all the machine learning software packages cannot handle design features such as unequal weights and stratification. For instance, some random forests algorithms may involve a bootstrapping procedure and/or a cross-validation procedure. To fully account for the sampling design, both procedures must be modified so as to account for the design features. One notable exception is the R package RPMS [40] that has the ability to incorporate sampling weights for CART and random forests. Not fully accounting for the sampling design may be viewed as a form of model misspecification. However, model-assisted estimation procedures remain design consistent even if the model is misspecified. In our experiments, several machine learning procedures (e.g. random forests, Cubist, XGboost) performed very well in most scenarios even though we did not modify the bootstrapping and cross-validation procedures to account for design features. In other words, it seems that, accounting for predictors that are highly predictive of the Y-variable, seems to be the preponderant factor with respect to the efficiency aspect of model-assisted estimators. We conjecture that fully accounting for the sampling design will likely lead to additional efficiency gains but that the predictive power of the model likely constitutes the 'determining factor'. Developing machine learning procedures that fully account for the sampling design is currently under investigation.

Note

[1] The data are available on request at: `https://www.ucd.ie/issda/data/commissionforenergyregulationcer/`.

Acknowledgments

We thank the Editor, an Associate Editor, and two referees for their comments and suggestions, which helped improving the paper substantially. We are very grateful to Professor Patrick Tardivel from the Université de Bourgogne for enlightening discussions about the lasso method.

Disclosure statement

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

Funding

The work of Mehdi Dagdoug was supported by grants of the Franche-Comté region and Médiamétrie. The work of David Haziza was supported by a grant of the Natural Sciences and Engineering Research Council of Canada.

References

[1] P. Bardsley and R. Chambers, *Multipurpose estimation from unbalanced samples*, Appl. Stat. 33 (1984), pp. 290–299.

[2] J.-F. Beaumont and C. Bocci, *Another look at ridge calibration*, Metron-Int. J. Statist. 66 (2008), pp. 260–262.

[3] F. Breidt, G. Claeskens, and J. Opsomer, *Model-assisted estimation for complex surveys using penalized splines*, Biometrika 92 (2005), pp. 831–846.

[4] F.-J. Breidt and J.-D. Opsomer, *Local polynomial regression estimators in survey sampling*, Ann. Statist. 28 (2000), pp. 1023–1053.

[5] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees*, Routledge, New York, 1984.

[6] L. Breiman, *Random forests*, Mach. Learn. 45 (2001), pp. 5–32.

[7] H. Cardot, A. Dessertaine, C. Goga, E. Josserand, and P. Lardin, *Comparison of different sample designs and construction of confidence bands to estimate the mean of functional data : an illustration on electricity consumption*, Surv. Methodol. 39 (2013), pp. 283–301.

[8] H. Cardot, C. Goga, and M.-A. Shehzad, *Calibration and partial calibration on principal components when the number of auxiliary variables is large*, Statist. Sin. 27 (2017), pp. 243–260.

[9] R. Chambers, *Robust case-weighting for multipurpose establishment surveys*, J. Off. Stat. 12 (1996), pp. 3–32.

[10] G. Chauvet and C. Goga, *Asymptotic efficiency of the calibration estimator in a high-dimensional data setting*, J. Statist. Plann. Inference 217 (2021), pp. 177–187.

[11] T. Chen and C. Guestrin, *XGBoost*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD 16*, 2016, ACM Press, New York.

[12] M. Dagdoug, C. Goga, and D. Haziza, *Model-assisted estimation through random forests in finite population sampling*, preprint (2022), to appear in J. Am. Stat. Assoc.

[13] M. Dagdoug, C. Goga, and D. Haziza, *Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison*, to appear in J. Survey Statist. Methodol. (2021).

[14] J.-C. Deville and C.-E. Särndal, *Calibration estimators in survey sampling*, J. Am. Stat. Assoc. 87 (1992), pp. 376–382.

[15] D. Firth and K. Bennett, *Robust models in probability sampling*, J. R. Statist. Soc. Ser. B 60 (1998), pp. 3–21.

[16] C. Goga, *Réduction de la variance dans les sondages en présence d'information auxiliaire: une approche non paramétrique par splines de régression*, Canad. J. Statist. 33 (2005), pp. 163–180.

[17] C. Goga and A. Ruiz-Gazen, *Efficient estimation of non-linear finite population parameters by using non-parametrics*, J. R. Statist. Soc. Ser. B 76 (2014), pp. 113–140.

[18] C. Goga and M.A. Shehzad, *Overview of ridge regression estimators in survey sampling*, Université de Bourgogne: Dijon, France, 2010.

[19] F. Guggemos and Y. Tillé, *Penalized calibration in survey sampling: design-based estimation assisted by mixed models*, J. Statist. Plann. Inference 140 (2010), pp. 3199–3212.

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer; New York: 2011.

[21] A.E. Hoerl and R.W. Kennard, *Ridge regression: biased estimation for nonorthogonal problems*, Technometrics 12 (1970), pp. 55–67.

[22] A.E. Hoerl and R.W. Kennard, *Ridge regression: biased estimation for nonorthogonal problems*, J. Am. Stat. Assoc. 42 (2000), pp. 80–86.

[23] D. Horvitz and D. Thompson, *A generalization of sampling without replacement from a finite universe*, J. Am. Stat. Assoc. 47 (1952), pp. 663–685.

[24] C.-T. Isaki and W.-A. Fuller, *Survey design under the regression superpopulation model*, J. Am. Stat. Assoc. 77 (1982), pp. 49–61.

[25] M. Kuhn and K. Johnson, *Applied Predictive Modelling*, Springer, New York, 2013.

[26] R. Lehtonen and A. Veijanen, *Logistic generalized regression estimators*, Surv. Methodol. 24 (1998), pp. 51–56.

[27] K. McConville and F.J. Breidt, *Survey design asymptotics for the model-assisted penalised spline regression estimator*, J. Nonparametric Regress. 25 (2013), pp. 745–763.

[28] K.S. McConville, F.J. Breidt, T.C. Lee, and G.G. Moisen, *Model-assisted survey regression estimation with the lasso*, J. Surv. Stat. Methodol. 5 (2017), pp. 131–158.

[29] K. McConville and D. Toth, *Automated selection of post-strata using a model-assisted regression tree estimator*, Scand. J. Statist. 46 (2019), pp. 389–413.

[30] G.E. Montanari and M.G. Ranalli, *Nonparametric model calibration in survey sampling*, J. Am. Stat. Assoc. 100 (2005), pp. 1429–1442.

[31] J.D. Opsomer, F.J. Breidt, G. Moisen, and G. Kauermann, *Model-assisted estimation of forest resources with generalized additive models*, J. Am. Stat. Assoc. 102 (2007), pp. 400–409.

[32] J. Quinlan, *Learning with continuous classes*, in *5th Australian Joint Conference on Artificial Intelligence*, Vol. 92, World Scientific, Singapore, 1992, pp. 343–348.

[33] J. Rao and A.C. Singh, *A ridge-shrinkage method for range-restricted weight calibration in survey sampling*, in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, 1997.

[34] P.M. Robinson and C. -E. Särndal, *Asymptotic properties of the generalized regression estimator in probability sampling*, Sankhyā Series B 45 (1983), pp. 240–248.

[35] C.-E. Särndal, *On the π -inverse weighting best linear unbiased weighting in probability sampling*, Biometrika 67 (1980), pp. 639–650.

[36] C.-E. Särndal, B. Swensson, and J Wretman, *Model Assisted Survey Sampling*, Springer Series in Statistics, Springer-Verlag, New York, 1992.

[37] C.-E. Särndal and R. Wright, *Cosmetic form of estimators in survey sampling*, Scand. J. Statist. 11 (1984), pp. 146–156.

[38] T. Ta, J. Shao, Q. Li, and L. Wang, *Generalized regression estimators with high-dimensional covariates*, Stat. Sin. 30 (2020), pp. 1135–1154.

[39] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Statist. Soc. Ser. B 58 (1996), pp. 267–288.

[40] D. Toth, *rpms: Recursive Partitioning for Modeling Survey Data*, 2021. R package version 0.5.1.

[41] D. Toth and J.L. Eltinge, *Building consistent regression trees from complex sample data*, J. Am. Stat. Assoc. 106 (2011), pp. 1626–1636.

[42] L. Wang and S. Wang, *Nonparametric additive model assisted estimation for survey data*, J. Multivar. Anal. 102 (2011), pp. 1126–1140.

[43] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. R. Statist. Soc. Ser. B 67 (2005), pp. 301–320.