# Convergence rates of model-assisted estimators in high-dimensional settings

Mehdi DAGDOUG[(a)], Camelia GOGA[(a)] and David HAZIZA[(b)]
(a) Université de Bourgogne Franche-Comté,
Laboratoire de Mathématiques de Besançon, Besançon, FRANCE
(b) University of Ottawa, Department of Mathematics and Statistics,
Ottawa, CANADA

March 21, 2021

**Abstract**

For decades, methodologies attempting to use efficiently the available auxiliary information have attracted a lot of attention in the survey sampling literature. At the estimation stage, model-assisted estimators have been widely used for that purpose. Many authors have shown that model-assisted estimators maintain important design properties such as asymptotic consistency and unbiasedness irrespective of whether or not the working model is correctly specified. Yet, nowadays, survey practitioners face the emergence of high-dimension data sets. Therefore, a more realistic framework would be to consider that the number of auxiliary variables grows as the population and sample sizes increase. In this article, we adopt this asymptotic framework and establish the consistency rate of model-assisted estimators based on linear unpenalized and penalized regression as well as on tree-based methods towards finite population totals. We also conducted a large simulation study on real electricity consumption data to compare the efficiency of various model-assisted estimators in high-dimensional models and several sampling designs.

*Key words: Design consistency; Elastic net; Lasso; Random forest; Ridge regression.*

## 1 Introduction

Over the last twenty years, survey practitioners witnessed the emergence of data sets of always increasing sizes. With the development of automatic data collection devices, it is no longer unusual to observe data sets containing a very large number of auxiliary variables which raise new estimation challenges. In this configuration, some traditional predictors (e.g. linear regression, k-nearest neighbors, ...) tend to break down as the number of covariates increases. The high-dimensional framework in statistical learning is an active area of research, with many open problems yet to be investigated. For some predictors such as linear regression ones, however, high-dimensional properties have been established, see e.g. Portnoy (1984), among others. As one could expect, it has been demonstrated that the number of covariates plays an important role in the asymptotic properties of the studied predictors. Since model-assisted estimators are constructed upon predictors, it should be expected that their properties depend on the dimension parameter as well. Nonetheless, to our knowledge, this research area has attracted only little attention yet. Notable exception are Cardot et al. (2017) who studied dimension reduction through principal component analysis and established the design consistency of the resulting calibration estimator and Ta et al. (2020) who investigated the properties of the GREG estimator and lasso model-assisted estimator from a model point of view. In this paper, we adopt a design approach and study the convergence rate in high-dimensional settings of the GREG estimator, commonly used penalized estimators such as lasso, ridge or elastic-net, and tree based methods with regression trees and random forests.

## 2 Set-up

Before stating our main results, a few notations are needed. Let $U$ denote the finite population of interest of size $N$. We denote by $Y$ the survey variable and we aim to estimate the finite population total $t_y = \sum_{i\in U} y_i$, where $y_i$ denotes the measurement of the survey variable $Y$ for element $i$. We select a sample $S$ of size $n$ according to a sampling design $p(\cdot)$. The first order inclusion probabilities are denoted by $\pi_i = \mathbb{P}(i \in S) > 0, i \in U$. Without additional information, one could use the design-unbiased Horvitz-Thompson estimator

$$\widehat{t}_{\text{ht}} = \sum_{i\in S} \frac{y_i}{\pi_i}.$$

However, in many cases, more information is available. In particular, we assume that we have access to the measurements of $p$ auxiliary variables $X_1, ..., X_p$ for all the population units; we denote by $\mathbf{x}_i = [x_{i1}, ...., x_{ip}]^\top$ the vector containing the measurements of the auxiliary variables for element $i \in U$. Model-assisted estimation starts with postulating the following working model:

$$\xi : \quad y_i = f(\mathbf{x}_i) + \epsilon_i, \quad i \in U, \tag{1}$$

where $f(\mathbf{x}_i) = \mathbb{E}_\xi [y_i|\mathbf{x}_i]$ and the errors $\epsilon_i$ are independent random variables such that $\mathbb{E}_\xi [\epsilon_i|\mathbf{x}_i] = 0$ and $\mathbb{V}_\xi (\epsilon_i|\mathbf{x}_i) = \sigma^2$ for all $i \in U$. The unknown function $f(\cdot)$ is estimated by $\widehat{f}(\cdot)$ and using the sample data $D_n = (\mathbf{x}_i, y_i)_{i\in S}$. The fitted values $\widehat{f}(\mathbf{x}_i), i \in U$, are then used to construct the model-assisted estimator for $t_y$ based on $\widehat{f}$ :

$$\widehat{t}_{\text{ma}}(\widehat{f}) = \sum_{i\in U} \widehat{f}(\mathbf{x}_i) + \sum_{i\in S} \frac{y_i - \widehat{f}(\mathbf{x}_i)}{\pi_i}. \tag{2}$$

## 3 Model-assisted estimators in high-dimensional settings

Linear regression is probably one of the most studied method in statistics and model-assisted estimators based on linear models have been extensively studied and used in practice (Särndal et al., 1992). The method is easy to analyze and it can be used to represent more abstract non-parametric predictors. In its simplest form, linear regression aims at estimating the unknown linear regression function $f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$ by a function $\widehat{f}_{\text{lr}}$ linear in the covariates, that is,

$$\widehat{f}_{\text{lr}}(\mathbf{x}) = \mathbf{x}^\top \widehat{\boldsymbol{\beta}}_{\text{lr}},$$

where the regression coefficient estimator $\widehat{\boldsymbol{\beta}}_{\text{lr}}$ is obtained from survey data $D_n$ by using a least-square criterion as follows:

$$\widehat{\boldsymbol{\beta}}_{\text{lr}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i\in S} \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{\pi_i}. \tag{3}$$

The well-known model-assisted estimator or GREG estimator is obtained by plugging $\widehat{f}_{\text{lr}}$ in (2), namely $\widehat{t}_{\text{greg}} = \widehat{t}_{ma}(\widehat{f}_{\text{lr}})$.

We are interested in studying the asymptotic behavior, such as asymptotic consistency and bias computed with respect to the sampling design $p(\cdot)$, of several model-assisted estimators in presence of a large number of auxiliary variables. We consider for that the asymptotic framework as introduced by Isaki and Fuller (1982) which allows for the population and sample sizes, $n_v, N_v$ to grow to infinity when $v \to \infty$. In this paper, we consider that the number of auxiliary variables $p_v$ is also growing to infinity. Very mild regularity conditions on the sampling design, the survey variable and the auxiliary information are also supposed, some of these assumptions being extensions to high-dimensional framework of those considered in Robinson and Särndal (1983) (see Dagdoug et al. (2020) for more details).

**Result 3.1.** *Let consider a sequence* $\{\widehat{t}_{\text{greg}}\}_{v\in\mathbb{N}}$ *of GREG-estimators for* $t_y$. *Then,*

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

In statistical learning, penalization methods are used for improving the ordinary least square estimator of the regression coefficient in presence of a large number of covariates. As before, the prediction at a point $\mathbf{x}$ is given by a linear combination of the auxiliary variables, but he unknown regression coefficient is estimated by a different criterion than the one used in (3). More precisely, we define $\widehat{f}_{\text{pen}}(\mathbf{x}) = \mathbf{x}^\top \widehat{\boldsymbol{\beta}}_{\text{pen}}$, where the regression coefficient estimator $\widehat{\boldsymbol{\beta}}_{\text{pen}}$ is obtained from survey data $D_n$ with the following penalized criterion:

$$\widehat{\boldsymbol{\beta}}_{\text{pen}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i\in S} \frac{1}{\pi_i} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \sum_{\ell=1}^{t} \lambda_\ell ||\boldsymbol{\beta}||_{\nu_\ell}^{\gamma_\ell}, \tag{4}$$

where $t \in \mathbb{N}, \nu_\ell \in \mathbb{N}$ and $\gamma_\ell \in \mathbb{R}_+$ are constants chosen before the estimation and $\lambda_\ell \in \mathbb{R}_+$ are the regularization parameters usually computed by cross-validation. Common choices include $t = 1$, $\gamma_1 = \nu_1 = 1$ for the lasso; $t = 1$, $\gamma_1 = \nu_1 = 2$ for ridge; $t = 2$, $\gamma_1 = \nu_1 = 1$, $\gamma_2 = \nu_2 = 2$ for the elastic-net. The lasso or the elastic-net methods have the effect of shrinking some coefficients to zero and therefore they can be seen as variable selection methods as well. Plugging $\widehat{f}_{\text{pen}}$ in (2) leads to the penalized model-assisted estimator denoted by $\widehat{t}_{\text{pen}} = \widehat{t}_{\text{ma}}(\widehat{f}_{\text{pen}})$.

Under the same assumptions as required by result 3.1, we can show that the penalized estimator $\widehat{t}_{\text{pen}} = \widehat{t}_{\text{ma}}(\widehat{f}_{\text{pen}})$ is consistent whenever the GREG estimator is, and that, they share the same convergence rate (Dagdoug et al., 2020). Under supplementary assumptions on the auxiliary information, it is possible to get the improved convergence rate $\sqrt{p_v/n_v}$ for the ridge estimator defined as $\widehat{t}_{\text{ridge}} = \widehat{t}_{\text{ma}}(\widehat{f}_{\text{ridge}})$.

**Result 3.2.** *Consider a sequence of penalized model-assisted estimators* $\{\widehat{t}_{\text{ridge}}\}_{v\in\mathbb{N}}$ *of* $t_y$. *Then,*

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

Linear regression and regularized regression are efficient whenever the regression function belongs to the set of functions linear in the auxiliary variables $X_1, X_2, ..., X_p$. However, when this is not the case, these method may break down. The aforementioned asymptotic results on $\widehat{t}_{\text{greg}}$ and $\widehat{t}_{\text{pen}}$ will hold even for misspecified models, but their design variance will be large in such cases. To remedy this issue, one could use non-parametric models. While some of them are known to be sensitive to the curse of dimensionality (e.g. k-nearest neighbors, splines based methods, kernels, ...), others, such as tree based methods, are more robust to high-dimensional frameworks. Interestingly, linear regression and tree-based methods are very closely linked. That is, a tree based method can be described as a two step process where the first step is dedicated to the creation of a new set of covariates $Z_1, Z_2, ..., Z_T$ based on $D_n$, and the second step consists in estimating the coefficients of the linear regression of $Y$ on these new covariates. More precisely, a regression tree algorithm can be implemented as follows:

1. Use a data partitioning algorithm (e.g. CART, C4.5, ...) which takes as input $D_n$ and outputs a partition $\mathcal{P} = \{\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_T\}$ of $\mathbb{R}^p$, where each element of $\mathcal{P}$, called a terminal node, contains at least $n_0$ sample observations.

2. Define new covariates $Z_1, Z_2, ..., Z_T$ based on terminal nodes, $Z_j = (z_{ij})_{i\in S}$ with $z_{ij} = \mathbb{1}_{\mathbf{x}_i\in\mathcal{A}_j}$ for all $j = 1, \ldots, T$. The estimation of $f$ with a regression tree and based on sample data $D_n$ is $\widehat{f}_{\text{tree}}(\mathbf{x}) = \mathbf{x}^\top \widehat{\boldsymbol{\beta}}_{\text{tree}}$ where
$$\widehat{\boldsymbol{\beta}}_{\text{tree}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^T}{\arg\min} \sum_{i\in S} \frac{1}{\pi_i}(y_i - \mathbf{z}_i^\top \boldsymbol{\beta})^2,$$
where $\mathbf{z}_i = (z_{ij})_{j=1}^T$.

In Result 3.2, ridge regression was shown to converge faster than linear regression. The intuition behind this result is that the length of the vector of estimated coefficients $\widehat{\boldsymbol{\beta}}_{\text{ridge}}$ is, in some sense, increasing slower than the length of the vector $\widehat{\boldsymbol{\beta}}_{\text{lr}}$. Interestingly, the length of $\widehat{\boldsymbol{\beta}}_{\text{tree}}$ is a function of $T$, the number of terminal nodes and not depending on the number of auxiliary variables. Next result uses this intuition to formalize a convergence rate independent of the dimension for a model-assisted estimator $\widehat{t}_{\text{tree}} = \widehat{t}_{\text{ma}}(\widehat{f}_{\text{tree}})$ based on a tree and obtained by plugging $\widehat{f}_{\text{tree}}$ in (2).

**Result 3.3.** *Consider a sequence of tree model-assisted estimators* $\{\widehat{t}_{tree}\}_{v\in\mathbb{N}}$ *for* $t_y$*. Then,*

$$\mathbb{E}_p\left[\frac{1}{N_v}\left|\widehat{t}_{\text{tree}} - t_y\right|\right] = \mathcal{O}\left(\frac{1}{\sqrt{n_v}}\right) + \mathcal{O}\left(\frac{1}{n_{0v}}\right).$$

Therefore, provided that there are enough elements in each terminal nodes and independently of the dimension, $\widehat{t}_{\text{tree}}$ is square root consistent. The same result can be shown also for more complex tree-based methods such as random forests, see Dagdoug et al. (2020) for more details.

## 4 Simulation study

For a more thorough investigation and comparison of model-assisted estimators, we have conducted a large simulation study on high-dimensional real data by considering many different linear or non-linear relationships between survey variables and the auxiliary ones. We were interested in estimating the finite population totals of the so created survey variables and we have considered a wide range of model-assisted estimators such as based on linear regression, penalized regression with lasso, ridge and elastic-net, principal component regression, regression trees, random forests, Cubist, gradient boosting, k-nearest neighbors. We have also tested the behavior of all these estimators, in terms of relative bias and relative efficiency with respect to the Horvitz-Thompson estimator, in presence of high number of auxiliary variables. We used both equal and unequal sampling designs in the simulation study.

Overall, Cubist and penalized regression estimators were the most efficient. Random forests, XGBoost, principal component regression, and, to a lesser extent, k-nearest neighbors, also improved on the Horvitz-Thompson estimator in most cases. Our results also illustrated several notable facts. First, whether or not the survey variable was linear in the auxiliary variables, the estimator based on linear regression was the most impacted in presence of a very large number of auxiliary variables whatever the sampling design. Another interesting finding was the fact that for unequal sampling designs, the model assisted estimator based on random forest may exhibit large bias if the hyper-parameters are not well-chosen (more precisely, if the design variables are not sufficiently taken into account). For more details on this topic, see Dagdoug et al. (2020).

## References

Cardot, H., Goga, C., and Shehzad, M.-A. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Statistica Sinica*, 27(243-260).

Dagdoug, M., Goga, C., and Haziza, D. (2020). Model-assisted estimation in high-dimensional settings for survey data. *arXiv preprint arXiv:2012.07385*.

Isaki, C.-T. and Fuller, W.-A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77:49–61.

Portnoy, S. (1984). Asymptotic behaviour of $m$-estimators of $p$-regression parameters when $p^2/n$ is large.i. consistency. *The Annals of Statistics*, 12(4):1298–1309.

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

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer Series in Statistics. Springer-Verlag, New York.

Ta, T., Shao, J., Li, Q., and Wang, L. (2020). Generalized regression estimators with high-dimensional covariates. *Statistica Sinica*.