Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

# Bagging non-differentiable estimators in complex surveys

by Jianqiang C. Wang, Jean D. Opsomer and Haonan Wang

Release date: December 19, 2014

Statistics Canada Statistique Canada

Canada

## How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

**email** at infostats@statcan.gc.ca,

**telephone**, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

| | |
|---|---|
| • Statistical Information Service | 1-800-263-1136 |
| • National telecommunications device for the hearing impaired | 1-800-363-7629 |
| • Fax line | 1-877-287-4369 |

**Depository Services Program**

| | |
|---|---|
| • Inquiries line | 1-800-635-7943 |
| • Fax line | 1-800-565-7757 |

## To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

## Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

## Standard symbols

The following symbols are used in Statistics Canada publications:

| | |
|---|---|
| . | not available for any reference period |
| .. | not available for a specific reference period |
| ... | not applicable |
| 0 | true zero or a value rounded to zero |
| $0^s$ | value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded |
| $^p$ | preliminary |
| $^r$ | revised |
| x | suppressed to meet the confidentiality requirements of the *Statistics Act* |
| $^E$ | use with caution |
| F | too unreliable to be published |
| * | significantly different from reference category ($p < 0.05$) |

Vol. 40, No. 2, pp. 189-209
Statistics Canada, Catalogue No. 12-001-X

# Bagging non-differentiable estimators in complex surveys

Jianqiang C. Wang, Jean D. Opsomer and Haonan Wang[1]

## Abstract

Bagging is a powerful computational method used to improve the performance of inefficient estimators. This article is a first exploration of the use of bagging in survey estimation, and we investigate the effects of bagging on non-differentiable survey estimators including sample distribution functions and quantiles, among others. The theoretical properties of bagged survey estimators are investigated under both design-based and model-based regimes. In particular, we show the design consistency of the bagged estimators, and obtain the asymptotic normality of the estimators in the model-based context. The article describes how implementation of bagging for survey estimators can take advantage of replicates developed for survey variance estimation, providing an easy way for practitioners to apply bagging in existing surveys. A major remaining challenge in implementing bagging in the survey context is variance estimation for the bagged estimators themselves, and we explore two possible variance estimation approaches. Simulation experiments reveal the improvement of the proposed bagging estimator relative to the original estimator and compare the two variance estimation approaches.

**Key Words:** Bootstrap; Distribution function; Quantile estimation.

# 1 Introduction

Bagging, short for “bootstrap aggregating”, is a resampling method originally introduced to improve “weak” learning algorithms. Bagging was proposed by Breiman (1996), who heuristically demonstrated how it improved the performance of tree-based predictors. Since then, bagging has been applied to a wide range of settings and analyzed by many authors. Bühlmann and Yu (2002) showed the smoothing effect of bagging and its variations on hard-decision classification algorithms, and formalized the notion of “unstable predictors”. Chen and Hall (2003) derived theoretical results on bagging estimators defined by estimating equations. Buja and Stuetzle (2006) considered bagging U-statistics, and claimed that bagging “often but not always decreases variance, whereas it always increases bias”. Friedman and Hall (2007) examined the impact of bagging on nonlinear estimators. More recently, Hall and Robinson (2009) discussed the effects of bagging on cross-validation choice of smoothing parameters, and presented intriguing results on improving the order of the cross-validation selected kernel bandwidth by bagging.

The aforementioned literature studied the effects of bagging on various estimators, especially nonlinear, non-differentiable estimators, under the *iid* (independent and identically distributed) sampling assumption. For dependent data, Lee and Yang (2006); Inoue and Kilian (2008) studied the effects of bagging on economic time series. The former authors studied the bagging effect on non-differentiable predictors like sign functions and quantiles, and the latter focused on bagging pretest predictors with application to U.S. consumer price inflation forecasting.

As this brief literature review shows, bagging is a promising method used to improve the efficiency of estimators. To date, however, bagging for survey estimators has not been considered. This article is a first exploration of the use of bagging in the survey context, including an evaluation of the potential efficiency gain, a number of theoretical results, and a discussion of implementation and variance estimation issues.

[1] Jianqiang C. Wang, Hewlett-Packard Labs, Palo Alto, CA 94304. Jean D. Opsomer, Department of Statistics, Colorado State University, Fort Collins, CO 80523. E-mail: jopsomer@stat.colostate.edu; Haonan Wang, Department of Statistics, Colorado State University, Fort Collins, CO 80523.

Corresponding with general survey practice, we will only consider estimators that can be written as functions of Horvitz-Thompson (HT) estimators. More specifically, we will consider the following three types of estimators. Firstly, many commonly used survey estimators can be written as differentiable functions of HT estimators. For instance, the Hajek estimator, ratio estimator, generalized regression estimator can all be regarded as differentiable functions of HT estimators. Secondly, there are other survey estimators that are non-differentiable, including the Dunstan and Chambers estimator (Dunstan and Chambers 1986), the Rao-Kovar-Mantel estimator (Rao, Kovar, and Mantel 1990), the endogenous post-stratification estimator (Breidt and Opsomer 2008), and estimators of low-income proportion (Berger and Skinner 2003), among others. Thirdly, other estimators are only defined as solutions to weighted estimating equations. For more information on estimating equations in the survey context, see Godambe and Thompson (2009); Fuller (2009), and references therein.

While bagging can be considered a type of replication method, it is quite different from bootstrapping and other replication methods that are designed for variance estimation. Unlike these other methods, bagging is introduced to improve the actual estimator itself. The bagging method can be naturally embedded in large-scale complex surveys, since we can take advantage of replication weights that are readily available in many practical surveys. In this paper, we will show how replicates created for bootstrap variance estimation can be modified and used in bagging the original estimator. Unfortunately, one difficulty in implementing bagging in surveys is the lack of a design-based variance estimator. We will discuss a number of proposals on how to estimate the variance of bagged survey estimators, but further work is still required in this area.

The remainder of this paper is organized as follows. We define our target survey estimators and introduce the bagged version of each estimator in Section 2. We then present the theoretical properties of the bagged estimators in Section 3. Section 4 shows how to use survey replicates to implement bagged versions of estimators, and addresses variance estimation for the resulting bagged estimators. We report on simulation results in Section 5, and conclude the paper with some final remarks in Section 6.

# 2 Bagging survey estimators

## 2.1 General approach

In this section, we discuss the implementation of bagging in the context of survey estimation. We first introduce necessary notation. Let $U$ represent a finite population of size $N,$ in which each element $i \in U$ is associated with a vector of measurements, $\mathbf{y}_i$, in the $q$-dimensional Euclidean space $\mathbb{R}^q$. The sampling design $p(\ )$ is used to draw a random sample $A \subseteq U$ of sample size $n.$ We denote by $\mathcal{Y} = \{\mathbf{y}_i \mid i \in A\}$ the collection of sample observations. Here, the sampling design could be simple random sampling without replacement (SRSWOR), Poisson sampling or a complex design with stratification and/or multiple stages. Under each design, the probability of an element $i$ being included in the sample is denoted by $\pi_i$.

The population mean of the measurement vector $\mathbf{y}$ is denoted by $\boldsymbol{\mu}$. It can be estimated by the Horvitz-Thompson (HT) estimator defined as

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i \in A} \frac{\mathbf{y}_i}{\pi_i} = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{y}_i}{\pi_i} I_i, \tag{2.1}$$

where $I_i$ is the **sample membership indicator** for the $i$-th element. More generally, let $\theta$ denote a population quantity of interest, and $\hat{\theta}(\mathcal{Y})$ is the estimator of $\theta$ based on the sample observations $\mathcal{Y}$. The estimator $\hat{\theta}(\mathcal{Y})$ will be abbreviated as $\hat{\theta}$ when there is no confusion. As noted in the previous section, we assume that $\hat{\theta}$ can be written as a function of simpler estimators of the form (2.1).

In its most general form, the bagging algorithm for survey estimation is as follows:

1. For $b = 1, 2, \ldots, B$:
   a. Draw resample $A_b$ from the random sample $A$, and denote the observations in the resample as $\mathcal{Y}_b^* = \{\mathbf{y}_i \mid i \in A_b\}$.
   b. Calculate the parameter estimate based on the resample $A_b$, denoted by $\hat{\theta}\left(\mathcal{Y}_b^*\right)$.

2. Average over the replicated estimates $\hat{\theta}\left(\mathcal{Y}_1^*\right)$, $\hat{\theta}\left(\mathcal{Y}_2^*\right)$, …, $\hat{\theta}\left(\mathcal{Y}_B^*\right)$ to obtain the bagged survey estimator,

$$\hat{\theta}_{bag} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}\left(\mathcal{Y}_b^*\right). \tag{2.2}$$

In the bagging literature, the resamples $A_b$ are often referred to as **bootstrap samples** (Breiman 1996), and we will do the same here despite the fact that we will not use them for variance estimation.

In the algorithm, the bootstrap samples could be drawn according to the sampling design rather than the empirical distribution of the sample observations, which is more commonly used in the ordinary bagging literature (Breiman 1996) and equivalent to simple random sampling (with or without replacement). For example, if the sample $A$ is drawn using stratified or cluster sampling, such design scheme could be taken into consideration when selecting the resamples. More generally in the survey context, step 1 of the proposed bagging algorithm can be treated in the framework of two-phase sampling: the first phase corresponds to the original sample $A$ and the second phase to the resample $A_b$. Thus the classical expansion estimator for two-phase designs Särndal, Swensson and Wretman (1997) is implemented in calculating the replicated estimator $\hat{\theta}\left(\mathcal{Y}_b^*\right)$. In the resample $A_b$, the pseudo inclusion probability for the $i$-th element is $\pi_i^* = \pi_i \pi_{i|A}$ where $\pi_{i|A} = \Pr\left(i \in A_b \mid i \in A\right)$ is the inclusion probability of the $i$-th element in resample $A_b$ given that it is included in sample $A$. Hence, the bagged estimator is an approximation to the expectation of the two-phase estimator with respect to the second sampling phase, which is also referred to as **bootstrap expectation** in ordinary bagging methods (Bühlmann and Yu 2002). Although a general design for the bootstrap samples is possible, in the theoretical portions of this article we will restrict ourselves to SRSWOR. To broaden the scope of our discussion, in the variance estimation and numerical section, we introduce the case in which the bootstrap samples are drawn by stratified SRSWOR with the same strata as the original sample $A$, which is a useful and realistic extension.

As an example, we consider the HT estimator as defined in (2.1). The bootstrap resampling from the realized sample $A$ is drawn under SRSWOR of size $k$. Under this resampling scheme, the replicated sample estimator is defined as

$$\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)=\frac{1}{N}\sum_{i\in A_b}\frac{\mathbf{y}_i}{\pi_i^*}, \tag{2.3}$$

where the pseudo inclusion probability $\pi_i^* = \pi_i \pi_{i|A} = k\pi_i / n$. Then the bagged version of the classical $\pi^*$-estimator can be calculated using (2.2). Straightforward calculation shows that the bagged estimator is identical to the original HT estimator if all SRSWOR samples of size $k$ are enumerated in calculating (2.2). The same result holds for any other linear survey estimator. In general, the calculation of the bagged estimator $\hat{\theta}_{bag}$ is not as easy. In the rest of this section, we will focus on such calculations for the three types of nonlinear survey estimators discussed in Section 1.

## 2.2 Bagging differentiable survey estimators

For the survey estimators that are differentiable functions of HT estimators, the population quantity of interest can also be written as a differentiable function of population means; that is, $\theta_d = m(\boldsymbol{\mu})$, where $m(\cdot)$ is a known differentiable function. The subscript "$d$" stands for **differentiable** in contrast to **non-differentiable** $(\theta_{nd})$ and **estimating equation** $(\theta_{ee})$ coming later. A direct plug-in estimator of $\theta_d$, based on sample observations $\mathcal{Y}$, can be written as

$$\hat{\theta}_d = m(\hat{\boldsymbol{\mu}}), \tag{2.4}$$

where $\hat{\boldsymbol{\mu}}$ is defined in (2.1). Thus, the replicated sample version of $\hat{\theta}_d$ can be expressed as

$$\hat{\theta}_d\left(\mathcal{Y}_b^*\right)=m\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)\right),$$

where $\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)$ is defined by (2.3). Then the bagged estimator of $\theta_d$, denoted by $\hat{\theta}_{d,bag}$, is defined using (2.2).

## 2.3 Bagging explicitly defined non-differentiable estimators

As an example of this type of estimators, consider the estimation of the proportion of households with income below the poverty line for a population. Such quantity can be written as $(1/N)\sum_{i=1}^{N} I\left(y_i \le \lambda_N\right)$, where $y_i$ is the income value for the $i$-th household in the population, and $\lambda_N$ is the population poverty line. It can be seen that this quantity of interest is the mean of indicator kernel functions, and the kernel function is non-differentiable with respect to $\lambda_N$. Here, we consider a more general class in which the kernel is an arbitrary non-differentiable but bounded function. This type of population quantity can be expressed as

$$\theta_{nd}=\frac{1}{N}\sum_{i=1}^{N}h\left(\mathbf{y}_i-\boldsymbol{\lambda}_N\right),$$

where $\boldsymbol{\lambda}_N$ is an unknown population parameter, for example, the mean, a quantile or other population quantity, and $h(\mathbf{y}-\boldsymbol{\lambda}):\mathbb{R}^p \to \mathbb{R}$ is a non-differentiable function of $\boldsymbol{\lambda}$. The population quantity $\theta_{nd}$ generalizes the notion of the proportion below an estimated level and resembles the general form of a U-statistic.

Wang and Opsomer (2011) studied a class of U-statistics-like estimators, namely, non-differentiable survey estimators,

$$\hat{\theta}_{nd} = \frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i}h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\right), \tag{2.5}$$

where $\hat{\boldsymbol{\lambda}}$ is a design-based estimator of $\boldsymbol{\lambda}_N$. In the non-survey context, estimators of this type are regarded as "non-differentiable functions of the empirical distribution" (Bickel, Götze and van Zwet 1997). The study of appropriate bootstrap procedures for such estimators was carried out by Beran and Srivastava (1985) and Dümbgen (1993), among others. We define the replicated version of $\hat{\theta}_{nd}$ based on resample $A_b$ as

$$\hat{\theta}_{nd}\left(\mathcal{Y}_b^*\right) = \frac{1}{N}\sum_{i\in A_b}\frac{1}{\pi_i^*}h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right),$$

where $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$ solely depends on the bootstrap resample $A_b$, and the bagged estimator is then defined by averaging replicated estimators. Suppose that the resampling process is SRSWOR of size $k$, and every subsample is selected in calculating the bagging estimator, then the bagging estimator takes the following form after manipulation,

$$\hat{\theta}_{nd,bag} = \frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i\binom{n-1}{k-1}}\sum_{A_b\ni i}h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right), \tag{2.6}$$

which replaces $h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\right)$ in (2.5) by a "smoothed" quantity $\sum_{A_b\ni i}h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right)\Big/\binom{n-1}{k-1}$, by averaging the "jumps" in the estimator. Very often, variance reduction can be achieved by this replacement. The summand $\sum_{A_b\ni i}h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right)\Big/\binom{n-1}{k-1}$ is the **bootstrap expectation** of $h\left(\mathbf{y}_i - \cdot\right)$ and can be approximated using the convolution of $h\left(\mathbf{y}_i - \cdot\right)$ with the sampling distribution of $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$. Study of the theoretical aspects of $\hat{\theta}_{nd,bag}$ is deferred until Section 3.

## 2.4 Bagging estimators defined by non-differentiable estimating equations

Finally, we explain how to bag estimators defined by non-differentiable estimating equations. For ease of presentation, we consider a one-dimensional parameter of interest. The population parameter $\theta_{ee}$ of interest is defined as

$$\theta_{ee} = \inf\left\{\gamma : S(\gamma) \geq 0\right\},$$

where

$$S(\gamma) = \frac{1}{N}\sum_{i=1}^{N}\psi\left(y_i - \gamma\right),$$

and $\psi(\cdot)$ is a non-differentiable real function. We can estimate the population parameter $\theta_{ee}$ by $\hat{\theta}_{ee}$, where

$$\hat{\theta}_{ee} = \inf\left\{\gamma : \hat{S}(\gamma) \geq 0\right\}$$

with

$$\hat{S}(\gamma) = \frac{1}{N}\sum_{i \in A}\frac{1}{\pi_i}\psi(y_i - \gamma).$$

A frequently encountered estimator of this type is the sample quantile defined by inverting the sample cumulative distribution function (Francisco and Fuller 1991), where $\psi(y_i - \gamma) = I_{(y_i \leq \gamma)} - \alpha$ for the $\alpha$-quantile.

Conceptually, there are two versions of bagging $\hat{\theta}_{ee}$, one is to solve the "bagged estimating equation" defined by bagging the score function, and another is to average over resampled estimates of $\hat{\theta}_{ee}$. Similarly to the discussion in Section 2.1, the first version results in an estimator equivalent to the original estimator, because the bootstrap expectation of bootstrap samples of $\hat{S}(\gamma)$ is equal to $\hat{S}(\gamma)$ for fixed $\gamma$. We therefore only consider the latter version. To define the bagged estimating equation estimator, we first define the replicated score function $\hat{S}_b(\gamma)$ based on resample $A_b$ as

$$\hat{S}_b(\gamma) = \frac{1}{N}\sum_{i \in A_b}\frac{1}{\pi_i^*}\psi(y_i - \gamma).$$

Then the replicated estimator based on $A_b$ is defined as $\hat{\theta}_{ee}\left(\mathcal{Y}_b^*\right) = \inf\left\{\gamma : \hat{S}_b(\gamma) \geq 0\right\}$. Thus the overall bagging estimator is defined as

$$\hat{\theta}_{ee,bag} = \frac{1}{\binom{n}{k}}\sum \hat{\theta}_{ee}\left(\mathcal{Y}_b^*\right), \tag{2.7}$$

where the average is over all possible without-replacement samples of size $k$ selected from $A$. Chen and Hall (2003) discussed bagging estimators defined by nonlinear estimating equations under the *iid* setup, and they stated that bagging does not always improve the precision of estimators under study.

# 3 Theoretical results

We begin by briefly describing the asymptotic analysis of the bagging estimators under general sampling design from a finite population, i.e. the design-based setting. We do this under the usual increasing-population framework, where we consider an increasing sequence of nested populations, say $U_N$, $N = 1, 2, \ldots,$ with finite population means $\boldsymbol{\mu}_N$. Associated with the sequence of populations is a sequence of sampling designs used to draw random sample $A_N \subseteq U_N$ of sample size $n_N$, with associated inclusion probabilities $\pi_{iN}$. As commonly done in the survey literature, we suppress the $N$ subscript in the sample $A$, the sample size $n$ and the inclusion probabilities $\pi_i$. For the sake of brevity, only design-based asymptotic results for bagging differentiable estimator $\hat{\theta}_d$ and non-differentiable $\hat{\theta}_{nd}$ are provided. The formal assumptions under which the results are obtained and the theorems for differentiable and non-differentiable estimators are in Appendix A.1. The main result we are able to obtain in this design-based setting is that, if we are starting from a design-consistent estimator and we let the number of bootstrap samples $k$ grow with $n,$ the bagged versions of the estimators are also design consistent. This is clearly a key property of these estimators, since there would be no reason to consider them unless they satisfied this design consistency.

Unfortunately, the above design-based results are quite limited and in particular, do not provide an asymptotic distribution with which one might be able to perform inference, another highly desirable property of survey estimators. We therefore also consider a model-based setting, under which we are able to obtain an asymptotic variance approximation. In presenting model-based results, we assume the sampling design selecting the original sample $A$ is an equal probability design, and the population characteristics can be regarded as an *iid* sample from a superpopulation distribution. Under this framework, the bagging estimator can be treated as a U-statistic. Thus we can apply the theory on U-statistics to obtain asymptotic expansion of bagging estimators. The analysis parallels that of Bühlmann and Yu (2002) and Buja and Stuetzle (2006). For the current paper, we restrict ourselves to bootstrap samples of size $k$ where $k$ is bounded and fixed. Under this asumption, the bagging estimators can be regarded as fixed-degree U-statistics, for which asymptotic theory has been well developed. A more interesting case is when the resample size $k$ grows with sample size $n,$ and this leads to infinite-degree U-statistics. Infinite-degree U-statistics have applications in studying the Kaplan-Meier estimator and $m$-out-of-$n$ bootstrap estimators, and the readers are referred to Frees (1989); Heilig (1997); Heilig and Nolan (2001), and the references therein on their statistical properties. Schick and Wefelmeyer (2004) studied the statistical properties of infinite-degree U-statistics constructed from moving averages of innovations in time series. The study of bagging estimators by viewing them as infinite-degree U-statistics is out of the scope of the current paper, and hence we limit ourselves to the case of fixed and bounded bootstrap sample size in the model-based case.

We first consider bagged estimator (2.5). Under SRSWOR, estimator (2.5) can be simplified to

$$\hat{\theta}_{nd} = \frac{1}{n}\sum_{i\in A} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\right)$$

and the bagged version of $\hat{\theta}_{nd}$ is defined as

$$\hat{\theta}_{nd,bag} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\binom{n-1}{k-1}} \sum_{A_b \ni i} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right) \tag{3.1}$$

where $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$ only depends on resample $A_b.$ For ease of presentation, we take $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$ as the sample mean. In this case, straightforward algebra reveals that

$$\hat{\theta}_{nd,bag} = \frac{1}{\binom{n}{k}} \sum_{A_b \in \mathcal{A}} \left\{ \frac{1}{k} \sum_{i \in A_b} h\left( \frac{k-1}{k}\mathbf{y}_i - \frac{1}{k}\sum_{j \neq i} \mathbf{y}_j \right) \right\},$$

where $\mathcal{A}$ is the collection of subsets of size $k$ from set $\{1,2,\ldots,n\}.$ The estimator $\hat{\theta}_{nd,bag}$ is a degree-$k$ U-statistic with kernel

$$g\left(y_1,\ldots,y_k\right)=\frac{1}{k}\sum_{i=1}^{k}h\left(\frac{k-1}{k}\mathbf{y}_i-\frac{1}{k}\sum_{\substack{j=1\\ j\neq i}}^{k}\mathbf{y}_j\right)$$

provided that $k$ remains finite.

One can see that the bagging estimator $\hat{\theta}_{nd,bag}$ is a symmetric statistic of $\mathbf{y}_i$, and standard theory on symmetric statistics (Lee 1990) applies. The results are stated in Theorem 1, with assumptions and proofs in Appendix A.2.

**Theorem 1** *Under Assumptions M.1-M.4 on the superpopulation distribution, sampling and resampling designs,*

$$\mathrm{AV}\left(\hat{\theta}_{nd,bag}\right)^{-1/2}\left(\hat{\theta}_{nd,bag}-\theta_{nd,\infty}\right)\overset{p}{\rightarrow}N\left(0,1\right),\tag{3.2}$$

*where the limiting value* $\theta_{nd,\infty}=\lim_{n\to\infty}\mathrm{E}\left[h\left(\mathbf{y}_i-\hat{\boldsymbol{\lambda}}\right)\right]$, *the asymptotic variance*

$$\mathrm{AV}\left(\hat{\theta}_{nd,bag}\right)=\frac{1}{n}\mathrm{Var}\left[u\left(\mathbf{y}_i\right)\right]+\frac{\left(k-1\right)^2}{n}\mathrm{Var}\left[v\left(\mathbf{y}_i\right)\right]+\frac{2\left(k-1\right)}{n}\mathrm{Cov}\left[u\left(\mathbf{y}_i\right),v\left(\mathbf{y}_i\right)\right],\tag{3.3}$$

*and*

$$u\left(\mathbf{y}\right)=\mathrm{E}\left[h\left(\mathbf{y}-\hat{\boldsymbol{\lambda}}\left(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1},\mathbf{y}\right)\right)\right],$$
$$v\left(\mathbf{y}\right)=\mathrm{E}\left[h\left(\mathbf{y}_1-\hat{\boldsymbol{\lambda}}\left(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1},\mathbf{y}\right)\right)\right].$$

As indicated by (3.3), the asymptotic variance of the bagging estimator depends on unknown functions $u\left(\mathbf{y}\right)$ and $v\left(\mathbf{y}\right)$, which are expectations of $h(\cdot)$ with respect to the superpopulation distribution. In $u\left(\mathbf{y}\right)$ and $v\left(\mathbf{y}\right)$, $\hat{\boldsymbol{\lambda}}\left(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1},\mathbf{y}\right)$ is calculated from $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1}$ together with an arbitrary vector $\mathbf{y}$. The expectation is with respect to the distribution of *iid* random vectors $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1}$. This high-dimensional expectation is difficult to calculate and may not have an explicit expression in general. The exact form of $u(\cdot)$ and $v(\cdot)$ can not be obtained but can be approximated via a resampling-based approach. The unknown functions $u(\cdot)$ and $v(\cdot)$ are defined as expectations of respective quantities with respect to the superpopulation distribution, which can be approximated by the expectation with respect to the empirical distribution.

The model-based asymptotic variance can be estimated along with the process of bagging. We can calculate integrands $h\left(\mathbf{y}-\hat{\boldsymbol{\lambda}}\left(\mathbf{y}_1^*,\mathbf{y}_2^*,\ldots,\mathbf{y}_{k-1}^*,\mathbf{y}\right)\right)$ and $h\left(\mathbf{y}_1-\hat{\boldsymbol{\lambda}}\left(\mathbf{y}_1^*,\mathbf{y}_2^*,\ldots,\mathbf{y}_{k-1}^*,\mathbf{y}\right)\right)$ based on each bootstrap sample, with $\mathbf{y}$ denoting where we want to evaluate $u(\cdot)$ and $v(\cdot)$, and $\mathbf{y}_1^*,\mathbf{y}_2^*,\ldots,\mathbf{y}_{k-1}^*$ denoting resampled values. Then we can average each quantity to approximate the expectation. Finally, the variance can be estimated by computing the sample variance of the expectations evaluated at each of the sample points. For nonsmooth estimators like the ones we are dealing with, it is often recommended to use smoothed bootstrap in variance approximation (Efron 1979; Davison and Hinkley 1997). We apply the smoothed bootstrap and add a small amount of noise to each resampled value to smooth the underlying function. The detailed algorithm will be explained in Section 5 through an example.

We now study the model-based result of bagging estimators defined by estimating equations (2.7). A special case in this framework is bagging sample quantiles, which was studied by Knight and Bassett (2002). Knight and Bassett (2002) considered both bootstrap and SRSWOR for resampling, and studied the effects of bagging on the remainder term in the Bahadur representation of quantiles (Bahadur 1966). We take a slightly different perspective and treat the bagging estimator as a U-statistic. Assumptions and proof are again in Appendix A.2. Note that Assumption M.5 requires that the non-differentiable estimating function have a smooth limit. In the next theorem, we provide linearization of the bagging estimating equation estimator and give an expression for the asymptotic variance.

**Theorem 2** *Under Assumptions M.1-M.3 and M.5, the following asymptotic result holds for the bagged estimating equation estimator (2.7),*

$$\mathrm{AV}\left(\hat{\theta}_{ee,bag}\right)^{-1/2}\left(\hat{\theta}_{ee,bag}-\theta_{ee,\infty}\right)\xrightarrow{p}N\left(0,1\right),\tag{3.4}$$

*where* $\theta_{ee,\infty}$ *denotes the asymptotic limit of population quantity* $\theta_{ee}$, *the asymptotic variance of* $\hat{\theta}_{ee,bag}$ *is*

$$\mathrm{AV}\left(\hat{\theta}_{ee,bag}\right)=\frac{k^2}{n}\mathrm{Var}\left[u\left(y_i\right)\right],\tag{3.5}$$

*and*

$$u\left(y\right)=\mathrm{E}\ \inf\left\{\gamma:\frac{1}{k}\sum_{i=1}^{k-1}\psi\left(y_i-\gamma\right)+\frac{1}{k}\psi\left(y-\gamma\right)\geq 0\right\}.\tag{3.6}$$

As we saw for the bagged estimator (3.1), the asymptotic results in Theorem 2 involve an unknown function. This function can again be computed using resampling that takes advantage of the available replicate samples.

# 4 Variance Estimation

While the model-based approach makes it possible to obtain asymptotic distributions and hence perform inference that is asymptotically correct, we are most interested here in the design-based applications of bagging. In the design-based context, the construction of the bagging estimator can be naturally combined with the variance estimation of the original statistic, by taking advantage of the replication samples released by the statistical agencies. In this article, we take stratified simple random sampling as a specific example, with a bootstrap sampling design of stratified SRSWOR.

We begin by applying a version of the Rao and Wu (1988) bootstrap procedure to estimate the variance of the survey estimators prior to bagging. Let $N_h$, $n_h$ and $k_h$ denote the population size, sample size and sub-sample size in the $h$-th stratum, $h=1,2,\ldots,H$. Here, $B$ bootstrap samples are drawn by stratified simple random sample without replacement of size $k_h$ for computing the bootstrap variance of the original statistic and the bagging estimator. For each bootstrap sample, we assign a weight of

$$\frac{N_h}{N}\left(1-k_h^{1/2}\left(n_h-1\right)^{-1/2}\left(1-\frac{n_h}{N_h}\right)^{1/2}\right)\frac{1}{n_h}+\frac{N_h}{N}k_h^{1/2}\left(n_h-1\right)^{-1/2}\left(1-\frac{n_h}{N_h}\right)^{1/2}\frac{1}{k_h}$$

to each sampled element in the $h$-th stratum, and

$$\frac{N_h}{N}\left(1-k_h^{1/2}\left(n_h-1\right)^{-1/2}\left(1-\frac{n_h}{N_h}\right)^{1/2}\right)\frac{1}{n_h}$$

to the nonsampled elements. We then use the ordinary variance of the replicated sample estimators as variance estimator. The aforementioned weighting scheme is algebraically identical to equation 4.1 of Rao and Wu (1988), in which the finite population correction is incorporated into replication weights. The resampling variance estimator derived from the weighting method reduces to ordinary variance estimator under stratified SRSWOR and guarantees design unbiasedness. In order to combine bagging with bootstrap variance estimator, we use the same bootstrap samples to construct the bagging estimators for the population quantities of interest.

Under the design-based framework, no analytic variance estimator is available for the bagged estimator in general. For now, we would suggest the following two variance estimation approaches in practice:

(Var. 1) Use the estimated variance of the original estimator even though the bagged estimator may have a smaller variance. This method provides confidence intervals of the same width but outperforms the original confidence interval in having larger coverage rate.

(Var. 2) Multiply the estimated variance of the original estimator by an adjustment factor accounting for the likely improvement in efficiency. One possible choice for such a factor is the efficiency gain assuming the sample is an *iid* sample from an infinite superpopulation. The factor can be determined by using the results of Theorems 1 and 2, or by a nonparametric bootstrap experiment. One such possible bootstrap procedure is double bootstrap, which is implemented by drawing ordinary bootstrap resamples to estimate the variance of the original estimator, and another level of SRSWOR resamples to determine the variance of the bagging estimator. One can estimate the ratio of the variance of bagging estimator to original estimator using these nested bootstrap samples, and multiply the design variance of the original estimator by this ratio.

We will explore both approaches in the simulations in Section 5, but this is clearly an area in which further research is warranted.

# 5 Simulations

To evaluate the practical behavior of bagging in the survey context, we generate a finite population of size $N = 2,000$ with three strata. The size of each stratum is denoted as $N_h$ with $h = 1, 2, 3,$ and the stratum proportions are fixed at $\left(N_1, N_2, N_3\right)/N = \left(0.5, 0.3, 0.2\right).$ The distribution of the target variable $y_i$

within each stratum is $y_{1i} \sim \left|N(-1,1)\right|$, $y_{2i} \sim \Gamma(1,1)$ and $y_{3i} \sim \left|N(3,2)\right|$. An auxiliary variable $x_i$ is generated via $x_i = A_0 + A_1 y_i + A_2 (G_i - \alpha/\beta)$ where $A_0 = A_1 = 2$, $A_2 = 1$, $\alpha = 2$, $\beta = 1$ and $G_i \overset{iid}{\sim} \Gamma(2,1)$. We repeatedly draw samples of size $n$ using stratified simple random sampling from the population of interest and the sample size allocation is $(n_1, n_2, n_3)/n = (0.3, 0.3, 0.4)$. In this set-up, the design is clearly informative, because the observations are not *iid* in the overall population and are correlated with the inclusion probabilities.

We are interested in three population quantities: a population $\alpha$-quantile, a population proportion below a given fraction of a population quantile (see Berger and Skinner 2003, for an example) and the Rao-Kovar-Mantel (RKM) estimator of the distribution function (Rao et al. 1990). The former is an example of a non-differentiable estimating equation-based estimator, while the latter two are explicitly defined non-differentiable estimators. The sample estimator of the quantile is found by inverting the estimated cumulative distribution function. The sample estimator of the proportion below a given fraction of a population quantile is the HT estimator of the proportion of observations below the sample median of a variable of interest times a constant $c$,

$$\hat{\theta}_{pr} = \frac{1}{N} \sum_{i \in A} \frac{1}{\pi_i} \mathrm{I}_{\left(y_i \le c\hat{\theta}_{\mathrm{med}}\right)},$$

where $\hat{\theta}_{\mathrm{med}}$ denotes the sample median of the $y_i$. The design-based RKM difference estimator based on a ratio model is

$$\hat{\theta}_{\mathrm{RKM}} = \frac{1}{N} \left\{ \sum_{i \in A} \frac{1}{\pi_i} \mathrm{I}_{(y_i \le t)} + \sum_{i=1}^{N} \mathrm{I}_{\left(\hat{R}x_i \le t\right)} - \sum_{i \in A} \frac{1}{\pi_i} \mathrm{I}_{\left(\hat{R}x_i \le t\right)} \right\}, \tag{5.1}$$

where $\hat{R}$ denotes the estimated ratio between $y$ and $x$.

The design variance of these non-differentiable estimators is somewhat cumbersome to estimate. For variance and interval calculations for sample quantiles, the readers are referred to Francisco and Fuller (1991), Sitter and Wu (2001), and references therein. For proportion below an estimated level, see Shao and Rao (1993) and Berger and Skinner (2003).

The design variances of the original estimators $\hat{\theta}_{qr}$, $\hat{\theta}_{pr}$ and $\hat{\theta}_{\mathrm{RKM}}$, are estimated via the without-replacement bootstrap procedure described in the previous section. We employ a bootstrap sample size of $k_h = n_h/2$. The so-constructed bagging estimators are often referred to as subagging estimators (Bühlmann and Yu 2002). It was established that without-replacement samples of size $n/2$ produces similar results to with replacement samples of size $n$ in bagging (Buja and Stuetzle 2006; Friedman and Hall 2007). We apply the two variance approaches for bagging estimators proposed in the previous section, i.e. one identical to that of unbagged estimator (Var. 1) and another one that multiplies the original variance estimate by a model-based adjustment factor (Var. 2). The factor is determined by double bootstrap on one particular sample. In principle, one should repeat the exercise for each sample, but this is precluded by the heavy computational burden. The confidence intervals of all three estimators are constructed by normal approximation. The confidence intervals for the proportion and the RKM estimator are constructed by normal approximation on *logit* transformed scale, $\log\left[\hat{\theta}/(1-\hat{\theta})\right]$ or $\log\left[\hat{\theta}_{bag}/(1-\hat{\theta}_{bag})\right]$, and then back transformation (Agresti 2002; Korn and Graubard 1998).

Table 5.1 summarizes the bias, standard deviation and MSE ratio of the original and bagged sample quantiles and Table 5.2 examines the variance estimators and confidence intervals. The sample sizes are chosen to be $n = 100$ and $200.$ From Table 5.1, we can see that the bagged quantile estimator is more efficient than the original estimator since the MSE ratio is less than one in this simulation experiment. The smoothing effects of bagging generally become more prominent as we decrease the sample size. In Table 5.2, we compare the two confidence intervals with bagging point estimator to that of original confidence intervals. As expected, the confidence interval constructed via method 1 has the same length and higher coverage than the original. In this example, the confidence intervals via method 2 are narrower but maintain coverage level close to nominal.

**Table 5.1**
**Bias, standard deviation and MSE ratios of sample quantiles and bagged sample quantiles; population size $N = 2{,}000,$ number of bootstraps $B = 2{,}000,$ and results are from $2{,}000$ simulations**

| | $n = 100,\ k = 50$ | | | | | $n = 200,\ k = 100$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | **0.2** | **0.3** | **0.5** | **0.7** | **0.8** | **0.2** | **0.3** | **0.5** | **0.7** | **0.8** |
| $\text{bias}\left(\hat{\theta}_{qt}\right)$ | 0.002 | 0.008 | 0.000 | -0.005 | -0.035 | -0.008 | 0.005 | 0.006 | 0.007 | -0.005 |
| $\text{bias}\left(\hat{\theta}_{qt,bag}\right)$ | 0.018 | 0.019 | -0.001 | -0.007 | -0.043 | -0.006 | 0.009 | 0.005 | 0.006 | -0.022 |
| $\text{sd}\left(\hat{\theta}_{qt}\right)$ | 0.093 | 0.124 | 0.149 | 0.181 | 0.212 | 0.070 | 0.076 | 0.103 | 0.136 | 0.148 |
| $\text{sd}\left(\hat{\theta}_{qt,bag}\right)$ | 0.089 | 0.112 | 0.138 | 0.167 | 0.197 | 0.065 | 0.073 | 0.099 | 0.127 | 0.139 |
| $\dfrac{MSE_p\left(\hat{\theta}_{qt,bag}\right)}{MSE_p\left(\hat{\theta}_{qt}\right)}$ | 0.946 | 0.844 | 0.859 | 0.854 | 0.875 | 0.866 | 0.924 | 0.919 | 0.862 | 0.912 |

**Table 5.2**
**Relative bias, coverage probability and confidence interval width of bootstrap variance estimators for sample quantiles and unadjusted $\left(\hat{V}_1\right)$ and adjusted $\left(\hat{V}_2\right)$ variance estimators for bagged sample quantiles; simulation setting is the same as in Table 5.1**

| | $n = 100,\ k = 50$ | | | | | $n = 200,\ k = 100$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | **0.2** | **0.3** | **0.5** | **0.7** | **0.8** | **0.2** | **0.3** | **0.5** | **0.7** | **0.8** |
| $\dfrac{\text{E}\left[\hat{V}_{boot}\left(\hat{\theta}_{qt}\right)\right]}{V\left(\hat{\theta}_{qt}\right)}$ | 1.208 | 1.091 | 1.099 | 1.135 | 1.205 | 1.067 | 1.117 | 1.093 | 1.098 | 1.180 |
| $\dfrac{\text{E}\left[\hat{V}_1\left(\hat{\theta}_{qt,bag}\right)\right]}{V\left(\hat{\theta}_{qt,bag}\right)}$ | 1.327 | 1.325 | 1.279 | 1.331 | 1.402 | 1.224 | 1.217 | 1.188 | 1.273 | 1.326 |
| $\dfrac{\text{E}\left[\hat{V}_2\left(\hat{\theta}_{qt,bag}\right)\right]}{V\left(\hat{\theta}_{qt,bag}\right)}$ | 1.307 | 1.217 | 1.196 | 1.184 | 1.383 | 1.245 | 1.249 | 1.392 | 1.107 | 1.104 |
| C.P.(C.I.) | 0.944 | 0.934 | 0.924 | 0.928 | 0.922 | 0.938 | 0.951 | 0.942 | 0.935 | 0.950 |
| $\text{C.P.}\left(\text{C.I.1.}_{bag}\right)$ | 0.950 | 0.946 | 0.938 | 0.938 | 0.939 | 0.942 | 0.950 | 0.946 | 0.943 | 0.954 |
| $\text{C.P.}\left(\text{C.I.2.}_{bag}\right)$ | 0.949 | 0.934 | 0.932 | 0.929 | 0.938 | 0.944 | 0.952 | 0.958 | 0.927 | 0.936 |
| Width(C.I.) | | | | | | | | | | |
| $\text{Width}\left(\text{C.I.1.}_{bag}\right)$ | 0.386 | 0.492 | 0.597 | 0.729 | 0.880 | 0.277 | 0.309 | 0.414 | 0.544 | 0.612 |
| $\text{Width}\left(\text{C.I.2.}_{bag}\right)$ | 0.383 | 0.472 | 0.577 | 0.688 | 0.874 | 0.279 | 0.313 | 0.448 | 0.508 | 0.559 |

Tables 5.3 and 5.4 summarize design-based results on the low-income proportion estimator. Based on the MSE ratio, we can see that the bagging estimator is uniformly more efficient than the original estimator, and the MSE of bagging estimator is less than $50\%$ of that of original estimator in a few cases (see $c = 1.2$). The likely reason for this is that the estimator involves two "levels" of non-differentiability: the sample median being a non-differentiable estimator, whose efficiency gain was shown in Table 5.1, and the low-income proportion being a non-differentiable function of the sample median. The "jumps" in the estimators are smoothed out by bagging, resulting in a more stable estimator. The confidence interval comparison in Table 5.4 leads to results similar to the quantile case.

**Table 5.3**
**Bias, standard deviation and MSE ratio of estimated proportion below a constant $c$ multiplied by estimated median and the bagged proportion estimator; population size $N = 2{,}000,$ number of bootstraps $B = 2{,}000,$ and results are from $2{,}000$ simulations**

| | $n = 100,\ k = 50$ | | | | | $n = 200,\ k = 100$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $c$ | **0.2** | **0.4** | **0.6** | **1.2** | **1.5** | **0.2** | **0.4** | **0.6** | **1.2** | **1.5** |
| $\text{bias}\left(\hat{\theta}_{pr}\right)$ | -0.002 | -0.002 | -0.003 | 0.011 | 0.006 | 0.000 | -0.002 | -0.005 | -0.004 | -0.004 |
| $\text{bias}\left(\hat{\theta}_{pr,bag}\right)$ | -0.004 | -0.004 | -0.007 | 0.017 | 0.009 | -0.001 | -0.005 | -0.009 | -0.001 | -0.004 |
| $\text{sd}\left(\hat{\theta}_{pr}\right)$ | 0.034 | 0.039 | 0.038 | 0.034 | 0.046 | 0.023 | 0.027 | 0.026 | 0.026 | 0.036 |
| $\text{sd}\left(\hat{\theta}_{pr,bag}\right)$ | 0.031 | 0.035 | 0.031 | 0.020 | 0.034 | 0.022 | 0.025 | 0.022 | 0.017 | 0.029 |
| $\dfrac{MSE_p\left(\hat{\theta}_{pr,bag}\right)}{MSE_p\left(\hat{\theta}_{pr}\right)}$ | 0.861 | 0.821 | 0.709 | 0.538 | 0.581 | 0.883 | 0.860 | 0.783 | 0.434 | 0.671 |

**Table 5.4**
**Relative bias, coverage probability and confidence interval width of bootstrap variance estimators for sample proportions and unadjusted $\left(\hat{V}_1\right)$ and adjusted $\left(\hat{V}_2\right)$ variance estimators for bagged sample proportions; simulation setting is the same as in Table 5.3. We use "C.I.T." to denote confidence intervals obtained with logit transformation**

| | $n = 100,\ k = 50$ | | | | | $n = 200,\ k = 100$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $c$ | **0.2** | **0.4** | **0.6** | **1.2** | **1.5** | **0.2** | **0.4** | **0.6** | **1.2** | **1.5** |
| $\dfrac{\text{E}\left[\hat{V}_{boot}\left(\hat{\theta}_{pr}\right)\right]}{V\left(\hat{\theta}_{pr}\right)}$ | 1.122 | 1.191 | 1.325 | 1.472 | 1.281 | 1.140 | 1.191 | 1.251 | 1.350 | 1.217 |
| $\dfrac{\text{E}\left[\hat{V}_1\left(\hat{\theta}_{pr,bag}\right)\right]}{V\left(\hat{\theta}_{pr,bag}\right)}$ | 1.323 | 1.471 | 1.959 | 4.095 | 2.307 | 1.293 | 1.428 | 1.766 | 3.064 | 1.821 |
| $\dfrac{\text{E}\left[\hat{V}_2\left(\hat{\theta}_{pr,bag}\right)\right]}{V\left(\hat{\theta}_{pr,bag}\right)}$ | 1.240 | 0.963 | 1.190 | 1.174 | 1.149 | 1.145 | 1.262 | 1.319 | 2.039 | 1.524 |
| C.P.(C.I.T.) | 0.969 | 0.970 | 0.984 | 0.991 | 0.980 | 0.964 | 0.974 | 0.977 | 0.983 | 0.946 |
| $\text{C.P.}\left(\text{C.I.T.1.}_{bag}\right)$ | 0.979 | 0.983 | 0.995 | 0.998 | 0.995 | 0.974 | 0.980 | 0.988 | 0.998 | 0.976 |
| $\text{C.P.}\left(\text{C.I.T.2.}_{bag}\right)$ | 0.976 | 0.944 | 0.973 | 0.922 | 0.942 | 0.962 | 0.969 | 0.968 | 0.993 | 0.957 |
| Width(C.I.T) | | | | | | | | | | |
| $\text{Width}\left(\text{C.I.T.1.}_{bag}\right)$ | 0.144 | 0.166 | 0.168 | 0.157 | 0.197 | 0.098 | 0.115 | 0.114 | 0.113 | 0.149 |
| $\text{Width}\left(\text{C.I.T.2.}_{bag}\right)$ | 0.139 | 0.134 | 0.131 | 0.085 | 0.140 | 0.093 | 0.108 | 0.099 | 0.092 | 0.136 |

Tables 5.5 and 5.6 summarize the design-based results on the RKM estimator. Again, we observe the efficiency gain by applying the bagging method, and the gain is between 2% and 12%. Both variance estimators of the bagging quantity perform quite well. Both versions of confidence intervals for bagging estimators have actual coverage rates close to 95%, and the confidence intervals using the adjustment factor approach (Var. 2) are slightly shorter than method 1.

**Table 5.5**
**Bias, standard deviation and MSE ratios of RKM estimator and bagging RKM estimator (5.1); population size $N = 2{,}000,$ number of bootstraps $B = 2{,}000,$ and results are from $2{,}000$ simulations**

| | $n = 100,\ k = 50$ | | | | | $n = 200,\ k = 100$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $t$ | **0.5** | **1.5** | **2.5** | **3.5** | **4.5** | **0.5** | **1.5** | **2.5** | **3.5** | **4.5** |
| $\text{bias}\left(\hat{\theta}_{\text{RKM}}\right)$ | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | -0.001 | 0.001 | 0.000 | 0.000 | 0.001 |
| $\text{bias}\left(\hat{\theta}_{\text{RKM},bag}\right)$ | -0.001 | 0.000 | -0.001 | 0.000 | 0.000 | -0.001 | 0.001 | 0.000 | 0.001 | 0.001 |
| $\text{sd}\left(\hat{\theta}_{\text{RKM}}\right)$ | 0.043 | 0.044 | 0.030 | 0.015 | 0.012 | 0.030 | 0.030 | 0.020 | 0.011 | 0.009 |
| $\text{sd}\left(\hat{\theta}_{\text{RKM},bag}\right)$ | 0.042 | 0.042 | 0.028 | 0.014 | 0.012 | 0.030 | 0.029 | 0.019 | 0.011 | 0.009 |
| $\dfrac{MSE_p\left(\hat{\theta}_{\text{RKM},bag}\right)}{MSE_p\left(\hat{\theta}_{\text{RKM}}\right)}$ | 0.965 | 0.911 | 0.877 | 0.914 | 0.917 | 0.976 | 0.928 | 0.917 | 0.918 | 0.981 |

**Table 5.6**
**Relative bias, coverage probability and confidence interval width of bootstrap variance estimators for the RKM estimator (5.1) and unadjusted $\left(\hat{V}_1\right)$ and adjusted $\left(\hat{V}_2\right)$ variance estimators for bagging RKM estimators; simulation setting is the same as in Table 5.5**

| | $n = 100,\ k = 50$ | | | | | $n = 200,\ k = 100$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $t$ | **0.5** | **1.5** | **2.5** | **3.5** | **4.5** | **0.5** | **1.5** | **2.5** | **3.5** | **4.5** |
| $\dfrac{\text{E}\left[\hat{V}_{boot}\left(\hat{\theta}_{\text{RKM}}\right)\right]}{V\left(\hat{\theta}_{\text{RKM}}\right)}$ | 1.081 | 1.192 | 1.078 | 1.082 | 1.078 | 1.016 | 1.045 | 1.138 | 1.121 | 1.016 |
| $\dfrac{\text{E}\left[\hat{V}_1\left(\hat{\theta}_{\text{RKM},bag}\right)\right]}{V\left(\hat{\theta}_{\text{RKM},bag}\right)}$ | 1.115 | 1.324 | 1.183 | 1.198 | 1.156 | 1.038 | 1.138 | 1.223 | 1.210 | 1.062 |
| $\dfrac{\text{E}\left[\hat{V}_2\left(\hat{\theta}_{\text{RKM},bag}\right)\right]}{V\left(\hat{\theta}_{\text{RKM},bag}\right)}$ | 1.087 | 1.117 | 0.962 | 1.042 | 1.019 | 1.009 | 1.083 | 1.106 | 1.118 | 1.002 |
| C.P.(C.I.) | 0.958 | 0.963 | 0.955 | 0.956 | 0.959 | 0.954 | 0.956 | 0.966 | 0.964 | 0.948 |
| $\text{C.P.}\left(\text{C.I.1.}_{bag}\right)$ | 0.958 | 0.968 | 0.958 | 0.967 | 0.964 | 0.958 | 0.964 | 0.970 | 0.970 | 0.956 |
| $\text{C.P.}\left(\text{C.I.2.}_{bag}\right)$ | 0.957 | 0.954 | 0.937 | 0.951 | 0.950 | 0.955 | 0.958 | 0.959 | 0.960 | 0.948 |
| Width(C.I.) | | | | | | | | | | |
| $\text{Width}\left(\text{C.I.1.}_{bag}\right)$ | 0.171 | 0.183 | 0.116 | 0.074 | 0.052 | 0.122 | 0.122 | 0.083 | 0.049 | 0.034 |
| $\text{Width}\left(\text{C.I.2.}_{bag}\right)$ | 0.169 | 0.168 | 0.105 | 0.069 | 0.049 | 0.120 | 0.120 | 0.079 | 0.047 | 0.033 |

In the context of nonsmooth estimators such as those considered here, it is often recommended that one uses a smoothed bootstrap instead of the simple bootstrap in variance estimation. We considered perturbing each resampled observation $y_{hi}^*$ in the $h$-th stratum to obtain,

$$\tilde{y}_{hi}^* = \overline{y}_h + \left(1+\sigma_Z^2\right)^{-1/2}\left(y_{hi}^* - \overline{y}_h + s_h Z^*\right), \tag{5.2}$$

where $\overline{y}_h,\ s_h$ denote the sample mean and standard deviation of the original sample stratum, $y_{hi}^*$ denotes the originally resampled value and $Z^*$ denotes random noise with $Z^* \overset{iid}{\sim} N\left(0, \sigma_Z^2\right)$. The variance of $Z^*$ controls the amount of smoothing. We applied this method to quantile estimation and the proportion below an estimated level, but it did not appear to improve the performance of the estimation procedure. One possible explanation is that noise contamination "jitters" duplicated observations arising from with-replacement sample and stabilizes subsequent variance estimator to some extent. Since we used without-replacement sampling, this problem was already mostly avoided. More careful study is necessary to understand the effect of smoothing in the context.

# 6 Conclusions

In this article, we have explored the use of bagging procedures for nonlinear and non-differentiable survey estimators. We presented theoretical results on bagging estimator both under design-based and model-based framework. The bagging estimator can be treated as the expectation of a two-phase estimator conditioning on the first phase, and this expectation smoothes out "jumps" in the non-differentiable estimator. The empirical study has revealed the potential of bagging non-differentiable survey estimators, and while the relative performance of bagging varies from one scenario to another, the results are certainly promising.

How to estimate the variance of bagged survey estimators remains an open question when the sampling design is a general complex design. We have proposed two ideas for variance estimation for practical use, but further theoretical study of variance estimation under design-based framework is certainly warranted.

# Appendix

## A.1 Design-based theory

Assumptions D.1-D.6 are used to show the design-based results given below (Theorems 3 and 4). Assumption D.1 specifies moment conditions on the study variable $\mathbf{y}_i$, and Assumption D.2 specifies conditions on the second order inclusion probability of the sampling design. Assumption D.3 guarantees that the size of each resample converges to infinity in the limit. Assumption D.4 specifies smoothness conditions on $m(\cdot)$ in the differentiable estimator. Assumptions D.5-D.6 are used to show the design consistency of bagging non-differentiable survey estimators.

(D.1) The study variable $\mathbf{y}_i$ has finite $2+\delta$ population moment for arbitrarily small $\delta > 0,$

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \| \mathbf{y}_i^{2+\delta} \| \ < \ \infty,$$

where each element of $\mathbf{y}_i^{2+\delta}$ is the original element raised to the power of $2+\delta$ and $\|\cdot\|$ denotes Euclidean norm.

(D.2) For all $N,$ $\min_{i\in U_N} \pi_i \geq \pi_N^* > 0,$ where $N\pi_N^* \to \infty,$ and

$$\lim_{N\to\infty} \sup n \cdot \max \left| \pi_{ij} - \pi_i \pi_j \right| < \infty,$$

where $\pi_{ij}$ denotes the joint inclusion probability of elements $i,\ j.$

(D.3) The resampling process generating $A_b$ is SRSWOR of size $k,$ with $k = O\left(n^\kappa\right),$ $\kappa \in (0,1].$ Further, every bootstrap resample of size $k$ is used in calculating the bagged estimator.

(D.4) The function $m(\cdot)$ is differentiable and has nontrivial continuous second derivative in a compact neighborhood of $\boldsymbol{\mu}_N.$

(D.5) The estimator $\hat{\boldsymbol{\lambda}}$ is $\sqrt{n}$-consistent for the population target $\boldsymbol{\lambda}_N,$ $\lim_{N\to\infty}\boldsymbol{\lambda}_N = \boldsymbol{\lambda}_\infty$ and the estimator $\hat{\boldsymbol{\lambda}}$ is a symmetric statistic.

(D.6) The function $h(\cdot)$ is bounded and the population quantity is "compactly differentiable in a weak sense" (Dümbgen 1993). There exists a function $g(\cdot)$ such that,

$$\sup_{\mathbf{s}\in C_\mathbf{s}} \left| \frac{1}{N}\sum_{i=1}^{N} h\left(\mathbf{y}_i - \boldsymbol{\lambda}_\infty - N^{-\alpha}\mathbf{s}\right) - \frac{1}{N}\sum_{i=1}^{N} h\left(\mathbf{y}_i - \boldsymbol{\lambda}_\infty\right) - g\left(\boldsymbol{\lambda}_\infty\right) N^{-\alpha}\mathbf{s} \right| \to 0,$$

where $C_\mathbf{s}$ is a large enough compact set in $\mathbb{R}^p,$ $0 < \alpha \leq 1/2$ and $g\left(\boldsymbol{\lambda}_\infty\right)$ is bounded.

The following theorem gives several asymptotic approximations for the bagged estimator, depending on the rate of convergence of $k$ relative to $n.$ In all three cases, the bagged estimator is design consistent. Intuitively speaking, the bagging estimator behaves like the original estimator when the resample size $k$ is large (approaches infinity no slower than $n^{1/2}$), but converges at a different speed when the resample size is small.

**Theorem 3** *Under Assumptions D.1-D.4, the bagged differentiable estimator* $\hat{\theta}_{d,bag}$ *admits the following second-order expansion,*

$$\hat{\theta}_{d,bag}-\theta_d=\begin{cases}\left\{m'(\boldsymbol{\mu}_N)\right\}^T(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_N)+o_p\left(n^{-1/2}\right), & \text{for } \kappa>1/2\\ \left\{m'(\boldsymbol{\mu}_N)\right\}^T(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_N)+ & \\ \qquad \dfrac{1}{2\binom{n}{k}}\sum\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)-\boldsymbol{\mu}_N\right)^T m''(\boldsymbol{\mu}_N)\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)-\boldsymbol{\mu}_N\right)+o_p\left(n^{-1/2}\right), & \text{for } \kappa=1/2\\ \dfrac{1}{2\binom{n}{k}}\sum\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)-\boldsymbol{\mu}_N\right)^T m''(\boldsymbol{\mu}_N)\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)-\boldsymbol{\mu}_N\right)+o_p\left(k^{-1}\right), & \text{for } \kappa<1/2\end{cases}$$

*where* $\kappa>0$ *is such that the resample size* $k=O\left(n^{\kappa}\right)$.

**Proof of Theorem 3:**

The proof easily follows from a Taylor expansion of the individual resample-based estimator $m\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)\right)$ around $\boldsymbol{\mu}_N$. The linear expansion term reduces to $\left\{m'(\boldsymbol{\mu}_N)\right\}^T(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_N)$ based on an earlier argument. Under D.1 and D.3, the quadratic term has the same order as the SRSWOR variance of $\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)$ and hence is $o_p(1/k)$.

Next, Theorem 4 gives the design consistency of the non-differentiable bagged estimator.

**Theorem 4** *Under Assumptions D.1-D.3 and D.5-D.6, the bagged non-differentiable estimator* $\hat{\theta}_{nd,bag}$ *is design consistent for its population target* $\theta_{nd}$, *i.e.,* $\hat{\theta}_{nd,bag}-\theta_{nd}=o_p(1)$.

**Proof of Theorem 4:**

We can establish that $(1/N)\sum_{i\in A}(1/\pi_i)h(\mathbf{y}_i-\boldsymbol{\lambda}_N)$ is design consistent for $\theta_{nd}$ as a result of D.2 and the fact that $h(\cdot)$ is bounded (D.6). Then it suffices to show that $\hat{\theta}_{nd,bag}-(1/N)\sum_{i\in A}(1/\pi_i)h(\mathbf{y}_i-\boldsymbol{\lambda}_N)=o_p(1)$, or

$$\frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i}\left\{\frac{1}{\binom{n-1}{k-1}}\sum_{A_b\ni i}h\left(\mathbf{y}_i-\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right)-h(\mathbf{y}_i-\boldsymbol{\lambda}_N)\right\}=o_p(1)$$

following (2.6). We can establish that the collection of resample-based estimators $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$ are uniformly contained in a neighborhood of $\boldsymbol{\lambda}_N$, or, $\sup_{A_b}\left|\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)-\boldsymbol{\lambda}_N\right|=O\left(N^{-\alpha}\mathbf{s}\right)$ for some $\alpha>0$. Then we can apply D.6 to conclude the design consistency of the bagging estimator.

## A.2 Model-based theory

Assumptions M.1-M.4 are used to show the model-based results (Theorems 1 and 2). Assumption M.1 specifies superpopulation distribution of population characteristics $\mathbf{y}_i$. Assumptions M.2 and M.3 assume simple random without replacement sampling for both the design and the resampling process. Assumption M.5 is needed for showing the model-based asymptotic results for the bagging estimator defined by estimating equations.

(M.1) The sequence of population characteristics $\mathbf{y}_i$ constitute an *iid* sample from a probability distribution with density $f_Y(\mathbf{y})$.

(M.2) The sampling design is ignorable, or equivalently, the sampled and unsampled observations are subject to the same distribution.

(M.3) The resampling process generating $A_b$ is SRSWOR of size $k$, where the bootstrap sample size $k$ is bounded. Further, every bootstrap resample of size $k$ is used in calculating the bagged estimator.

(M.4) The function $h(\cdot)$ is bounded.

(M.5) Let $S_\infty(\gamma) = \mathrm{E}\psi(y_i - \gamma)$ be a continuous function of $\gamma$, and $\theta_{ee,\infty}$ be the smallest root of $S_\infty(\gamma) = 0$; for an arbitrary $y$ in the support of the random variable $y_i$, the quantity

$$\inf\left\{\gamma : \frac{1}{k}\sum_{i=1}^{k-1}\psi(y_i - \gamma) + \frac{1}{k}\psi(y - \gamma) \geq 0\right\}$$

belongs to a compact set with probability 1.

**Proof of Theorem 1:**

The bagging estimator $\hat{\theta}_{nd,bag}$ is a symmetric statistic, provided that $\hat{\boldsymbol{\lambda}}$ is symmetric (Lee 1990). We can project it onto a single dimension, say, $\mathbf{y}_1$. But projections onto other observations are equivalent due to symmetry,

$$\begin{aligned}
&\mathrm{E}\left\{\hat{\theta}_{nd,bag} \middle| \mathbf{y}_1 = \mathbf{y}\right\} \\
&=\mathrm{E}\left\{\frac{1}{n}\frac{1}{\binom{n-1}{k-1}}\sum_{A_b \ni 1} h\left(\mathbf{y}_1 - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right) \middle| \mathbf{y}_1 = \mathbf{y}\right\} + \mathrm{E}\left\{\frac{n-1}{n}\frac{1}{\binom{n-1}{k-1}}\sum_{A_b \ni \{i,1\}, i \neq 1} h\left(\mathbf{y}_1 - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right) \middle| \mathbf{y}_1 = \mathbf{y}\right\} \\
&=\frac{1}{n}u(\mathbf{y}) + \frac{k-1}{n}v(\mathbf{y}).
\end{aligned}$$

Then we can derive the following linearization of bagging estimator using the theory of symmetric statistics,

$$\hat{\theta}_{nd,bag} - \theta_{nd,\infty} = \frac{1}{n}\sum_{i=1}^{n}\left\{u\left(\mathbf{y}_i\right) - \theta_{nd,\infty}\right\} + \frac{k-1}{n}\sum_{i=1}^{n}\left\{v\left(\mathbf{y}_i\right) - \theta_{nd,\infty}\right\} + o_p\left(n^{-1/2}\right),$$

where $u(\cdot)$, $v(\cdot)$ and $\theta_{nd,\infty}$ are defined in Theorem 1. The asymptotic variance (3.3) can be easily derived given the $iid$ sampling assumption.

**Proof of Theorem 2:**

The bagged estimator defined in (2.7) can be treated as a one-sample $k$-th order U-statistic, with kernel function

$$h\left(y_1, y_2, \ldots, y_k\right) = \inf\left\{\gamma : \frac{1}{k}\sum_{i=1}^{k}\psi\left(y_i - \gamma\right) \geq 0\right\}.$$

We can directly apply a well-known formula for linearizing U-statistic (Serfling 1980 and van der Vaart 1998, p. 161) to obtain the linearization

$$\hat{\theta}_{ee,bag} - \theta_{ee,\infty} = \frac{k}{n}\sum_{i=1}^{n}\left\{u\left(y_i\right) - \theta_{ee,\infty}\right\} + o_p\left(n^{-1/2}\right),$$

where

$$\begin{aligned} u\left(y\right) &= \mathrm{E}\ h\left(y, y_1, y_2, \ldots, y_{k-1}\right) \\ &= \mathrm{E}\ \inf\left\{\gamma : \frac{1}{k}\sum_{i=1}^{k-1}\psi\left(y_i - \gamma\right) + \frac{1}{k}\psi\left(y - \gamma\right) \geq 0\right\}. \end{aligned}$$

The bagged estimating equation estimator (2.7) can be linearized as

$$\hat{\theta}_{ee,bag} - \theta_{ee,\infty} = \frac{k}{n}\sum_{i=1}^{n}\left\{u\left(y_i\right) - \theta_{ee,\infty}\right\} + o_p\left(n^{-1/2}\right). \tag{A.1}$$

The asymptotic variance of $\hat{\theta}_{ee,bag}$ can be directly obtained from linearization (A.1).

# References

Agresti, A. (2002). *Categorical Data Analysis*. Second Edition, New York: John Wiley and Sons.

Bahadur, R.R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics*, 37, 577-580.

Beran, R. and Srivastava, M. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *The Annals of Statistics*, 13, 95-115.

Berger, Y.G. and Skinner, C.J. (2003). Variance estimation for a low income proportion. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 52 (4), 457-468.

Bickel, P., Götze, F. and van Zwet, W. (1997). Resampling fewer than $n$ observations: gains, losses and remedies for losses. *Statistica Sinica*, 7, 1-31.

Breidt, F. and Opsomer, J. (2008). Endogenous post-stratification in surveys: classifying with a sample-fitted model. *Annals of Statistics*, 36, 403-427.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123-140.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, 30 (4), 927-961.

Buja, A. and Stuetzle, W. (2006). Observations on bagging. *Statistica Sinica*, 16 (2), 323-351.

Chen, S.X. and Hall, P. (2003). Effects of bagging and bias correction on estimators defined by estimating equations. *Statistica Sinica*, 13 (1), 97-109.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, 95, 125-140.

Dunstan, R. and Chambers, R.L. (1986). Model-based confidence intervals in multipurpose surveys. *Applied Statistics*, 35, 276-280.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7, 1-26.

Francisco, C.A. and Fuller, W.A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics*, 19, 454-469.

Frees, E.W. (1989). Infinite order U-statistics. *Scandinavian Journal of Statistics*, 16, 29-45.

Friedman, J.H. and Hall, P. (2007). On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, 137 (3), 669-683.

Fuller, W. (2009). *Sampling Statistics*. John Wiley and Sons.

Godambe, V. and Thompson, M. (2009). Estimating functions and survey sampling. In C. Rao and D. P. (editors) (Eds.), *Handbook of Statistics, vol. 29: Sample Surveys: Inference and Analysis*, 669-687. Elsevier/North-Holland.

Hall, P. and Robinson, A. (2009). Reducing variability of crossvalidation for smoothing parameter choice. *Biometrika*, 96, 175-186.

Heilig, C.M. (1997). *An Empirical Process Approach to U-processes of Increasing Degree*. Ph. D. thesis, University of California, Berkeley.

Heilig, C.M. and Nolan, D. (2001). Limit theorems for the infinite-degree U-process. *Statistica Sinica*, 11, 289-302.

Inoue, A. and Kilian, L. (2008). How useful is bagging in forecasting economic time series? A case study of U.S. consumer price inflation. *Journal of the American Statistical Association*, 103 (482), 511-522.

Knight, K. and Bassett, J.G. (2002). Second order improvements of sample quantiles using subsamples. Unpublished manuscript.

Korn, E.L. and Graubard, B.I. (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology*, 24 (2), 193-201.

Lee, A.J. (1990). *U-statistics: Theory and Practice*. Marcel Dekker Inc.

Lee, T.-H. and Yang, Y. (2006). Bagging binary and quantile predictors for time series. *Journal of Econometrics*, 135 (1-2), 465-497.

Rao, J., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.

Rao, J. and Wu, C. (1988). Resampling inference with complex surveys. *Journal of the American Statistical Association*, 83 (401), 231-241.

Särndal, C.-E., Swensson, B. and Wretman, J.H. (1997). *Model Assisted Survey Sampling*. Springer-Verlag Inc (Berlin; New York).

Schick, A. and Wefelmeyer, W. (2004). Estimating invariant laws of linear processes by U-statistics. *The Annals of Statistics*, 32, 603-632.

Sering, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley and Sons.

Shao, J. and Rao, J. (1993). Standard errors for low income proportions estimated from stratified multistage samples. *Sankhya B,* 55, 393-414.

Sitter, R.R. and Wu, C. (2001). A note on Woodruff confidence intervals for quantiles. *Statistics & Probability Letters*, 52 (4), 353-358.

van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Wang, J.C. and Opsomer, J.D. (2011). On the asymptotic normality and variance estimation of nondifferentiable survey estimators. *Biometrika*, 98, 91-106.