

# ESTIMATING THE DISTANCE DISTRIBUTION OF SUBPOPULATIONS FOR A LARGE-SCALE COMPLEX SURVEY

BY JIANQIANG WANG\*, JEAN D. OPSOMER

*Iowa State University and Colorado State University*

Many finite populations targeted by sample surveys comprise a relatively small number of homogenous subpopulations. In on-going survey operations, it is often of interest to be able to assess whether a new observation belongs to one of those subpopulations or should be flagged as not belonging to any of them. Because the homogeneity of the subpopulations depends on potentially large numbers of survey variables interacting in complex ways, we define a distance measure in the space induced by the survey variables, and consider the distribution of these distances within the subpopulation as a summary of the distributional characteristics of the subpopulation. In this article, we propose a sample-based estimator for the subpopulation distribution functions of the distances between the elements and the subpopulation centers. We allow for a general distance measure, and consider both multivariate means and medians as centers. We describe the design-based asymptotic properties of the estimator under weak assumptions on the finite population. We investigate several approaches for design-based variance estimation, including a combination of kernel regression and replication variance estimation. The practical properties of the procedures are evaluated in a simulation study.

**1. Introduction.** A common issue in large-scale complex surveys is the detection of outliers in the data. Such outliers can be caused by frame imperfections, which can lead to ineligible units being selected, or by errors during data collection. If these outliers remain in the survey dataset, they can cause inference based on the survey to be invalid for the population of interest. Most survey operations therefore incorporate data editing and validation as part of the post-data collection steps, where they attempt to identify suspicious observations and either remove or correct them. Lyberg et al. (1997) contains a collection of articles discussing data collection

---

\*This work was supported by Cooperative Agreement No. 68-3A75-4-122 between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University.

*AMS 2000 subject classifications:* Primary 62D05, 62G05; secondary 62G09

*Keywords and phrases:* jackknife, kernel estimation, estimating equations, generalized median, elliptical distribution

and post survey processing to improve data quality. When outliers exhibit “extreme” values on one or several survey variables, they can be detected relatively easily. However, because surveys often collect large numbers of variables, there is the potential for other outliers which are not extreme on any single variable. Detecting such outliers is more difficult, and identifying unusual or suspicious patterns in the data often requires substantial subject-matter knowledge. For literature on detecting statistical outliers in general, the readers are referred to Barnett and Lewis (1994) and Hawkins (1980).

In practice, many finite populations targeted by surveys consist of a number of relatively homogeneous subpopulations, and this structure can be exploited to develop a statistical approach for flagging suspicious observations that does not require detailed knowledge of the relationships between the variables. The type of surveys we are targeting here is one in which recurring surveys are made over time, and the characteristics and composition of the subpopulations remains relatively stable between surveys. An example of such a survey is the National Resources Inventory (NRI) conducted by the US Natural Resources Conservation Service. The NRI is a longitudinal survey of all non-federally owned land in the US and territories. Its primary mode of data collection is a combination of aerial photography and photo-interpretation, and the main variables in the NRI measure various aspects of land use, farming practices and environmentally important variables such as wetland status and erosion. Because the survey is longitudinal, the NRI measures both the level and the change over time in these variables. Many of these variables are related to each other, so that it is indeed possible to identify broadly similar “clusters” within the population. We refer to Nusser and Goebel (1997) for further information about the NRI.

In a survey context like the one just described, we have the opportunity to develop a characterization of the subpopulations based on past years’ surveys, and develop a method with which to flag suspicious observations in the current year’s survey, potentially as early as while the data are being collected. For the method to be practical, it should be relatively easy to compute, yet account for complex interactions between the survey variables, not require significant input from subject-matter experts, and allow for statistical analysis.

The focus of the current article is to propose an estimator for the “outlyingness” of an observation relative to a subpopulation, based on the “distance” between an observation and the “center” of the subpopulation (all these terms will be more precisely described below), and to derive its statistical properties in an asymptotic design-based context. In doing this, we will assume that the subpopulations have been identified based on prior years’

surveys, in the sense that each observation in those surveys has been assigned to a subpopulation. This subpopulation identification problem and the implementation of a complete outlier detection protocol for the NRI are further described in Wang (2008).

In order to construct the proposed estimator, we consider each element in the population as located in a multidimensional space, with each dimension corresponding to one of the variables being measured in the survey. In this space, it is possible to define a distance measure and to compute the distance between any two locations. An example of such distance measure is the Minkowski distance, with special cases as Euclidean distance and Manhattan distance. Mahalanobis distance can also be used, which takes the shape of subpopulation into account. The approach we will describe later in the paper is valid for general distance measures, so that it is possible to customize the measure for the particular survey under investigation.

A simple definition for the center of a subpopulation is the arithmetic mean vector among all the elements who belong to that subpopulation. Another definition, which we will also consider in this article, is the multidimensional median for the subpopulation. The median for a multivariate variable has been explored in the literature of multivariate statistics. Brown (1983) discusses the spatial median as a location measure minimizing the Euclidean distance in the context of spatial data, and Small (1990) gives a review of multidimensional medians.

Given a distance measure and a subpopulation center definition, one could in principle compute the distance between each observation in the subpopulation and its corresponding center, and construct the “distance distribution function” for the subpopulation. The set of finite population-level distance distribution functions represents our target measure of outlyingness: for a new observation that needs to be assigned to one of the subpopulations, we would like to compute its distance relative to each of the subpopulation centers and obtain the corresponding tail probabilities on the subpopulation distribution functions for this observation. If the probabilities are small for all subpopulations, we conclude that this observation is unlikely to belong to any of the subpopulations and flag it for further investigation. Note that the term “probability” in this context refers to the distribution of elements in the finite population, and does not assume the existence of a probability model for the data.

The target measure just described is infeasible, because we do not have access to the full population and hence neither the subpopulation centers nor the subpopulation distance distributions functions are available. We will therefore estimate both quantities, and the resulting measure of outlying-

ness will be the sample-based estimated probabilities of belonging to each of the subpopulations, as measured by the estimated distance distribution functions.

The estimation of distribution functions using survey data has been well explored in the literature. Dunstan and Chambers (1986) offered a model-based perspective and proposed an estimator under a ratio model with known variance. Rao et al. (1990) furnished a thorough treatment of estimators of distribution functions. They proposed a ratio estimator, a difference estimator and another estimator that is asymptotically both design-unbiased and model-unbiased (often referred to as the RKM estimator). Design normality of distribution function and quantile estimators was proved by Francisco and Fuller (1991) under stratified cluster sampling. The properties of the Dunstan and Chambers estimator and the RKM estimator are further investigated by Chambers et al. (1992). Dorfman (2007) provides an overall review of estimating distributions and quantiles in the survey context.

While our approach is similarly based on estimating (sub)population distribution functions, the results of the above authors are not directly applicable to our situation, because we are considering the distribution of distances from a center that is itself estimated. In order to obtain valid inference results for our estimator, we need to incorporate the uncertainty in estimation of the subpopulation centers as well as that in estimating the distribution function. A critical technical difficulty is that the distribution function estimator is not a smooth function with respect to the subpopulation center, so that the standard tools from classical design-based asymptotic inference cannot be used. This substantially complicates the derivations of the asymptotic results, and also requires us to investigate alternative methods for variance estimation.

A number of methodological innovations are introduced to handle these complications, which might be useful in other survey estimation contexts. First, we provide a method of proof for the design consistency and asymptotic normality of a non-smooth function of a survey estimator, under the assumption that the non-smooth function has a smooth limit. The method of proof is broadly similar to that of Randles (1982), who considered  $U$ -statistics with estimated parameters, and that of Breidt and Opsomer (2008). Unlike in our situation, their results are derived in a model-based context.

Second, in order for our design-based results to hold, a number of regularity conditions are required to hold for the sequence of finite populations (among others, the smooth limit of the non-smooth function mentioned above). As a complement to stating those as population-level assumptions, we show that these regularity conditions hold with probability one under a

set of model assumptions. In other words, we show that the assumed behavior of the finite population sequence is “reasonable” in the sense that deviations from that behavior only happen on a set of (model) probability 0 under stated model assumptions. This explicit connection between sufficient model assumptions and the finite population assumptions is to our knowledge not commonly used in design-based inference, but it provides a formal assessment of the generality of the population specifications, and hence of the subsequent design-based results.

Third, we use a novel combination of nonparametric regression and replication to estimate the design-based variance of the estimated subpopulation distance distribution functions. A kernel-based method is used to obtain an estimate of the derivative of the limiting smooth function mentioned above, and is combined with a traditional design-based replication method such as jackknife to estimate the variance. The method is straightforward to implement, because the replicates do not require recomputing the distances between the observations and the estimated center, and the nonparametric regression step does not have to be repeated across the replicates.

The remainder of the paper is as follows. Section 2 defines notation and establishes preliminary results for design-based inference. Section 3 lists our general design assumptions and assumptions on the sequence of finite populations. Section 4 presents our theoretical results showing asymptotic properties of distance distribution functions using either means or medians as subpopulation centers. Section 5 discusses simulation results to evaluate the finite sample performance our estimators and variance estimators.

**2. Distance Distribution Functions and Estimators.** For the theoretical study of our estimators, we will follow the framework of Isaki and Fuller (1982) in which the properties of estimators are established under a fixed sequence of populations and a corresponding sequence of random samples. Suppose therefore that we have an increasing sequence of finite populations  $(U_\nu)_{\nu=1}^\infty$  of sizes  $N_\nu$  with  $N_\nu < N_{\nu+1}$ . Associated with the  $i$ -th population element is a  $p$ -dimensional vector of observations

$$(1) \quad \mathbf{y}_i = (y_{i,1}, \dots, y_{i,p}),$$

and let  $\mathcal{F}_\nu$  be the power set of  $\nu$ -th finite population  $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{N_\nu}\}$ . The finite population  $U_\nu$  can be partitioned into  $G$  subpopulations, with  $U_\nu = \bigcup_{g=1}^G U_{\nu g}$ . The knowledge of this structure of finite population is usually available for longitudinal surveys of stable populations, and we assume the partition into subpopulations is provided. But in the process of deriving the theoretical result, it is essentially the same as assuming one population, so

for simplicity we will work with one population in the theoretical portions of this article.

We take a sample  $\mathcal{S}_\nu$  of size  $n_\nu$  from population  $U_\nu$ , and the sampling design generating  $\mathcal{S}_\nu$  may be a complex design with stratification or multi-stage sampling. We let  $\pi_i = P(i \in \mathcal{S}_\nu)$  and  $\pi_{ij} = P(i, j \in \mathcal{S}_\nu)$  denote the one-way and two-way inclusion probabilities under this sampling design, where we suppress the dependence on  $\nu$  of both quantities. Because in general the sample size  $n_\nu$  is random, we use  $n_\nu^* = E(n_\nu | \mathcal{F}_\nu)$  to denote the expected sample size over repeated sampling from  $U_\nu$ .

Let  $\|\cdot\|$  denote a norm in the space defined by the  $\mathbf{y}_i$  in (1). We define the population-level distance distribution function as

$$(2) \quad D_{\nu,d}(\boldsymbol{\gamma}_\nu) = \frac{1}{N_\nu} \sum_{U_\nu} \mathbf{I}(\|\mathbf{y}_i - \boldsymbol{\gamma}_\nu\| \leq d),$$

where  $\boldsymbol{\gamma}_\nu$  is some measure of center of  $U_\nu$ . Given a sample  $\mathcal{S}_\nu$ ,  $D_{\nu,d}(\boldsymbol{\gamma}_\nu)$  is estimated by the Hajek-type estimator

$$(3) \quad \widehat{D}_{\nu,d}(\widehat{\boldsymbol{\gamma}}_\nu) = \frac{1}{\widehat{N}_\nu} \sum_{\mathcal{S}_\nu} \frac{1}{\pi_i} \mathbf{I}(\|\mathbf{y}_i - \widehat{\boldsymbol{\gamma}}_\nu\| \leq d),$$

where  $\widehat{\boldsymbol{\gamma}}_\nu$  is an estimator of  $\boldsymbol{\gamma}_\nu$ , some measure of the center of subpopulation  $g$  formed from sample data  $\mathcal{S}_\nu$  and  $\widehat{N}_\nu = \sum_{i \in \mathcal{S}_\nu} \frac{1}{\pi_i}$  is an estimator of the (sub)population size. For future reference, we also define a Horvitz-Thompson version of  $\widehat{D}_{\nu,d}(\widehat{\boldsymbol{\gamma}}_\nu)$  in which  $N_\nu$  is known as

$$(4) \quad \widetilde{D}_{\nu,d}(\widehat{\boldsymbol{\gamma}}_\nu) = \frac{1}{N_\nu} \sum_{\mathcal{S}_\nu} \frac{1}{\pi_i} \mathbf{I}(\|\mathbf{y}_i - \widehat{\boldsymbol{\gamma}}_\nu\| \leq d).$$

In equation (2), we assume  $\boldsymbol{\gamma}_\nu$  is a nonrandom sequence of finite population centers but  $\widehat{\boldsymbol{\gamma}}_\nu$  is random due to sampling mechanism. Quantities  $D_{\nu,d}(\boldsymbol{\gamma}_\nu)$  and  $\widehat{D}_{\nu,d}(\widehat{\boldsymbol{\gamma}}_\nu)$  are step functions of  $d$  and  $\boldsymbol{\gamma}$  with jumps of size  $O(\frac{1}{N_\nu})$  and  $O_p(\frac{1}{n_\nu^*})$ , respectively, which will significantly complicate the study of their theoretical properties.

As a measure of center  $\boldsymbol{\gamma}_\nu$  in equations (2)-(4), we can use the usual mean vector

$$(5) \quad \boldsymbol{\mu}_\nu = \frac{1}{N_\nu} \sum_{U_\nu} \mathbf{y}_i,$$

which is estimated by,

$$(6) \quad \widehat{\boldsymbol{\mu}}_\nu = \frac{1}{\widehat{N}_\nu} \sum_{\mathcal{S}_\nu} \frac{\mathbf{y}_i}{\pi_i}.$$

Alternatively, we can also use a generalized version of multivariate median. The definition of spatial median was given in Brown (1983) and Small (1990). We generalize this idea and define the multivariate median of a finite population to be the location with smallest overall distance to all population units with respect to some norm  $\|\cdot\|$ . For a given norm, we define the median for population  $U_\nu$  as

$$(7) \quad \mathbf{q}_\nu = \arg \inf_{\gamma} \sum_{U_\nu} \|\mathbf{y}_i - \gamma\|,$$

and the sample-based estimator of  $\mathbf{q}_\nu$  is defined as

$$(8) \quad \hat{\mathbf{q}}_\nu = \arg \inf_{\gamma} \sum_{i \in S_\nu} \frac{1}{\pi_i} \|\mathbf{y}_i - \gamma\|.$$

In what follows, we will use  $\gamma_\nu$  to denote a general measure of population center, which can be mean vector  $\boldsymbol{\mu}_\nu$  or median  $\mathbf{q}_\nu$ , and  $\hat{\gamma}_\nu$  to denote the estimator of  $\gamma_\nu$ .

While the distribution function estimator in (3) can be of interest in its own right, our primary purpose is to use it in the detection of potential survey outliers. To do so, we consider a (possibly future) observation as a location  $\mathbf{y}^0$  in  $\mathfrak{R}^p$ , and we would like to evaluate the distance between  $\mathbf{y}^0$  and the population center,  $\|\mathbf{y}^0 - \gamma_\nu\|$ , in relation to the distribution of distances in the population. Hence, for any location  $\mathbf{y}^0 \in \mathfrak{R}^p$ , we define a measure of outlyingness of this location with respect to the finite population as

$$(9) \quad \Omega_\nu(\mathbf{y}^0) = \frac{1}{N_\nu} \sum_{U_\nu} \mathbf{I}_{(\|\mathbf{y}_i - \gamma_\nu\| \leq \|\mathbf{y}^0 - \gamma_\nu\|)}.$$

The population outlyingness measure is generally not known and it is our target population parameter. We propose a sample estimator of  $\Omega_\nu(\mathbf{y}^0)$  defined as

$$(10) \quad \hat{\Omega}_\nu(\mathbf{y}^0) = \frac{1}{\hat{N}_\nu} \sum_{S_\nu} \frac{1}{\pi_i} \mathbf{I}_{(\|\mathbf{y}_i - \hat{\gamma}_\nu\| \leq \|\mathbf{y}^0 - \hat{\gamma}_\nu\|)}.$$

The closer this outlyingness measure is to one for a given observation, the more suspicious it is with respect to the finite population. One possible way to use this in a particular survey is to define a cutoff value  $\alpha^*$  and flag all observations  $k$  with value of  $\hat{\Omega}_\nu(\mathbf{y}_k) > 1 - \alpha^*$  for closer scrutiny.

**3. Assumptions.** In this paper, we estimate the distribution of population distances under a design-based framework. We assume the sequence of finite populations to be fixed and randomness only comes from the sampling mechanism. We do not want to restrict our attention to a specific sampling design but make rather general assumptions to cover various sampling schemes. In Assumption 3.1, we assume that the sample size grows with increasing population size, with a lower bound on its rate of growth for technical convenience. Assumptions 3.2 and 3.3 ensure the design consistency and asymptotic normality of our estimator.

ASSUMPTION 3.1. *The sample size  $n_\nu$  can be fixed or random for any  $\nu$ , but we require  $n_\nu = O_p(N_\nu^\beta)$  with  $\beta \in (\frac{2p}{2p+1}, 1]$ .*

ASSUMPTION 3.2. *The following conditions hold for inclusion probabilities  $\pi_i$  and design variance of Horvitz-Thompson estimator of the mean,*

1.  $K_L \leq \frac{N_\nu}{n_\nu} \pi_i \leq K_U$  for all  $i$ , where  $K_L$  and  $K_U$  are positive constants.
2. For any vector  $\mathbf{z}$  with finite  $2 + \delta$  moments, define  $\bar{\mathbf{z}}_{\nu,\pi} = \frac{1}{N_\nu} \sum_{S_\nu} \frac{\mathbf{z}_i}{\pi_i}$  as the Horvitz-Thompson estimator of  $\bar{\mathbf{z}}_\nu = \frac{1}{N_\nu} \sum_{U_\nu} \mathbf{z}_i$ . We assume

$$\text{Var}(\bar{\mathbf{z}}_{\nu,\pi} | \mathcal{F}_\nu) \leq K_1 \text{Var}_{SRS}(\bar{\mathbf{z}}_{\nu,\pi} | \mathcal{F}_\nu),$$

for some constant  $K_1$ , where  $\text{Var}_{SRS}(\bar{\mathbf{z}}_{\nu,\pi} | \mathcal{F}_\nu)$  is the design variance-covariance matrix of  $\bar{\mathbf{z}}_{\nu,\pi}$  under simple random sampling of size  $n_\nu^*$ .

It is readily shown that under Assumption 3.2(2),  $\frac{n_\nu}{n_\nu^*} \xrightarrow{p} 1$  by bounding its design variance.

ASSUMPTION 3.3. *For any  $\mathbf{z}$  with positive variance-covariance matrix and finite fourth population moment,*

$$(11) \quad n_\nu^{*1/2} (\bar{\mathbf{z}}_{\nu,\pi} - \bar{\mathbf{z}}_\nu) | \mathcal{F}_\nu \xrightarrow{d} N(\mathbf{0}, \Sigma_{\mathbf{z}\mathbf{z}}),$$

and

$$(12) \quad [V(\bar{\mathbf{z}}_{\nu,\pi} | \mathcal{F}_\nu)]^{-1} \widehat{V}_{HT}\{\bar{\mathbf{z}}_{\nu,\pi}\} - \mathbf{I}_{p \times p} = O_p(n_\nu^{*-1/2}),$$

where  $\Sigma_{\mathbf{z}\mathbf{z}}$  is a nonnegative definite matrix,  $V(\bar{\mathbf{z}}_{\nu,\pi} | \mathcal{F}_\nu)$  is the design variance-covariance matrix of estimator  $\bar{\mathbf{z}}_{\nu,\pi}$ , and  $\widehat{V}_{HT}\{\bar{\mathbf{z}}_{\nu,\pi}\}$  is the Horvitz-Thompson estimator of the variance of  $\bar{\mathbf{z}}_{\nu,\pi} | \mathcal{F}_\nu$ .

Assumption 3.3 on the normality of the Horvitz-Thompson estimator for a general design and for a general vector with moment conditions is similar to assumptions in Fuller (2007), among others.

We also need to assume a further number of more specific regularity conditions on the sequence of finite populations. Assumption 3.4 specifies population moment conditions on population vectors  $\mathbf{y}_i$ . In Assumption 3.5, we assume the existence of the limiting function of  $D_{\nu g, d}(\boldsymbol{\gamma})$  and that the limiting function satisfies certain smoothness. Assumption 3.6 provides limits for a number of key population quantities needed to derive the design-based results.

ASSUMPTION 3.4. *The sequence of population vectors  $\mathbf{y}_i$ 's has bounded  $4 + \delta$  population moments,*

$$\lim_{N_\nu \rightarrow \infty} N_\nu^{-1} \sum_{i \in U_\nu} |\mathbf{y}_i|^{4+\delta} < \infty,$$

for some  $\delta > 0$ .

ASSUMPTION 3.5. *1. The population-level distance distribution converges to a limiting distance distribution,*

$$\lim_{\nu \rightarrow \infty} D_{\nu, d}(\boldsymbol{\gamma}) = \mathcal{D}_d(\boldsymbol{\gamma})$$

on  $(d, \boldsymbol{\gamma}) \in [0, \infty) \times \mathfrak{R}^p$ .

*2. The limiting function  $\mathcal{D}_d(\boldsymbol{\gamma})$  is continuous in  $d \in [0, \infty)$  and  $\boldsymbol{\gamma} \in \mathfrak{R}^p$ . Additionally, the derivatives  $\frac{\partial \mathcal{D}_d(\boldsymbol{\gamma})}{\partial d}$ ,  $\frac{\partial \mathcal{D}_d(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}$  and  $\frac{\partial^2 \mathcal{D}_d(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^2}$  all exist and are finite in  $(d, \boldsymbol{\gamma}) \in [0, +\infty) \times \mathfrak{R}^p$ .*

ASSUMPTION 3.6. *The following population quantities converge to zero:*

*1.*

$$\sqrt{N_\nu} \left\{ \frac{1}{N_\nu} \sum_{U_\nu} \mathbf{I}_{(d < \|\mathbf{y}_i - \boldsymbol{\gamma}\| \leq d + h_{N_\nu})} - \frac{\partial \mathcal{D}_d(\boldsymbol{\gamma})}{\partial d} h_{N_\nu} \right\} \rightarrow 0,$$

where  $h_{N_\nu} = O(N_\nu^{-\alpha})$  and  $\alpha \in (\frac{1}{4}, 1)$ .

*2. let*

$$(13) \quad R_i = \mathbf{I}_{(\|\mathbf{y}_i - \boldsymbol{\gamma} - n_\nu^{-1/2} \mathbf{s}\| \leq d)} - \mathbf{I}_{(\|\mathbf{y}_i - \boldsymbol{\gamma}\| \leq d)} - \mathcal{D}_d(\boldsymbol{\gamma} + n_\nu^{*-1/2} \mathbf{s}) + \mathcal{D}_d(\boldsymbol{\gamma}),$$

then

$$\frac{n_\nu^{*1/2}}{N_\nu} \sum_{U_\nu} R_i \rightarrow 0,$$

uniformly for  $\boldsymbol{\gamma} \in \mathfrak{R}^p$  and  $\mathbf{s} \in C_{\mathbf{s}}$ , a large enough compact set in  $\mathfrak{R}^p$ .

The reasonableness of the population requirements in Assumptions 3.5 and in particular 3.6 is somewhat difficult to evaluate as stated. Therefore, we justify the assumptions by working under a superpopulation model, under which the  $\mathbf{y}_i$  are generated through a probabilistic mechanism. As further discussed in Appendix A, the statements in these population assumptions are shown to hold almost surely under the superpopulation model based on a set of more readily interpretable model assumptions. We here assume that the specific population sequence from which we are sampling is such that these results hold, without the almost sure condition. Stated differently, we exclude population sequences violating Assumptions 3.5 3.6, but this only happens with model probability 0.

The next set of assumptions are necessary when deriving median-based results, where we need stronger conditions on the distribution of finite population elements and restrictions on the norm. Assumption 3.7 guarantees the uniqueness of population median  $\mathbf{q}_\nu$ , and Assumption 3.8(1) gives smoothness conditions on the norm. Assumption 3.8(2) restricts our concern to an inner product space, and Assumption 3.8(3) will be used in showing asymptotic normality of  $\hat{\mathbf{q}}_\nu$ .

ASSUMPTION 3.7. *There exist no  $(\lambda, \gamma_1, \gamma_2)$ , where  $\lambda \in (0, 1), \gamma_1, \gamma_2 \in \mathbb{R}^p$ , such that the finite population  $U_\nu$  puts all points on*

$$(14) \quad L_{\gamma_1, \gamma_2} = \{\mathbf{y} \mid \|\mathbf{y} - \gamma_1 - \gamma_2\| = \|\lambda\mathbf{y} - \gamma_1\| + \|(1 - \lambda)\mathbf{y} - \gamma_2\|\}.$$

ASSUMPTION 3.8. *The following conditions hold for the norm  $\|\cdot\|$ ,*

1. *The norm  $\|\cdot\|$  is continuous on  $\mathbb{R}^p$ , with a continuous gradient vector  $\psi(\gamma)$ , and bounded second derivative matrix  $H_s(\gamma)$ .*
2. *For any  $\gamma$  in a neighborhood of  $\mathbf{q}_\nu$ ,  $\frac{1}{N_\nu} \sum_{U_\nu} H_s(\mathbf{y}_i - \gamma)$  is a nonsingular matrix. Further, the sequence of  $H_s(\mathbf{y}_i - \gamma_\nu)$  has bounded first two moments at population level.*
3. *The gradient function  $\psi(\mathbf{y}_i - \gamma_\nu)$  has bounded fourth population moments,*

$$\frac{1}{N_\nu} \sum_{U_\nu} |\psi(\mathbf{y}_i - \gamma_\nu)|^4 < \infty.$$

The set  $L_{\gamma_1, \gamma_2}$  in Assumption 3.7 includes all the points  $\mathbf{y}$  such that  $\lambda\mathbf{y} - \gamma_1$  and  $(1 - \lambda)\mathbf{y} - \gamma_2$  are on the same “line” through the origin. The definition of “line” differs from one norm to another, where in  $L_2$  norm, it is the usual definition of a line, but in  $L_1$  norm, it says  $\lambda\mathbf{y} - \gamma_1$  and  $(1 - \lambda)\mathbf{y} - \gamma_2$  are in the same orthant.

Finally, we give some conditions on the kernel function and bandwidth that will be used in constructing a variance estimator, where we will use a kernel regression estimator to estimate the derivative of limiting smooth function as needed in the expression of asymptotic variance.

ASSUMPTION 3.9. *The following conditions hold for the kernel function  $K(t)$  and bandwidth  $h$ ,*

1. *The kernel function  $K(t)$  is symmetric with  $\int_{-\infty}^{\infty} K(t)dt = 1$ , and  $K(t)$  is an absolutely continuous function with finite derivative  $K'(t)$ . Further, let  $R(K) = \int_{-\infty}^{\infty} K^2(t)dt < \infty$  and  $\sigma_K^2 = \int_{-\infty}^{\infty} t^2 K(t)dt < \infty$ .*
2. *The bandwidth  $h \rightarrow 0$ , and  $N_\nu h (\log N_\nu)^{-1} \rightarrow \infty$ , as  $N_\nu \rightarrow \infty$ .*
3. *There exists a constant  $c$ , such that  $\left| \frac{1}{h^2} K' \left( \frac{x}{h} \right) \right| \leq c$ , for any  $x \neq 0$  and  $h$  arbitrarily small.*

#### 4. Main results.

4.1. *Mean-based distance distribution.* In this section, we use the mean vector as a measure of center, and show the asymptotic properties of the sample distance distribution function estimator. The sample distance distribution function (3) is a nonsmooth function of the estimated center, so the uncertainty due to estimating the mean can not be directly quantified through linearization. The idea to tackle the problem is to replace the nonsmooth function by a smooth limiting function, and do linearization on the smooth function. Lemma 1 establishes an intermediate result, which allows us to replace the nonsmooth function by the limiting smooth function, similar to (1.6) of Randles (1982) but in the design-based asymptotic framework. Then we can linearize the estimator (Theorem 1) and derive asymptotic normality of the estimator (Theorem 2).

LEMMA 1. *Under Assumptions 3.1–3.2 and 3.5–3.6,*

$$(15) \quad n_\nu^{*1/2} \left( \widehat{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_\nu) - \widehat{D}_{\nu,d}(\boldsymbol{\mu}_\nu) - \mathcal{D}_{g,d}(\hat{\boldsymbol{\mu}}_\nu) + \mathcal{D}_{g,d}(\boldsymbol{\mu}_\nu) \right) \xrightarrow{p} 0$$

*in design.*

PROOF. See Appendix B. □

THEOREM 1. *Under Assumptions 3.1–3.2 and 3.5–3.6, the sample-based quantity  $\widehat{D}_{\nu,g,d}(\hat{\boldsymbol{\mu}}_\nu)$  is  $\sqrt{n_\nu^*}$ -consistent for the corresponding population quantity  $D_{\nu,g,d}(\boldsymbol{\mu}_\nu)$ , namely,*

$$n_\nu^{*1/2} \left( \widehat{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_\nu) - D_{\nu,d}(\boldsymbol{\mu}_\nu) \right) = O_p(1).$$

Further,

$$(16) \quad \begin{aligned} \widehat{D}_{\nu,d}(\widehat{\boldsymbol{\mu}}_{\nu}) - D_{\nu,d}(\boldsymbol{\mu}_{\nu}) &= \left( \widehat{D}_{\nu,d}(\boldsymbol{\mu}_{\nu}) - D_{\nu,d}(\boldsymbol{\mu}_{\nu}) \right) \\ &+ (\mathcal{D}_d(\widehat{\boldsymbol{\mu}}_{\nu}) - \mathcal{D}_d(\boldsymbol{\mu}_{\nu})) + o_p(n_{\nu}^{*-1/2}). \end{aligned}$$

PROOF. We use the following decomposition,

$$\begin{aligned} &n_{\nu}^{*1/2} \left( \widehat{D}_{\nu,d}(\widehat{\boldsymbol{\mu}}_{\nu}) - D_{\nu,d}(\boldsymbol{\mu}_{\nu}) \right) \\ &= n_{\nu}^{*1/2} \left( \widehat{D}_{\nu,d}(\widehat{\boldsymbol{\mu}}_{\nu}) - \widehat{D}_{\nu,d}(\boldsymbol{\mu}_{\nu}) - \mathcal{D}_{g,d}(\widehat{\boldsymbol{\mu}}_{\nu}) + \mathcal{D}_{g,d}(\boldsymbol{\mu}_{\nu}) \right) \\ &\quad + n_{\nu}^{*1/2} \left( \widehat{D}_{\nu,d}(\boldsymbol{\mu}_{\nu}) - D_{\nu,d}(\boldsymbol{\mu}_{\nu}) \right) + n_{\nu}^{*1/2} (\mathcal{D}_d(\widehat{\boldsymbol{\mu}}_{\nu}) - \mathcal{D}_d(\boldsymbol{\mu}_{\nu})), \end{aligned}$$

where the first term is  $o_p(1)$  by Lemma 1, and the last two terms are both  $O_p(1)$  by assumption 3.2 and standard linearization methods.  $\square$

Equation (16) allows us to obtain the asymptotic distribution of  $\widehat{D}_{\nu,d}(\widehat{\boldsymbol{\mu}}_{\nu})$ . The first term on the RHS contains  $\widehat{D}_{\nu,d}(\boldsymbol{\mu}_{\nu})$ , which is a ratio of two Horvitz-Thompson estimators. In the second term,  $\mathcal{D}_d(\widehat{\boldsymbol{\mu}}_{\nu})$  applies a smooth function to the ratio estimator of population mean, whose variance can be evaluated through linearizing  $\mathcal{D}_d(\cdot)$ .

Assumptions 3.3 and 3.4 implies the following multivariate normality,

$$(17) \quad \frac{n_{\nu}^{*1/2}}{N_{\nu}} \sum_{U_{\nu}} \underbrace{\begin{bmatrix} \mathbb{I}(\|\mathbf{y}_i - \boldsymbol{\mu}_{\nu}\| \leq d) \\ 1 \\ \mathbf{y}_i \end{bmatrix}}_{\mathbf{b}_{\boldsymbol{\mu},i}} \left[ \frac{\mathbb{I}(i \in \mathcal{S}_{\nu})}{\pi_i} - 1 \right] \Bigg| \mathcal{F}_{\nu} \xrightarrow{d} N(0, \Sigma_{\boldsymbol{\mu},d}),$$

where

$$(18) \quad \Sigma_{\boldsymbol{\mu},d} = \frac{n_{\nu}^{*}}{N_{\nu}^2} \sum_{U_{\nu}} \sum_{U_{\nu}} (\pi_{ij} - \pi_i \pi_j) \frac{\mathbf{b}_{\boldsymbol{\mu},i} \mathbf{b}'_{\boldsymbol{\mu},i}}{\pi_i \pi_j}.$$

This leads immediately to the following result.

**THEOREM 2.** *Under Assumptions 3.1–3.6,*

$$n_{\nu}^{1/2} \left[ V \left( \widehat{D}_{\nu,d}(\widehat{\boldsymbol{\mu}}_{\nu}) \right) \right]^{-1/2} \left( \widehat{D}_{\nu,d}(\widehat{\boldsymbol{\mu}}_{\nu}) - D_{\nu,d}(\boldsymbol{\mu}_{\nu}) \right) \Bigg| \mathcal{F}_{\nu} \xrightarrow{d} N(0, 1),$$

where

$$(19) \quad V \left( \widehat{D}_{\nu,d}(\widehat{\boldsymbol{\mu}}_{\nu}) \right) = \mathbf{a}'_{\boldsymbol{\mu}} \Sigma_{\boldsymbol{\mu},d} \mathbf{a}_{\boldsymbol{\mu}},$$

$$(20) \quad \mathbf{a}_\mu = \left( 1, -D_{\nu,d}(\boldsymbol{\mu}_\nu) - \left( \frac{\partial \mathcal{D}_d(\boldsymbol{\mu}_\nu)}{\partial \boldsymbol{\mu}_\nu} \right)' \boldsymbol{\mu}_\nu, \left( \frac{\partial \mathcal{D}_d(\boldsymbol{\mu}_\nu)}{\partial \boldsymbol{\mu}_\nu} \right)' \right)',$$

and  $\Sigma_{\boldsymbol{\mu},d}$  is defined in (18).

PROOF. Use decomposition (16) and Slutsky's Theorem.  $\square$

The asymptotic variance of  $\widehat{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_\nu)$  consists of two components, the first one from using estimator  $\widehat{D}_\nu(\boldsymbol{\mu}_\nu)$  and the second one due to the uncertainty of estimating population center. The first component is essentially the variance of a ratio estimator and can be easily estimated using plug-in estimator or replication procedure, but the second component involves an unknown derivative of limiting smooth function. The derivative can be estimated by kernel smoothing and incorporated into either plug-in or replication estimator. We will discuss estimation of  $V\left(\widehat{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_\nu)\right)$  in Section 5.

4.2. *Median-based distance distribution.* Now let us look at the case where we use median as a measure of subpopulation center. We introduce the following estimating equations at the population and sample level, respectively,

$$(21) \quad \sum_{U_\nu} \boldsymbol{\psi}(\mathbf{y}_i - \boldsymbol{\gamma}) = \mathbf{0},$$

and

$$(22) \quad \sum_{i \in \mathcal{S}_\nu} \frac{\boldsymbol{\psi}(\mathbf{y}_i - \boldsymbol{\gamma})}{\pi_i} = \mathbf{0}.$$

The medians defined by (7)-(8) are solutions to the estimating equations (21)-(22), if the sequence of population medians  $\mathbf{q}_\nu$  are interior points in  $\mathbb{R}^p$ , which is true for nondegenerate populations. Lemma 2 states the uniqueness of  $\mathbf{q}_\nu$  under certain regularity conditions, and is followed by discussion of the uniqueness of sample median  $\hat{\mathbf{q}}_\nu$ . As in the mean-based case above, we obtained model-based results ensuring existence and uniqueness of the population version of Lemma 2, but omit those here for brevity. Readers are referred to Wang (2008) for detailed model-based results.

LEMMA 2. *Under Assumptions 3.7, for a large enough population,  $\sum_{U_\nu} \|\mathbf{y}_i - \boldsymbol{\gamma}\|$  has only one local minimum, which is also its global minimum.*

If we make Assumption 3.7 on the sample  $\mathcal{S}_\nu$ , then it is obvious that  $\sum_{\mathcal{S}_\nu} \frac{\|\mathbf{y}_i - \boldsymbol{\gamma}\|}{\pi_i}$  has a unique global minimizer and no other local minimizers.

But under a complex sampling design, we may have nonzero probability of selecting a sample where all points are on the same line. But considering the increasing sequence of finite populations and sequence of sampling designs, there is high probability of selecting a large enough sample so that it is guaranteed that not all points are on a line.

Although the sequence of sample medians  $\hat{\mathbf{q}}_\nu$  need not be unique, we can show that any sequence of  $\hat{\mathbf{q}}_\nu$  is design consistent for population median  $\mathbf{q}_\nu$  and asymptotically normally distributed, as will be stated and proved in Theorems 3 and 4. Then the distances can be defined using any one of these sequences, but sample distance distribution estimators will remain consistent and asymptotically normally distributed.

In establishing the weak convergence of  $\hat{\mathbf{q}}_\nu$  for  $\mathbf{q}_\nu$ , we adopt the definition of weak convergence of Billingsley (1968, p.24) and use a general norm  $\|\cdot\|$  as a discrepancy measure. Here, we say  $\hat{\mathbf{q}}_\nu$  converges to  $\mathbf{q}_\nu$  weakly if and only if  $\forall \epsilon > 0$ ,

$$(23) \quad \lim_{\nu \rightarrow \infty} P(\|\hat{\mathbf{q}}_\nu - \mathbf{q}_\nu\| > \epsilon) = 0.$$

**THEOREM 3.** *Under Assumptions 3.1–3.2 and 3.7, any sequence  $\hat{\mathbf{q}}_\nu$  that satisfies (8) is design consistent for  $\mathbf{q}_\nu$ .*

**PROOF.** Uses negation, and a key result we need is to show that

$$(24) \quad \sup_{\gamma \in C} |Q_n(\gamma)| \xrightarrow{p} 0,$$

where  $Q_n(\gamma) = \frac{1}{N_\nu} \sum_{i \in \mathcal{S}_\nu} \frac{\|\mathbf{y}_i - \gamma\|}{\pi_i} - \frac{1}{N_\nu} \sum_{U_\nu} \|\mathbf{y}_i - \gamma\|$ , through covering technique, similar to the approach we use to show (48) in Lemma 1 in the appendix.  $\square$

**THEOREM 4.** *Under Assumptions 3.1–3.3, 3.8(1), and assuming that  $\hat{\mathbf{q}}_\nu$  is a design consistent sequence, then we have the following asymptotic normality for the sample median,*

$$(25) \quad n_\nu^{1/2}(\hat{\mathbf{q}}_\nu - \mathbf{q}_\nu) \xrightarrow{d} N_p(\mathbf{0}, \Sigma_{\nu g, \mathbf{q}}),$$

where  $\Sigma_{\nu g, \mathbf{q}} = \mathbf{A} \Sigma_{\nu g, \psi} \mathbf{A}'$ ,  $\mathbf{A} = \left[ \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} H_s(\mathbf{y}_i - \mathbf{q}_\nu) \right]^{-1}$ , and

$$\Sigma_{\nu g, \psi} = \frac{n_\nu}{N_\nu^2} \sum_{U_\nu} \sum_{U_\nu} (\pi_{ij} - \pi_i \pi_j) \frac{\psi(\mathbf{y}_i - \mathbf{q}_\nu)}{\pi_i} \frac{\psi(\mathbf{y}_j - \mathbf{q}_\nu)}{\pi_j}.$$

PROOF. The consistency of  $\hat{\mathbf{q}}_\nu$  for  $\mathbf{q}_\nu$  gives,

$$\begin{aligned} \sum_{i \in \mathcal{S}_\nu} \frac{1}{\pi_i} \boldsymbol{\psi}(\mathbf{y}_i - \hat{\mathbf{q}}_\nu) &= \mathbf{0} \\ \Leftrightarrow \sum_{i \in \mathcal{S}_\nu} \frac{1}{\pi_i} \{ \boldsymbol{\psi}(\mathbf{y}_i - \mathbf{q}_\nu) - H_s(\mathbf{y}_i - \mathbf{q}_\nu)(\hat{\mathbf{q}}_\nu - \mathbf{q}_\nu) + o_p(\hat{\mathbf{q}}_\nu - \mathbf{q}_\nu) \} &= \mathbf{0}, \end{aligned}$$

which implies

$$(26) \quad \hat{\mathbf{q}}_\nu = \mathbf{q}_\nu + \left[ \frac{1}{N_\nu} \sum_{i \in \mathcal{S}_\nu} \frac{H_s(\mathbf{y}_i - \mathbf{q}_\nu)}{\pi_i} \right]^{-1} \frac{1}{N_\nu} \sum_{i \in \mathcal{S}_\nu} \frac{\boldsymbol{\psi}(\mathbf{y}_i - \mathbf{q}_\nu)}{\pi_i} + o_p(\hat{\mathbf{q}}_\nu - \mathbf{q}_\nu)$$

after using the non-singularity condition in Assumption 3.8(2).

It is easy to argue that

$$(27) \quad \left[ \frac{1}{N_\nu} \sum_{i \in \mathcal{S}_\nu} \frac{H_s(\mathbf{y}_i - \mathbf{q}_\nu)}{\pi_i} \right]^{-1} \xrightarrow{p} \left[ \frac{1}{N_\nu} \sum_{U_\nu} H_s(\mathbf{y}_i - \mathbf{q}_\nu) \right]^{-1},$$

and

$$(28) \quad \frac{n_\nu^{*1/2}}{N_\nu} \sum_{i \in \mathcal{S}_\nu} \frac{\boldsymbol{\psi}(\mathbf{y}_i - \mathbf{q}_\nu)}{\pi_i} \xrightarrow{d} N_p(\mathbf{0}, \Sigma_{\nu g, \boldsymbol{\psi}}),$$

where  $\Sigma_{\nu g, \boldsymbol{\psi}} = \frac{n_\nu^*}{N_\nu^2} \sum_{U_\nu} \sum_{U_\nu} (\pi_{ij} - \pi_i \pi_j) \frac{\boldsymbol{\psi}(\mathbf{y}_i - \mathbf{q}_\nu)}{\pi_i} \frac{\boldsymbol{\psi}(\mathbf{y}_j - \mathbf{q}_\nu)}{\pi_j}$ .

The proof is then completed by applying Slutsky's Theorem.  $\square$

**THEOREM 5.** *Suppose Assumptions 3.1–3.2 and 3.7–3.8 are satisfied, then for any sequence  $\hat{\mathbf{q}}_\nu$  that satisfies (8), the estimated distance distribution  $\hat{D}_{\nu, d}(\hat{\mathbf{q}}_\nu)$  is  $\sqrt{n_\nu^*}$ -consistent for the corresponding population quantity  $D_{\nu, d}(\mathbf{q}_\nu)$ , namely,*

$$n_\nu^{*1/2} \left( \hat{D}_{\nu, d}(\hat{\mathbf{q}}_\nu) - D_{\nu, d}(\mathbf{q}_\nu) \right) = O_p(1).$$

Further,

$$(29) \quad \begin{aligned} \hat{D}_{\nu, d}(\hat{\mathbf{q}}_\nu) - D_{\nu, d}(\mathbf{q}_\nu) &= \left( \hat{D}_{\nu, d}(\mathbf{q}_\nu) - D_{\nu, d}(\mathbf{q}_\nu) \right) \\ &+ (\mathcal{D}_d(\hat{\mathbf{q}}_\nu) - \mathcal{D}_d(\mathbf{q}_\nu)) + o_p(n_\nu^{*-1/2}). \end{aligned}$$

PROOF. Similar to the proof of Theorem 2.  $\square$

Assumption 3.8(4) together with Assumption 3.3 gives

$$(30) \quad \frac{n_\nu^*}{N_\nu} \sum_{U_\nu} \left[ \begin{array}{c} \mathbf{I}(\|\mathbf{y}_i - \mathbf{q}_\nu\| \leq d) \\ 1 \\ \underbrace{\psi(\mathbf{y}_i - \mathbf{q}_\nu)}_{\mathbf{b}_{\mathbf{q},i}} \end{array} \right] \left[ \frac{\mathbf{I}(i \in \mathcal{S}_\nu)}{\pi_i} - 1 \right] \Bigg| \mathcal{F}_\nu \xrightarrow{d} N(0, \Sigma_{\mathbf{q},d}),$$

where

$$(31) \quad \Sigma_{\mathbf{q},d} = \frac{n_\nu^*}{N_\nu^2} \sum_i \sum_j (\pi_{ij} - \pi_i \pi_j) \frac{\mathbf{b}_{\mathbf{q},i} \mathbf{b}_{\mathbf{q},j}'}{\pi_i \pi_j}.$$

**THEOREM 6.** *Under Assumptions 3.1-3.3 and 3.7-3.8, for any sequence  $\hat{\mathbf{q}}_\nu$  satisfying (8),*

$$n_\nu^{1/2} \left[ V \left( \hat{D}_{\nu,d}(\hat{\mathbf{q}}_\nu) \right) \right]^{-1/2} \left( \hat{D}_{\nu,d}(\hat{\mathbf{q}}_\nu) - D_{\nu,d}(\mathbf{q}_\nu) \right) \Bigg| \mathcal{F}_\nu \xrightarrow{d} N(0, 1),$$

where

$$(32) \quad V \left( \hat{D}_{\nu,d}(\hat{\mathbf{q}}_\nu) \right) = \mathbf{a}_{\mathbf{q}}' \Sigma_{\mathbf{q},d} \mathbf{a}_{\mathbf{q}},$$

and

$$(33) \quad \mathbf{a}_{\mathbf{q}} = \left( 1, -D_{\nu,d}(\mathbf{q}_\nu), \left( \frac{\partial D_{\nu,d}(\mathbf{q}_\nu)}{\partial \mathbf{q}_\nu} \right)' H_{s,N_\nu}^{-1} \right)',$$

where  $H_{s,N_\nu} = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} H_s(\mathbf{y}_i - \mathbf{q}_\nu)$  and  $\Sigma_{\mathbf{q},d}$  is defined in (31).

**PROOF.** The proof is similar to that of Theorem 2. But as there is no  $\hat{N}_\nu$  in the expansion of  $\hat{\mathbf{q}}_\nu$ , so we do not have a complicated second term in  $\mathbf{a}_{\mathbf{q}}$ .  $\square$

**4.3. Measures of outlyingness.** In this section, we use a general measure of center, either mean vector or generalized median, and show that the estimated outlyingness measure  $\hat{\Omega}_\nu(\mathbf{y}^0)$  is design consistent for  $\Omega_\nu(\mathbf{y}^0)$  and asymptotically normally distributed, where  $\mathbf{y}^0$  is an arbitrary location in  $\mathbb{R}^p$  (these results do not follow directly from those in Sections 4.1 and 4.2, because of the presence of an additional estimated center in  $\hat{\Omega}_\nu(\mathbf{y}^0)$ ). Lemma 3 is similar but a stronger version of Lemma 1, and it leads to the conclusion that quantifying the distance between  $\mathbf{y}^0$  and the population as  $\|\mathbf{y}^0 - \hat{\gamma}_\nu\|$

will not affect the leading term in the asymptotic variance, although there is error in estimating  $\gamma_\nu$ . No proofs are given here because they follow similar arguments as those in Sections 4.1 and 4.2, but details are available in Wang (2008).

LEMMA 3. *Under Assumptions 3.1–3.2, 3.5–3.8,*

(34) 
$$n_\nu^{*1/2} \left( \widehat{D}_{\nu, \|\mathbf{y}^0 - \hat{\gamma}_\nu\|}(\gamma) - \widehat{D}_{\nu, \|\mathbf{y}^0 - \gamma_\nu\|}(\gamma) - \mathcal{D}_{g, \|\mathbf{y}^0 - \hat{\gamma}_\nu\|}(\gamma) + \mathcal{D}_{g, \|\mathbf{y}^0 - \gamma_\nu\|}(\gamma) \right) \xrightarrow{p} 0$$

*uniformly for  $\gamma$  in a neighbourhood of  $\gamma_\nu$ , where the measures of center  $\gamma_\nu$  and  $\hat{\gamma}_\nu$  can be either mean vector or generalized median.*

THEOREM 7. *Under Assumptions 3.1–3.8, for any sequence  $\hat{\gamma}_\nu$  satisfying (6) or (8),*

$$n_\nu^{*1/2} \left[ V \left( \widehat{\Omega}_\nu(\mathbf{y}^0) \right) \right]^{-1/2} \left( \widehat{\Omega}_\nu(\mathbf{y}^0) - \Omega_\nu(\mathbf{y}^0) \right) \Big|_{\mathcal{F}_\nu} \xrightarrow{d} N(0, 1),$$

where

(35) 
$$V \left( \widehat{\Omega}_\nu(\mathbf{y}^0) \right) = \mathbf{a}'_\gamma \Sigma_{\gamma, \|\mathbf{y}^0 - \gamma_\nu\|} \mathbf{a}_\gamma,$$

and  $\mathbf{a}_\gamma$  is defined in (20) or (33), and  $\Sigma_{\gamma, \|\mathbf{y}^0 - \gamma_\nu\|}$  is defined in either (18) or (31) with  $d$  replaced by  $\|\mathbf{y}^0 - \gamma_\nu\|$ .

Finally, we briefly consider an additional extension of the outlier detection estimator, to the situation in which the distance measure is itself sample-dependent. As an example of this, consider a population measure of outlyingness defined as

(36) 
$$\Omega_\nu^*(\mathbf{y}^0) = \frac{1}{N_\nu} \sum_{\mathcal{S}_\nu} \mathbf{I}_{(\|\mathbf{y}_i - \gamma_\nu\|_{\Sigma_\nu} \leq \|\mathbf{y}^0 - \gamma_\nu\|_{\Sigma_\nu})},$$

with

$$\|\mathbf{y}_i - \gamma_\nu\|_{\Sigma_\nu} = \sqrt{(\mathbf{y}_i - \gamma_\nu)' \Sigma_\nu^- (\mathbf{y}_i - \gamma_\nu)}, \Sigma_\nu = \frac{1}{N_\nu - 1} \sum_{i=1}^{N_\nu} (\mathbf{y}_i - \gamma_\nu)(\mathbf{y}_i - \gamma_\nu)',$$

where the generalized inverse of  $\Sigma_\nu$  is used if the population variance-covariance matrix is degenerate. This approach is attractive because the norm adjusts for the distribution and correlation of the survey variables,

which might be beneficial in surveys with numerous related variables. Because  $\Sigma_\nu$  is typically unknown, the corresponding sample estimator

$$(37) \quad \widehat{\Omega}_\nu^*(\mathbf{y}^0) = \frac{1}{\widehat{N}_\nu} \sum_{\mathcal{S}_\nu} \frac{1}{\pi_i} \mathbf{I}(\|\mathbf{y}_i - \hat{\gamma}_\nu\|_{\widehat{\Sigma}_\nu} \leq \|\mathbf{y}^0 - \hat{\gamma}_\nu\|_{\widehat{\Sigma}_\nu}),$$

is now used, with

$$\|\mathbf{y}_i - \hat{\gamma}_\nu\|_{\widehat{\Sigma}_\nu} = \sqrt{(\mathbf{y}_i - \hat{\gamma}_\nu)' \widehat{\Sigma}_\nu^{-1} (\mathbf{y}_i - \hat{\gamma}_\nu)}, \quad \widehat{\Sigma}_\nu = \frac{1}{\widehat{N}_\nu - 1} \sum_{i \in \mathcal{S}_\nu} \frac{1}{\pi_i} (\mathbf{y}_i - \hat{\gamma}_\nu)(\mathbf{y}_i - \hat{\gamma}_\nu)'$$

The estimator (37) contains a sample-dependent norm, so that the previous theory does not apply directly in this case. To show the design consistency and asymptotic normality of this outlyingness measure with both center and shape estimated, we need to restate or strengthen the following assumptions,

1. The estimated shape measure  $\widehat{\Sigma}_\nu$  is  $\sqrt{n^*}$ -consistent for population quantity  $\Sigma_\nu$ .
2. In Assumption 3.5, we need the limiting function to be a continuous function of shape measure, with suitably defined derivatives.
3. A stronger uniform convergence assumption is needed in Assumption 3.6 (2).

The asymptotic properties of estimator (37) could then be derived following the same approach as in this article, but we will not do so explicitly here.

4.4. *Variance Estimation.* This section deals with estimating the variances of  $\widehat{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_\nu)$ ,  $\widehat{D}_{\nu,d}(\hat{\mathbf{q}}_\nu)$  and the outlyingness measures of Section 4.3. We will introduce several different estimators whose appropriateness will depend on the specific survey context. A first naive estimator ignores the error in estimating the population center. When that component of the error cannot be ignored, we propose to estimate it by kernel smoothing, after which it can be included in an analytic variance estimator. An alternative approach is to incorporate a kernel smoothing term in a jackknife variance estimator. These estimators are discussed below.

4.4.1. *Naive estimator.* In general, in order to estimate  $V(\widehat{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_\nu))$  and  $V(\widehat{D}_{\nu,d}(\hat{\mathbf{q}}_\nu))$  as defined in (19) and (32), we need to estimate the gradient vectors  $\frac{\partial \mathcal{D}_d(\boldsymbol{\mu}_\nu)}{\partial \boldsymbol{\mu}_\nu}$  and  $\frac{\partial \mathcal{D}_d(\mathbf{q}_\nu)}{\partial \mathbf{q}_\nu}$ . However, for certain population distributions and norms, these gradients vanish and hence a simple variance estimator is sufficient. We investigate this case here.

We will show in Lemma 4 and 5 that the gradient vectors  $\frac{\partial \mathcal{D}_d(\boldsymbol{\mu}_\nu)}{\partial \boldsymbol{\mu}_\nu}$  and  $\frac{\partial \mathcal{D}_d(\mathbf{q}_\nu)}{\partial \mathbf{q}_\nu}$  are negligible if the population acts like a sample from an elliptical distribution. We say a random variable  $\mathbf{Y}$  is distributed according to an *elliptically contoured distribution* with parameters  $\boldsymbol{\mu}$ ,  $\Lambda$  and  $\phi$ , denoted  $EC_p(\boldsymbol{\mu}, \Lambda, \phi)$ , if the characteristic function of  $\mathbf{Y}$  has the form  $\exp(it'\boldsymbol{\mu})\phi(t'\Lambda\mathbf{y})$ , where  $\boldsymbol{\mu}$  is a  $p \times 1$  vector and  $\Lambda$  is a non-negative definite matrix (Fang and Zhang 1980). The proofs of Lemmas 4 and 5 are given in Wang (2008).

LEMMA 4. *Assume random variable  $\mathbf{Y} \sim EC_p(\boldsymbol{\mu}, \Lambda, \phi)$  with mean vector  $\boldsymbol{\mu}$ , and  $\Lambda$  is a non-negative definite matrix. We define the norm as*

$$(38) \quad \|\mathbf{u}\| = \sqrt{\mathbf{u}'\mathbf{B}\mathbf{u}},$$

where  $\mathbf{B}$  is a non-negative definite matrix. Then

1. the partial derivative evaluated at superpopulation mean is  $\mathbf{0}$ ,  $\frac{\partial \mathcal{D}_d(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = \mathbf{0}$ .
2. mean  $\boldsymbol{\mu}$  coincide with median  $\mathbf{q}$ ,  $\boldsymbol{\mu} = \mathbf{q}$ .

LEMMA 5. *Assume 3.5,  $\frac{\partial \mathcal{D}_d(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = \mathbf{0}$  for some constant vector  $\boldsymbol{\gamma}$ , and the sequence of population centers  $\boldsymbol{\gamma}_\nu$  converges to  $\boldsymbol{\gamma}$ ,  $\lim_{\nu \rightarrow \infty} \boldsymbol{\gamma}_\nu = \boldsymbol{\gamma}$ . Then  $\frac{\partial \mathcal{D}_d(\boldsymbol{\gamma}_\nu)}{\partial \boldsymbol{\gamma}_\nu} = o(1)$ .*

Lemma 4 and 5 imply that the extra variability due to estimating the population center can be ignored in elliptical distributions with a norm specified by (38). This special case is similar to case A of Randles (1982), where we can “pretend” that we are using true population center  $\boldsymbol{\gamma}_\nu$  without affecting the leading variance term. So we propose the following naive plug-in variance estimator for the leading term

$$(39) \quad \widehat{V}_{NV} \left( \widehat{D}_{\nu,d}(\hat{\boldsymbol{\gamma}}_\nu) \right) = \left( 1, -\widehat{D}_{\nu,d}(\hat{\boldsymbol{\gamma}}_\nu) \right) \widehat{\Sigma}_{\boldsymbol{\gamma},d,NV} \left( 1, -\widehat{D}_{\nu,d}(\hat{\boldsymbol{\gamma}}_\nu) \right)',$$

with

$$(40) \quad \widehat{\Sigma}_{\boldsymbol{\gamma},d,NV} = \frac{n_\nu}{\widehat{N}_\nu^2} \sum_i \sum_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{\left[ \frac{\mathbf{I}(\|\mathbf{y}_i - \hat{\boldsymbol{\gamma}}_\nu\| \leq d)}{1} \right]}{\pi_i \pi_j} \left[ \frac{\mathbf{I}(\|\mathbf{y}_i - \hat{\boldsymbol{\gamma}}_\nu\| \leq d), 1}{1} \right],$$

and we use (39) as an estimator of asymptotic variances  $V \left( \widehat{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_\nu) \right)$  and  $V \left( \widehat{D}_{\nu,d}(\hat{\mathbf{q}}_\nu) \right)$  as defined in (19) and (32), where we plug in the estimated center  $\hat{\boldsymbol{\gamma}}_\nu$  in place of true center  $\boldsymbol{\gamma}_\nu$ .

4.4.2. *Estimating the effect of population center estimation error.* The naive estimator (39) ignores the variance due to estimating the population center and hence will tend to underestimate the true variance for a general population. So we need to estimate  $\frac{\partial \mathcal{D}_d(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}$  and incorporate the extra component of variance. Let  $\boldsymbol{\zeta}_d(\boldsymbol{\gamma}) = \frac{\partial \mathcal{D}_d(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}$ , and a population-level estimator of  $\boldsymbol{\zeta}_d(\boldsymbol{\gamma})$  using kernel smoothing is given by

$$(41) \quad \boldsymbol{\zeta}_{\nu,d}(\boldsymbol{\gamma}) = \frac{1}{N_\nu h} \sum_{U_\nu} K\left(\frac{d - \|\mathbf{y}_i - \boldsymbol{\gamma}\|}{h}\right) \boldsymbol{\psi}(\mathbf{y}_i - \boldsymbol{\gamma}).$$

We propose an sample-based estimator of  $\boldsymbol{\zeta}_d(\boldsymbol{\gamma})$  and thus  $\boldsymbol{\zeta}_{\nu g,d}(\boldsymbol{\gamma})$ , defined as

$$(42) \quad \hat{\boldsymbol{\zeta}}_{\nu,d}(\boldsymbol{\gamma}) = \frac{1}{\hat{N}_\nu h} \sum_{S_\nu} K\left(\frac{d - \|\mathbf{y}_i - \boldsymbol{\gamma}\|}{h}\right) \boldsymbol{\psi}(\mathbf{y}_i - \boldsymbol{\gamma}) \frac{1}{\pi_i}.$$

The idea of estimator  $\hat{\boldsymbol{\zeta}}_{\nu,d}(\boldsymbol{\gamma})$  is to estimate  $\mathcal{D}_d(\boldsymbol{\gamma})$  using a primitive function of kernel  $K(\cdot)$ , and then take derivatives with respect to the components of  $\boldsymbol{\gamma}$ .

LEMMA 6. *Under Assumptions 3.1, 3.2, 3.8(1,3) and 3.9(1,2), the estimator  $\hat{\boldsymbol{\zeta}}_{\nu,d}(\boldsymbol{\gamma})$  is design consistent for  $\boldsymbol{\zeta}_{\nu,d}(\boldsymbol{\gamma})$ .*

LEMMA 7. *Under Assumptions 3.1, 3.2, 3.8(1,2) and 3.9 (1-3), and assume the sequence of populations is such that*

$$(43) \quad \sup_{\boldsymbol{\gamma} \in \mathbb{R}^p} \left| \boldsymbol{\zeta}_{\nu,d}(\boldsymbol{\gamma}) - \boldsymbol{\zeta}_d(\boldsymbol{\gamma}) \right| \rightarrow \mathbf{0},$$

*the kernel estimator  $\hat{\boldsymbol{\zeta}}_{\nu,d}(\hat{\boldsymbol{\gamma}}_\nu)$  is design consistent for  $\boldsymbol{\zeta}_d(\boldsymbol{\gamma}_\nu)$  for every  $d$ .*

Similarly to what we did in Section 4.1, we have established uniform strong consistency of  $\boldsymbol{\zeta}_{\nu,d}(\boldsymbol{\gamma})$  for  $\boldsymbol{\zeta}_d(\boldsymbol{\gamma})$  in Wang (2008) under appropriate superpopulation assumptions. The assumption in (43) assumes that we are not working with the populations where  $\boldsymbol{\zeta}_{\nu,d}(\boldsymbol{\gamma})$  does not converge to  $\boldsymbol{\zeta}_d(\boldsymbol{\gamma})$ , which is a zero-probability event under the superpopulation assumptions.

We can directly plug the estimator (42) and sample estimates of  $D_{\nu,d}(\boldsymbol{\gamma}_\nu)$  and  $\Sigma_{\boldsymbol{\gamma},d}$  into variance expressions (19) and (32) to get design-consistent analytic variance estimators for the mean-based and median-based cases, respectively. But unlike naive estimator (39), this variance estimator incorporates the extra variability due to estimating the center. In the next section, we use  $\hat{\boldsymbol{\zeta}}_{\nu,d}(\boldsymbol{\gamma})$  in the construction of a replication-based variance estimator.

4.4.3. *Jackknife variance estimator.* This section only applies formally to mean-based estimator, but can be modified to include median-based estimator by using delete- $d$  jackknife or other replication methods such as Balanced Repeated Replication (BRR). To introduce the jackknife variance estimator for our application, we start by assuming there already exists a design consistent jackknife variance estimator for simple linear estimators. Then we define jackknife replicates in our case and show the design consistency of the resulting variance estimator. This approach is also used by Fuller and Kim (2005) and Da Silva and Opsomer (2006). We will use a number of assumptions from the latter article, and not state them here fully for the sake of brevity. The proof of the following theorem can be found in Wang (2008).

THEOREM 8. *Let  $\hat{\theta}$  be a linear estimator with*

$$\hat{\theta} = \sum_{S_\nu} w_i z_i,$$

where  $w_i$  is the survey weight and  $z_i$  has bounded  $4 + \delta$  moments. Assume there is a jackknife replication procedure that generates  $L$  replicated estimates

$$\hat{\theta}^{(l)} = \sum_{S_\nu} w_i^{(l)} z_i,$$

with  $l = 1, 2, \dots, L$  and  $w_i^{(l)}$  is replication weight for unite  $i$  in the  $l$ -th replicate. The replication variance estimator is defined as

$$(44) \quad \hat{V}_{JK}(\hat{\theta}) = \sum_{l=1}^L c_l \left( \hat{\theta}^{(l)} - \hat{\theta} \right)^2,$$

where  $c_l$  is a set of constants for the  $l$ -th replicate. Assumptions similar to (D1)-(D4) and (D6) in Da Silva and Opsomer (2006) are assumed.

We define the  $l$ -th jackknife replicate as

$$(45) \quad \hat{D}^{(l)}(\hat{\boldsymbol{\mu}}_\nu) = \hat{D}_{\nu,d}^{(l)}(\hat{\boldsymbol{\mu}}_\nu) + \hat{\boldsymbol{\zeta}}_{\nu,d}(\hat{\boldsymbol{\mu}}_\nu)(\hat{\boldsymbol{\mu}}_\nu^{(l)} - \hat{\boldsymbol{\mu}}_\nu),$$

where  $\hat{D}_{\nu,d}^{(l)}(\hat{\boldsymbol{\mu}}_\nu) = \frac{1}{\hat{N}_\nu^{(l)}} \sum_{i \in S_\nu} w_i^{(l)} \mathbf{I}_{(\|\mathbf{y}_i - \hat{\boldsymbol{\mu}}_\nu\| \leq d)}$ ,  $\hat{N}_\nu^{(l)} = \sum_{i \in S_\nu} w_i^{(l)}$  and  $\hat{\boldsymbol{\mu}}_\nu^{(l)} = \frac{1}{\hat{N}_\nu^{(l)}} \sum_{i \in S_\nu} w_i^{(l)} \mathbf{y}_i$ , and  $\hat{\boldsymbol{\zeta}}_{\nu,d}(\hat{\boldsymbol{\mu}}_\nu)$  as defined in (42). Then the jackknife variance estimator

$$(46) \quad \hat{V}_{JK} \left( \hat{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_\nu) \right) = \sum_{l=1}^L c_l \left( \hat{D}^{(l)}(\hat{\boldsymbol{\mu}}_\nu) - \hat{D}(\hat{\boldsymbol{\mu}}_\nu) \right)^2$$

is design consistent for  $V \left( \hat{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_\nu) \right)$  in (19).

This replication variance estimator is readily interpreted by considering the composition of the replicate in (45). We first ignore the second term in (45) and compare the resulting jackknife estimator with the naive estimator (39). Both estimators consistently estimate the asymptotic variance of  $\widehat{D}_\nu(\boldsymbol{\mu}_\nu)$ , corresponding to setting  $\frac{\partial \mathcal{D}_d(\boldsymbol{\mu}_\nu)}{\partial \boldsymbol{\mu}_\nu} = 0$  in (20). The second term in (45) uses the combination of the kernel estimator and the replication method to estimate the effect of estimating the subpopulation center. It should be noted that the naive “full” jackknife variance estimator which recalculates both the mean and the fraction of distances in each replicate,  $\widehat{D}^{(l)}(\hat{\boldsymbol{\mu}}_\nu^{(l)})$ , results in an inconsistent variance estimator, as  $\widehat{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_\nu)$  is a nonsmooth function of  $\hat{\boldsymbol{\mu}}_\nu$ .

In many practical situations, a significant advantage of the replication-based approach is that we do not need to estimate the covariance matrix (18), which can be complicated in a large-scale complex survey. In jackknife variance estimation, we start from an existing jackknife procedure for simple linear estimators, as could for instance be provided by a statistical agency as part of the survey dataset. We then estimate the gradient vector using kernel regression based on the whole sample only once, while  $\widehat{D}_{\nu,d}^{(l)}(\hat{\boldsymbol{\mu}}_\nu)$  and  $\hat{\boldsymbol{\mu}}_\nu^{(l)}$  can be computed based on the replicate weights  $w_i^{(l)}$ . Then, the replicates  $\widehat{D}^{(l)}(\hat{\boldsymbol{\mu}}_\nu)$  are computed as in (45).

Delete- $d$  jackknife can be used in complex surveys or in case of using median as measures of center. As has been pointed out by Shao and Wu (1989), the number of deleted points  $d$  has to go to infinity at some rate depending on the nonsmoothness of the estimator. By choosing an appropriate  $d$ , we can account for the variation in  $\hat{\boldsymbol{q}}_\nu$  as well as the nonsmoothness of  $\widehat{D}_{\nu,d}(\cdot)$ . We do not explore this issue further here.

## 5. Simulation study.

5.1. *Distribution Function Estimation.* When deriving theoretical results in Section 4, we ignored the subpopulation structure in our finite population. But the knowledge of subpopulation structures is very helpful in detecting suspicious points, and it is usually a practical concern. So in our simulation study, we will re-introduce the subpopulations and examine the estimator and variance estimators for subpopulations with different distribution structures.

We generate a finite population of bivariate variables with 2 subpopulations of equal size  $N_{\nu g} = 1000$ ,  $g = 1, 2$  defined as follows:

- Subpopulation 1 is simulated from a truncated Pearson type VII dis-

tribution with,

$$\mathbf{Y}_1 \sim \begin{bmatrix} -5 \\ -6 \end{bmatrix} + R_1 \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} \mathbf{U}^{(2)},$$

where  $R_1 = \frac{Z}{1-Z} \mathbf{I}_{(\frac{Z}{1-Z} < 5)}$  with  $Z \sim \text{Beta}(1, 1)$  independent of  $\mathbf{U}^{(2)}$ , which is uniformly distributed on unit circle in  $\mathbb{R}^2$ .

- Subpopulation 2 is generated from

$$\mathbf{Y}_2 \sim \begin{bmatrix} 5 \\ -6 \end{bmatrix} + \begin{bmatrix} R(\theta) \cos \theta \\ R(\theta) \sin \theta \end{bmatrix},$$

with  $\theta \sim \text{Unif}[0, 2\pi]$  and  $R(\theta) = |\theta - \pi| \times \chi_2^2$ .

A more complete simulation experiment with five subpopulations is reported in Wang (2008).

The distribution of  $\mathbf{Y}_1$  is elliptically contoured, and subpopulation 2 is a skewed population with an irregular shape. Figure 1 shows the scatterplot of two subpopulations. Figure 2 shows the empirical distribution function of subpopulation distances using both mean and median as measures of center. For elliptical subpopulation 1, the empirical distribution of distances using mean or median do not differ much but the two distributions differ noticeably for a skewed cluster like subpopulation 2.

We conducted simulation studies to assess the consistency and asymptotic normality of our estimators, using sampling schemes including simple random sampling, stratified simple random sampling with differing sampling fractions, and Poisson sampling. We only show the result from stratified sampling. For complete simulation results, see Wang (2008). We partition the overall population into two strata, based on if the second coordinate is smaller than or greater than  $-5$ . We draw a simple random samples from each stratum with sample sizes  $0.4n_\nu$  and  $0.6n_\nu$ , respectively, and overall sample size  $n_\nu = 200, 400$  or  $1000$ . The stratification cuts across both subpopulations, with 84% of subpopulation 1 and 80% of subpopulation 2 falling into stratum 1, and the remainders in stratum 2.

To evaluate the performance of estimator  $\hat{D}_{\nu,d}(\hat{\gamma}_\nu)$ , we will compare it with  $\hat{D}_{\nu,d}(\gamma_\nu)$ , for which we use the true population center. The comparison is on the basis of bias and standard deviation under three different sample sizes at distances  $d = 0.707, 1.0, 1.414, 2.45$  and  $3.873$ . The simulation results are shown in Tables 1 and 2 with mean and median as measures of center, respectively.

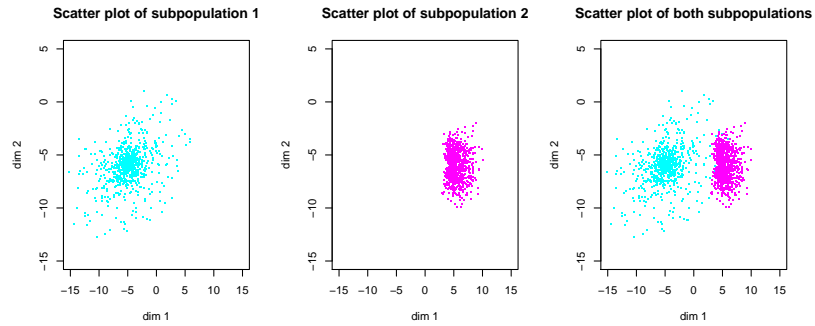


FIG 1. Scatterplot of two subpopulations; the first 2 plots show each individual subpopulation and the last plot shows them both in different colors.

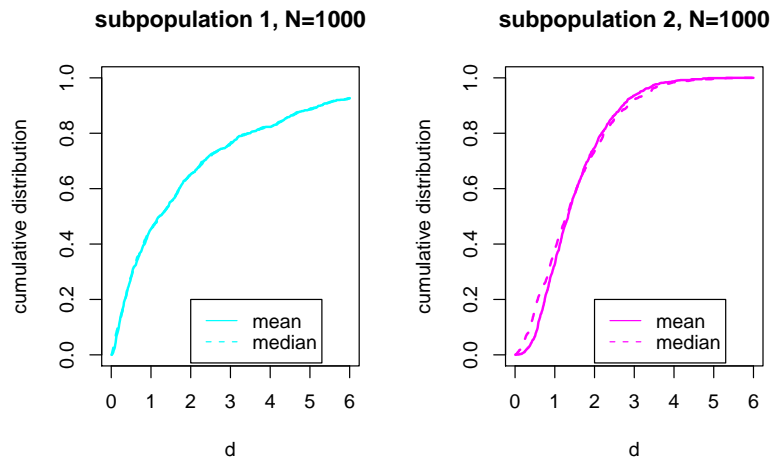


FIG 2. Distribution of subpopulation distances, with solid line indicating distances to mean vector and dash line distances to median.

The bias of  $\widehat{D}_{\nu,d}(\boldsymbol{\mu}_\nu)$  is negligible compared to its standard deviation in all three sample sizes. The bias of  $\widehat{D}_{\nu,d}(\widehat{\boldsymbol{\mu}}_\nu)$  has same magnitude as its standard deviation for some  $d$ 's under sample sizes  $n_\nu = 400$  and  $200$ . The bias diminishes as  $d$  increases. Due to estimating subpopulation center, we have introduced extra bias and variance to the estimator, in general. The increase in variance can be seen by comparing  $sd(\widehat{D}_\nu(\widehat{\boldsymbol{\mu}}_\nu))$  with  $sd(\widehat{D}_\nu(\boldsymbol{\mu}_\nu))$ . The difference between  $sd(\widehat{D}_\nu(\widehat{\boldsymbol{\mu}}_\nu))$  and  $sd(\widehat{D}_\nu(\boldsymbol{\mu}_\nu))$  is nearly negligible for subpopulation 1 under large sample sizes, confirming the use of naive variance estimator in elliptical distributions. But this discrepancy remains obvious in a skewed subpopulation even for the largest sample size.

Table 2 shows the results of using the median as a measure of center. For subpopulation 1, the bias and standard deviation of  $\widehat{D}_\nu(\mathbf{q}_\nu)$  are very close to that of  $\widehat{D}_\nu(\boldsymbol{\mu}_\nu)$  as the two centers are very close. The bias of  $\widehat{D}_\nu(\widehat{\mathbf{q}}_\nu)$  is generally smaller than  $\widehat{D}_\nu(\widehat{\boldsymbol{\mu}}_\nu)$ , everything else being the same, as a result of the median being more robust against “bad” samples. And  $sd(\widehat{D}_\nu(\widehat{\mathbf{q}}_\nu))$  is very close to  $sd(\widehat{D}_\nu(\mathbf{q}_\nu))$  under all three sample sizes for elliptical subpopulations, but the deviance remains obvious for skewed subpopulations.

Another interesting finding is that as  $d$  increases, the extra variance due to estimating the center will diminish, consistent with our intuition that the effect of estimating the center will decrease as we examine a larger region around the center. The simulation result confirms that the contribution to variance of estimating the center is much less in the tail of the distribution.

Next, we report on a comparison of the performance of the naive variance estimator defined by (39) and the kernel and jackknife combined variance estimator for the same set of distances  $d$ . We only report on the results for the mean-based case, as this is the case for which we proved the consistency of the jackknife estimator.

Table 3 shows that the Monte Carlo variance increases and then decreases as  $D_{\nu,d}(\boldsymbol{\mu}_\nu)$  goes from 0 to 1. The naive variance estimator (39) performs satisfactorily if the underlying distribution is elliptically contoured (subpopulation 1), but severely underestimates the variance in skewed distributions as in subpopulation 2. Incorporating the kernel smoothing term enables us to include the additional variance component, but we need to be careful in choosing the bandwidth. The variance estimate generally decreases as bandwidth  $h$  increases, and the “optimal” bandwidth depends not only on the features of the cluster but also on how far away from center we are estimating  $D_{\nu,d}(\boldsymbol{\mu}_\nu)$ . In additional simulation experiments reported in Wang (2008), we also confirmed that the jackknife estimator that re-calculates  $\widehat{D}_{\nu,d}(\widehat{\boldsymbol{\mu}}_\nu)$  for each replicate tends to severely overestimate the variance.

TABLE 1  
 Comparison of estimator  $\widehat{D}_{\nu,d}(\widehat{\mu}_{\nu})$  and  $\widehat{D}_{\nu,d}(\mu_{\nu})$  in terms of bias and standard deviation under three different simple random samples;  $G = 2$  is the number of subpopulations;  $bias(\widehat{D}(\widehat{\mu}))$ ,  $bias(\widehat{D}(\mu))$  represent the bias of our estimator when the true population mean is estimated or known;  $sd(\widehat{D}(\widehat{\mu}))$ ,  $sd(\widehat{D}(\mu))$  represent the standard deviation of our estimator when the true population mean is estimated or known.

$N_{\nu} = 2000$										
	Subpopulation 1					Subpopulation 2				
$d$	0.707	1.0	1.414	2.45	3.873	0.707	1.0	1.414	2.45	3.873
$D_{\nu g,d}(\mu_{\nu})$	.322	.422	.526	.708	.82	.13	.28	.508	.852	.99
$n_{\nu} = 400$										
$bias(\widehat{D}(\mu))$	.0004	.0002	.0001	.0004	0	.0003	.0006	.001	.0003	.0002
$bias(\widehat{D}(\widehat{\mu}))$	-.0005	-.0025	.0028	-.001	-.0005	.0059	.0042	-.0027	.0006	0
$sd(\widehat{D}(\mu))$	.017	.02	.021	.022	.02	.027	.035	.038	.028	.008
$sd(\widehat{D}(\widehat{\mu}))$	.018	.02	.022	.022	.021	.038	.043	.044	.026	.008
$n_{\nu} = 160$										
$bias(\widehat{D}(\mu))$	.0002	.0007	.0005	.0009	.0011	-.0011	-.0012	-.0012	-.0001	-.0002
$bias(\widehat{D}(\widehat{\mu}))$	-.0105	-.0072	-.0008	-.0009	.0029	.0071	.0047	-.0034	.0028	-.0002
$sd(\widehat{D}(\mu))$	.038	.038	.04	.04	.036	.06	.075	.074	.043	.013
$sd(\widehat{D}(\widehat{\mu}))$	.033	0.037	.04	.039	.035	.043	.058	.065	.046	.013
$n_{\nu} = 80$										
$bias(\widehat{D}(\mu))$	.0024	.003	.0029	.0019	.0018	-.0017	-.0013	-.0034	.0006	0
$bias(\widehat{D}(\widehat{\mu}))$	-.0253	-.0145	-.0033	-.0004	.0065	.0121	.0091	0	.0079	.0007
$sd(\widehat{D}(\mu))$	.05	.054	.057	.056	.051	.061	.084	.093	.065	.019
$sd(\widehat{D}(\widehat{\mu}))$	.063	.059	.058	.058	.051	.084	.107	.106	.062	.018

TABLE 2

Comparison of estimator  $\widehat{D}_{\nu,d}(\widehat{\mathbf{q}}_{\nu})$  and  $\widehat{D}_{\nu,d}(\mathbf{q}_{\nu})$  in terms of bias and standard deviation under three sample sizes;  $G = 2$  is the number of subpopulations;  $bias(\widehat{D}(\widehat{\mathbf{q}})), bias(\widehat{D}(\mathbf{q}))$  represent the bias with true population median estimated or known;  $sd(\widehat{D}(\widehat{\mathbf{q}})), sd(\widehat{D}(\mathbf{q}))$  represent the standard deviation with true population median estimated or known.

$N_{\nu} = 2000$										
	Subpopulation 1					Subpopulation 2				
$d$	0.707	1.0	1.414	2.45	3.873	0.707	1.0	1.414	2.45	3.873
$D_{\nu,d}(\mathbf{q}_{\nu})$	.322	.421	.53	.708	.817	.217	.347	.526	.825	.981
$n_{\nu} = 400$										
$bias(\widehat{D}(\mathbf{q}))$	.0004	.0002	.0002	.0004	0	.0002	.0002	.0012	.0004	.0003
$bias(\widehat{D}(\widehat{\mathbf{q}}))$	.0012	.0025	.0006	-.0005	.0002	-.0016	.0028	.0014	.0037	.0008
$sd(\widehat{D}(\mathbf{q}))$	.017	0.02	.022	.022	.02	.032	.036	.039	.029	.011
$sd(\widehat{D}(\widehat{\mathbf{q}}))$	.017	.02	.022	0.022	.02	.04	.045	.038	.027	.011
$n_{\nu} = 160$										
$bias(\widehat{D}(\mathbf{q}))$	.0002	.0006	.0005	.001	.0011	-.0005	-.0019	-.0008	-.0001	-.0004
$bias(\widehat{D}(\widehat{\mathbf{q}}))$	.0023	.003	.0009	.0002	.0018	-.0027	.0032	.0013	.0071	.0006
$sd(\widehat{D}(\mathbf{q}))$	.033	.037	.04	.039	.035	.053	.061	.064	.049	.018
$sd(\widehat{D}(\widehat{\mathbf{q}}))$	.033	.037	.04	.039	.035	.067	.073	.065	.045	.018
$n_{\nu} = 80$										
$bias(\widehat{D}(\mathbf{q}))$	.0024	.0031	.0029	.0021	.0019	-.0009	-0.002	-.0015	.0007	-.0002
$bias(\widehat{D}(\widehat{\mathbf{q}}))$	.0059	.0062	.0043	.0016	.0032	.0011	.009	.0046	.0122	.0014
$sd(\widehat{D}(\mathbf{q}))$	.05	.054	.057	.056	.051	.077	.09	.093	.07	.025
$sd(\widehat{D}(\widehat{\mathbf{q}}))$	.05	.055	.058	.056	.051	.097	.105	.095	.064	.024

TABLE 3

Comparison of variance estimators:  $V_{MC}$  denotes the Monte Carlo estimate of variance from 2000 simulations,  $\widehat{V}_{NV}(\widehat{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_{\nu}))$  denotes the naive estimator,  $\widehat{V}_{SM}(\widehat{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_{\nu}))$  denotes the analytic variance estimator using the kernel estimator, and  $\widehat{V}_{JK}(\widehat{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_{\nu}))$  denotes the jackknife variance estimator using the kernel estimator.

$N_{\nu} = 2000, n_{\nu} = 400, G = 2$										
	Subpopulation 1					Subpopulation 2				
$d$	0.707	1.0	1.414	2.45	3.873	0.707	1.0	1.414	2.45	3.873
$D_{\nu,d}(\boldsymbol{\mu}_{\nu})$	0.331	0.425	0.517	0.674	0.813	0.183	0.324	0.544	0.868	0.987
$V_{MC} \times 1000$	0.741	0.79	0.8	0.786	0.555	0.869	1.013	0.925	0.423	0.043
$\frac{\widehat{V}_{NV}(\widehat{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_{\nu}))}{V_{MC}}$	0.931	0.981	1.021	0.962	0.969	0.523	0.68	0.891	0.961	1.105
$\frac{\widehat{V}_{SM}(\widehat{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_{\nu}))}{V_{MC}}$										
$h = 0.05$	1.432	1.403	1.275	1.109	1.129	1.113	1.137	1.135	1.057	1.133
$h = 0.2$	1.132	1.131	1.099	0.998	1.007	0.942	1.002	1.055	0.961	1.028
$h = 0.4$	1.155	1.086	1.066	0.981	0.987	0.873	0.998	1.07	0.911	1.011
$\frac{\widehat{V}_{JK}(\widehat{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_{\nu}))}{V_{MC}}$										
$h = 0.05$	1.186	1.067	1.058	0.966	1.057	1.034	0.98	0.946	0.94	1.108
$h = 0.2$	0.944	0.87	0.903	0.868	0.954	0.885	0.869	0.881	0.881	1.028
$h = 0.4$	0.952	0.844	0.863	0.841	0.938	0.818	0.871	0.883	0.857	1.015

5.2. *Outlier Detection.* In this simulation, we evaluate the performance of our proposed approach in detecting potential outliers in a survey. We will apply the outlyingness measure  $\widehat{\Omega}_{g,\nu}^*(\mathbf{y}^0)$  which is now defined at the subpopulation level and takes the shape of the subpopulation into account, as discussed in Section 4.3. In setting up this experiment, we first need to determine which population observations to label as “potential outliers.” Instead of creating a population and adding artificial outliers, we decided to generate a population with five subpopulations (see Figure 3) and label all points that are “far” from all of the subpopulation centers as outliers (see below). Hence, they are not outliers in the sense of not belonging to the population, but they are unusual with respect to the population distribution and therefore good candidates for closer scrutiny if they happen to be included in the sample.

The first four subpopulations are simulated from elliptical symmetric distributions and subpopulation 5 possesses some nonsymmetry. Subpopulations 4 and 5 are subpopulations 1 and 2 in Section 5.1, respectively. We used subpopulation means as centers in this experiment. In defining suspicious points according to the reasoning above, we prespecify a threshold  $\alpha$  equal to 0.02 in this case, and define the population points with  $\Omega_{g,\nu}^*(\mathbf{y}^0) \geq 1 - \alpha$

for  $g = 1, \dots, 5$  to be the set of outliers. Figure 3 shows the points identified as outliers. Given this set of suspicious points, the goal of the simulation experiment is to assess how well a sample-based estimator is able to identify the same points as being potential outliers.

The subpopulation structure is not used in the design but will be assumed known for the purpose of constructing the outlyingness measure. Probability samples are drawn under a complex design with 3 strata, and different designs for each stratum. We created a stratification variable  $z_i = y_{1i} + y_{2i} + \epsilon_i$  with  $\epsilon_i \sim N(0, 1)$ , and use -5 and 5 as cutoff points on  $z_i$  for determining stratum membership for each element  $i \in U_\nu$ . Stratum 1 contains the units where  $z_i \leq -5$ , and we draw a Poisson sample with inclusion probability proportional to  $|z_i|$  and anticipated sample size  $n/4$ . Stratum 2 has the units where  $-5 < z_i < 5$ , and we equally partition the range of  $z_i$  into 500 intervals, and select clusters using simple random sampling to obtain an anticipated stratum sample size of  $n/2$ . Finally, we draw a simple random sample of size  $n/4$  from stratum 3.

The scenario we are interested in investigating is the classification of a new point relative to the sample-fitted subpopulation distributions. Hence, we calculate  $\hat{\Omega}_\nu^*(\mathbf{y}^0)$  for each point in a sample by leaving it out in constructing the distance distributions. If the outlyingness measures of that point with respect to all five subpopulations are greater than  $1 - \alpha$ , then this point is labeled as an outlier in the sample.

We consider four settings: using mean or median as the center of subpopulation, and large or small sample sizes ( $n = 1000$  or  $n = 400$ ). Then we calculate the average number of true outliers in the sample, average number of outliers identified by the outlyingness measure, and the fraction of true outliers caught by the sample-based rule. Table 4 shows that we can correctly flag at least 74% of the true outliers using sample-based rule in the four scenarios, and the results improve with the sample size. The mean and median based inferences similar in this case, but can be expected to differ more if the subpopulations are more skewed. Additional simulations results are in Wang (2008).

### References.

- Barnett, V. and T. Lewis (1994). *Outliers in Statistical Data*. John Wiley & Sons.
- Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley & Sons.
- Breidt, F. and J. Opsomer (2008). Endogenous post-stratification in surveys: classifying with a sample-fitted model. *Annals of Statistics* 36,

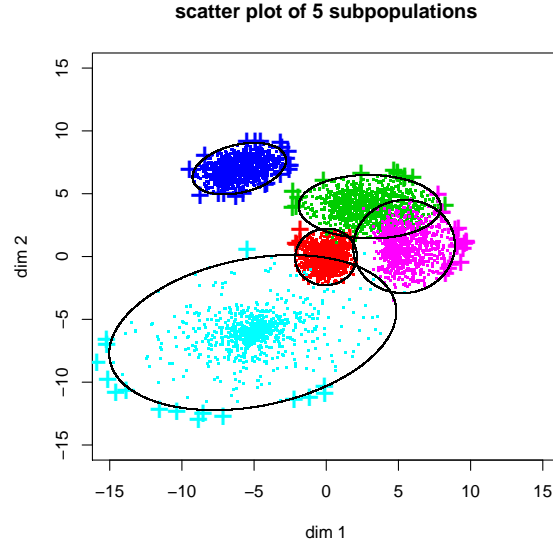


FIG 3. Bivariate population with 5 subpopulations; points shown as “plus” signs are population outliers, defined by  $\Omega_{g,\nu}^*(\mathbf{y}^0) \geq 1 - \alpha$  for  $g = 1, \dots, 5$ . We use mean vectors as measures of center and the threshold  $\alpha$  is chosen as 0.02.

TABLE 4

Summary table of simulation results to examine the performance of detecting population outliers. We use either mean or median as measure of population center and population outliers are defined by  $\Omega_{g,\nu}^*(\mathbf{y}^0) \geq 1 - \alpha$  for  $g = 1, \dots, 5$ . Sample outliers are the points  $\mathbf{y}^0$  such that  $\hat{\Omega}_{g,\nu}^*(\mathbf{y}^0) \geq 1 - \alpha$  for  $g = 1, \dots, 5$ .

	mean number of true outliers in sample	mean number of outliers identified	fraction of true outliers correctly identified
mean, n=1000	17.7	19.8	0.81
mean, n=400	7.1	9.3	0.74
median, n=1000	17.6	19.6	0.82
median, n=400	7.0	9.2	0.75

403–427.

- Brown, B. M. (1983). Statistical uses of the spatial median. *Journal of the Royal Statistical Society, Series B: Methodological* 45, 25–30.
- Chambers, R. L., A. H. Dorfman, and P. Hall (1992). Properties of estimators of the finite population distribution function. *Biometrika* 79, 577–582.
- Da Silva, D. and J. Opsomer (2006). A kernel smoothing method to adjust for unit nonresponse in sample surveys. *the Canadian Journal of Statistics* 34, 563–579.
- Dorfman, A. H. (2007). Inference on distributions and quantiles. In C. Rao and D. P. (editors) (Eds.), *Handbook of Statistics Volume 29: Sample Surveys: Theory, Methods and Inference*, Elsevier/North-Holland [Elsevier Science Publishing Co., New York; North-Holland Publishing Co., Amsterdam].
- Dunstan, R. and R. L. Chambers (1986). Model-based confidence intervals in multipurpose surveys. *Applied Statistics* 35, 276–280.
- Fang, K.-T. and Y.-T. Zhang (1980). *Generalized Multivariate Analysis*. Springer-Verlag.
- Francisco, C. A. and W. A. Fuller (1991). Quantile estimation with a complex survey design. *The Annals of Statistics* 19, 454–469.
- Fuller, W. (2007). *Sampling Statistics*.
- Fuller, W. A. and J. K. Kim (2005). Hot deck imputation for the response model. *Survey Methodology* 31(2), 139–149.
- Hawkins, D. M. (1980). *Identification of Outliers*. Chapman & Hall Ltd.
- Isaki, C. and W. Fuller (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association* 77, 89–96.
- Lyberg, L., P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (1997). *Survey Measurement and Process Quality*. John Wiley & Sons.
- Nusser, S. M. and J. J. Goebel (1997). The National Resources Inventory: A long-term multi-resource monitoring programme. *Environmental and Ecological Statistics* 4, 181–204.
- Opsomer, J. D. and D. Ruppert (1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics* 25(1), 186–211.
- Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag Inc.
- Randles, R. H. (1982). On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics* 10, 462–474.

- Rao, J., J. G. Kovar, and H. J. Mantel (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika* 77, 365–375.
- Shao, J. and C. F. J. Wu (1989). A general theory for jackknife variance estimation. *The Annals of Statistics* 17, 1176–1197.
- Small, C. G. (1990). A survey of multidimensional medians. *International Statistical Review* 58, 263–277.
- Wang, J. (2008). *Estimating the distance distribution of subpopulations for a large-scale complex survey*. Ph. D. thesis, Iowa State University.

## APPENDIX A: MODEL-BASED RESULTS

In Section 3, we have assumed several regularity conditions on the sequence of finite populations to derive the design-based results. In this section, we provide sufficient conditions under a superpopulation model to make it possible to evaluate the “reasonableness” of these population-level regularity conditions. We will show that the assumptions made in Section 3 hold with probability one under the superpopulation model.

In the model-based context, we assume that each subpopulation is an independent and identically distributed sample from a superpopulation model with cumulative distribution function  $F_g(\mathbf{y})$ . Let  $\mathbf{Y}_g$  denote the random variable from model component  $g$ . We state a model version of Assumption 3.5 below, and show that the statements in Assumption 3.6 hold with probability one under that assumption. Proofs are given in Appendix B.

ASSUMPTION A.1. *The distribution of  $\|\mathbf{Y}_g - \boldsymbol{\gamma}\|$  can be written as,*

$$\text{EI}_{(\|\mathbf{Y}_g - \boldsymbol{\gamma}\| \leq d)} = \text{P}(\|\mathbf{Y}_g - \boldsymbol{\gamma}\| \leq d) = \mathcal{D}_d(\boldsymbol{\gamma}),$$

which is continuous in  $d \in [0, \infty)$  and  $\boldsymbol{\gamma} \in \mathbb{R}^p$ . Additionally, the derivatives  $\frac{\partial \mathcal{D}_d(\boldsymbol{\gamma})}{\partial d}$ ,  $\frac{\partial \mathcal{D}_d(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}$  and  $\frac{\partial^2 \mathcal{D}_d(\boldsymbol{\gamma})}{\partial d^2}$  all exist in  $(d, \boldsymbol{\gamma}) \in [0, +\infty) \times \mathbb{R}^p$ .

LEMMA A.1. *Under Assumption A.1,*

$$\sqrt{N_\nu} \left\{ \frac{1}{N_\nu} \sum_{U_\nu} \mathbf{I}_{(d < \|\mathbf{y}_i - \boldsymbol{\gamma}\| \leq d + h_{N_\nu})} - \frac{\partial \mathcal{D}_d(\boldsymbol{\gamma})}{\partial d} h_{N_\nu} \right\} \xrightarrow{a.s.} 0$$

where  $h_{N_\nu} = O(N_\nu^{-\alpha})$  and  $\alpha \in (\frac{1}{4}, 1)$ .

LEMMA A.2. *Under Assumption A.1, and assuming that  $N_\nu h_{N_\nu} (\log N_\nu)^{-1} \rightarrow \infty$ , as  $N_\nu \rightarrow \infty$ , let*

$$(47) \quad R_{2i}(\boldsymbol{\gamma}) = \mathbf{I}_{(\|\mathbf{y}_i - \boldsymbol{\gamma} - h_{N_\nu} \mathbf{s}\| \leq d)} - \mathbf{I}_{(\|\mathbf{y}_i - \boldsymbol{\gamma}\| \leq d)} - \mathcal{D}_d(\boldsymbol{\gamma} + h_{N_\nu} \mathbf{s}) + \mathcal{D}_d(\boldsymbol{\gamma})$$

then

$$\sup_{(\gamma, \mathbf{s}) \in \mathbb{R}^p \times C_{\mathbf{s}}} \left| \frac{h_{N_\nu}}{N_\nu} \sum_{U_\nu} R_{2i}(\gamma) \right| \xrightarrow{a.s.} 0$$

for any compact region  $C_{\mathbf{s}}$  in  $\mathbb{R}^p$ , and  $h_{N_\nu} \rightarrow 0$  as  $N_\nu \rightarrow \infty$ .

**COROLLARY A.1.** *In Lemma A.2, we can replace  $\gamma$  with a stochastically bounded sequence  $\gamma_\nu$  and obtain,*

$$\frac{h_{N_\nu}}{N_\nu} \sum_{U_\nu} R_{2i}(\gamma_\nu) \xrightarrow{a.s.} 0$$

uniformly for  $\mathbf{s} \in C_{\mathbf{s}}$ , where  $R_{2i}(\cdot)$  is defined in (47).

## APPENDIX B: PROOFS

### Proof of Lemma 1

**PROOF.** Let us define  $\tilde{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_\nu)$  similar to (4). It suffices to show that,

$$n_\nu^{*1/2} \left( \tilde{D}_{\nu,d}(\hat{\boldsymbol{\mu}}_\nu) - \tilde{D}_{\nu,d}(\boldsymbol{\mu}_\nu) - \mathcal{D}_{g,d}(\hat{\boldsymbol{\mu}}_\nu) + \mathcal{D}_{g,d}(\boldsymbol{\mu}_\nu) \right) \xrightarrow{P} 0.$$

Define  $Q_n(\mathbf{s}) =$

$$n_\nu^{*1/2} \left( \tilde{D}_{\nu,d}(\boldsymbol{\mu}_\nu + n_\nu^{*-1/2} \mathbf{s}) - \tilde{D}_{\nu,d}(\boldsymbol{\mu}_\nu) - \mathcal{D}_{g,d}(\boldsymbol{\mu}_\nu + n_\nu^{*-1/2} \mathbf{s}) + \mathcal{D}_{g,d}(\boldsymbol{\mu}_\nu) \right),$$

As  $\hat{\boldsymbol{\mu}}_\nu = \boldsymbol{\mu}_\nu + O_p\left(\frac{1}{\sqrt{n_\nu^*}}\right)$  following Assumption 3.2, it suffices to prove that

$$\sup_{\mathbf{s} \in C} |Q_n(\mathbf{s})| \xrightarrow{P} 0,$$

for some compact set  $C$ . As

$$\begin{aligned} & |Q_n(\mathbf{s})| \\ & \leq \left| n_\nu^{*1/2} \left( \tilde{D}_{\nu,d}(\boldsymbol{\mu}_\nu + n_\nu^{*-1/2} \mathbf{s}) - \tilde{D}_{\nu,d}(\boldsymbol{\mu}_\nu) - D_{\nu,d}(\boldsymbol{\mu}_\nu + n_\nu^{*-1/2} \mathbf{s}) + D_{\nu,d}(\boldsymbol{\mu}_\nu) \right) \right| \\ & \quad + \left| n_\nu^{*1/2} \left( D_{\nu,d}(\boldsymbol{\mu}_\nu + n_\nu^{*-1/2} \mathbf{s}) + D_{\nu,d}(\boldsymbol{\mu}_\nu) - \mathcal{D}_{g,d}(\boldsymbol{\mu}_\nu + n_\nu^{*-1/2} \mathbf{s}) + \mathcal{D}_{g,d}(\boldsymbol{\mu}_\nu) \right) \right|. \end{aligned}$$

The second term converges to zero uniformly for  $\mathbf{s} \in C_{\mathbf{s}}$  by Assumption 3.6(2). It remains to show that

$$(48) \quad \sup_{\mathbf{s} \in C} \left| \frac{1}{N_\nu} \sum_{U_\nu} \left( \frac{I_{(i \in S_\nu)}}{\pi_i} - 1 \right) a_i(\mathbf{s}) \right| \xrightarrow{P} 0,$$

where  $a_i(\mathbf{s}) \hat{=} n_\nu^{*1/2} \left[ \mathbf{I}_{(\|\mathbf{y}_i - \boldsymbol{\mu}_\nu - n_\nu^{*-1/2} \mathbf{s}\| \leq d)} - \mathbf{I}_{(\|\mathbf{y}_i - \boldsymbol{\mu}_\nu\| \leq d)} \right]$ .

For any  $1 - \beta < \xi < \frac{\beta}{2p}$  where  $\beta$  is defined in Assumption 3.2(1), use the following partition of  $C$ ,

$$C = C_1 \cup C_2 \cup \dots \cup C_{N_\nu^{\xi p}}, C_j \cap C_{j'} = \emptyset, \forall j \neq j',$$

where  $\text{Diam}(C_j) = O(N_\nu^{-\xi}), \forall j = 1, 2, \dots, N_\nu^{\xi p}$ .

Select a set of  $\mathbf{s}_j \in C_j, j = 1, 2, \dots, N_\nu^{\xi p}$ ,

$$\begin{aligned} & \mathbb{P} \left( \max_j \left| \frac{1}{N_\nu} \sum_{U_\nu} \left( \frac{\mathbf{I}_{(i \in S_\nu)}}{\pi_i} - 1 \right) a_i(\mathbf{s}_j) \right| > \epsilon \mid \mathcal{F}_\nu \right) \\ & \leq \sum_{j=1}^{N_\nu^{\xi p}} \mathbb{P} \left( \left| \frac{1}{N_\nu} \sum_{U_\nu} \left( \frac{\mathbf{I}_{(i \in S_\nu)}}{\pi_i} - 1 \right) a_i(\mathbf{s}_j) \right| > \epsilon \mid \mathcal{F}_\nu \right) \\ & \leq \frac{1}{\epsilon^2} \sum_{j=1}^{N_\nu^{\xi p}} \text{Var} \left( \frac{1}{N_\nu} \sum_{U_\nu} \left( \frac{\mathbf{I}_{(i \in S_\nu)}}{\pi_i} - 1 \right) a_i(\mathbf{s}_j) \mid \mathcal{F}_\nu \right) \\ & \leq \frac{N_\nu^2 K_1}{N_\nu^2 \epsilon^2} \frac{1}{n_\nu^* - 1} \frac{1}{N_\nu - 1} \sum_{j=1}^{N_\nu^{\xi p}} \sum_i a_i^2(\mathbf{s}_j) \\ & \leq \frac{K_1' N_\nu^{\xi p}}{N_\nu \epsilon^2} \sum_{U_\nu} \mathbf{I}_{(\|d - n_\nu^{*-1/2} \max_j \|\mathbf{s}_j\| \leq \|\mathbf{y}_i - \boldsymbol{\mu}_\nu\| \leq d + n_\nu^{*-1/2} \max_j \|\mathbf{s}_j\|)} \end{aligned}$$

where  $K_1$  and  $K_1'$  are both constants. The last term converges to zero by Assumption 3.6(1).

Additionally,

$$\begin{aligned} & \sup_{\mathbf{s} \in C_j} \left| \frac{1}{N_\nu} \sum_{U_\nu} \left( \frac{\mathbf{I}_{(i \in S_\nu)}}{\pi_i} - 1 \right) (a_i(\mathbf{s}) - a_i(\mathbf{s}_j)) \right| \\ & \leq \frac{1}{n_\nu^{*1/2}} \sup_{\mathbf{s} \in C_j} \sum_{U_\nu} |a_i(\mathbf{s}) - a_i(\mathbf{s}_j)| \\ & \leq K_2 \frac{N_\nu}{n_\nu^{*1/2}} \frac{1}{N_\nu} \sup_{\mathbf{s} \in C_j} \sum_{U_\nu} \mathbf{I}_{(d - n_\nu^{*-1/2} \|\mathbf{s}_j - \mathbf{s}\| \leq \|\mathbf{y}_i - \boldsymbol{\mu}_\nu - n_\nu^{*-1/2} \mathbf{s}_j\| \leq d + n_\nu^{*-1/2} \|\mathbf{s}_j - \mathbf{s}\|)} \\ & \rightarrow 0, \end{aligned}$$

uniformly for all  $j = 1, 2, \dots, N_\nu^{\xi p}$ , as a result of Assumption 3.6(1), where  $K_2$  is a positive constant.

Hence, the proof of (48) is completed by bounding the LHS of (48) by  $\max_j \left| \frac{1}{N_\nu} \sum_{U_\nu} \left( \frac{I_{(i \in S_\nu)}}{\pi_i} - 1 \right) a_i(\mathbf{s}_j) \right|$   
and  $\max_j \sup_{\mathbf{s} \in C_j} \left| \frac{1}{N_\nu} \sum_{U_\nu} \left( \frac{I_{(i \in S_\nu)}}{\pi_i} - 1 \right) (a_i(\mathbf{s}) - a_i(\mathbf{s}_j)) \right|$ .  $\square$

### Proof of Lemma A.1

PROOF. Use Borel-Cantelli Lemma.  $\square$

### Proof of Lemma A.2

PROOF. Define

$$\begin{aligned} X_i &= \left[ \mathbf{I}_{(\|\mathbf{y}_i - \gamma - h_{N_\nu} \mathbf{s}\| \leq d)} - \mathbf{I}_{(\|\mathbf{y}_i - \gamma\| \leq d)} - \mathcal{D}_d(\gamma + h_{N_\nu} \mathbf{s}) + \mathcal{D}_d(\gamma) \right] \\ &= \left[ \mathbf{I}_{(\|\mathbf{y}_i - \gamma - h_{N_\nu} \mathbf{s}\| \leq d, \|\mathbf{y}_i - \gamma\| > d)} - \mathbf{P}(\|\mathbf{y}_i - \gamma - h_{N_\nu} \mathbf{s}\| \leq d, \|\mathbf{y}_i - \gamma\| > d) \right] \\ &\quad - \left[ \mathbf{I}_{(\|\mathbf{y}_i - \gamma - h_{N_\nu} \mathbf{s}\| \leq d, \|\mathbf{y}_i - \gamma\| > d)} - \mathbf{P}(\|\mathbf{y}_i - \gamma - h_{N_\nu} \mathbf{s}\| \leq d, \|\mathbf{y}_i - \gamma\| > d) \right] \\ &\cong X_{1i} - X_{2i} \end{aligned}$$

and  $T_{j, N_\nu} = \frac{h_{N_\nu}}{N_\nu} \sum_{i=1}^{N_\nu} X_{ji}, j = 1, 2$ .

Now, let us show that  $\sup_{(\gamma, \mathbf{s}) \in \mathfrak{R}^p \times C_s} |T_{1, N_\nu}| \xrightarrow{a.s.} 0$ . It is easy to establish that

$$\mathbf{E} X_{1i} = 0 \text{ and } \mathbf{E} \left( \mathbf{I}_{(\|\mathbf{y}_i - \gamma - h_{N_\nu} \mathbf{s}\| \leq d, \|\mathbf{y}_i - \gamma\| > d)} \right) = O(h_{N_\nu}).$$

Without loss of generality, we assume  $\mathbf{E} \left( \mathbf{I}_{(\|\mathbf{y}_i - \gamma - h_{N_\nu} \mathbf{s}\| \leq d, \|\mathbf{y}_i - \gamma\| > d)} \right) \leq h_{N_\nu}$ .

Define

$$g_{\gamma, \mathbf{s}}(\mathbf{y}) = \mathbf{I}_{(\|\mathbf{y} - \gamma - h_{N_\nu} \mathbf{s}\| \leq d, \|\mathbf{y} - \gamma\| > d)},$$

with  $|g_{\gamma, \mathbf{s}}(\mathbf{y})| \leq 1$  and  $\mathbf{E} g_{\gamma, \mathbf{s}}^2(\mathbf{y}_i) \leq h_{N_\nu}$ .

Now we define the graph of  $g_{\gamma, \mathbf{s}}(\mathbf{y})$  as,

$$\begin{aligned} \text{gr}(g_{\gamma, \mathbf{s}}) &= \{(\mathbf{y}, t) | 0 \leq t \leq g_{\gamma, \mathbf{s}}(\mathbf{y})\} \\ &= \{(\mathbf{y}, t) | 0 \leq t \leq \mathbf{I}_{(\|\mathbf{y} - \gamma\| > d)}\} \cap \{(\mathbf{y}, t) | 0 \leq t \leq \mathbf{I}_{(\|\mathbf{y} - \gamma - h_{N_\nu} \mathbf{s}\| \leq d)}\}. \end{aligned}$$

Both of the two sets of graphs are translation families of  $\{(\mathbf{y}, t) | 0 \leq t \leq \mathbf{I}_{(\|\mathbf{y} - \gamma\| > d)}\}$ , which has polynomial discrimination in  $\mathfrak{R}^p$  by Lemma B.1 (ii) of Opsomer and Ruppert (1997). Lemma A.4 (ii) of the same paper states that both  $\{(\mathbf{y}, t) | 0 \leq t \leq \mathbf{I}_{(\|\mathbf{y} - \gamma\| > d)}\}$  and  $\{(\mathbf{y}, t) | 0 \leq t \leq \mathbf{I}_{(\|\mathbf{y} - \gamma - h_{N_\nu} \mathbf{s}\| \leq d)}\}$  have

polynomial discrimination as well. That  $\text{gr}(g_{\gamma,s})$  has polynomial discrimination can be asserted given Lemma II.15 and Corollary II.17 of Pollard (1984).

Now, everything is set up for applying Theorem II.37 of Pollard (1984) to show the uniform almost sure convergence.  $\square$

DEPARTMENT OF STATISTICS AND STATISTICAL LABORATORY  
IOWA STATE UNIVERSITY  
AMES, IA, 50011-1210  
USA

DEPT OF STATISTICS  
COLORADO STATE UNIVERSITY  
FORT COLLINS, CO 80523-1877  
USA  
E-MAIL: jqwang@iastate.edu  
jopsomer@stat.colostate.edu