Project description
Applications of quasi-Monte Carlo methods in survey estimation

1 Introduction

This project aims at proposing a new method for estimating variances of complex survey estimators based on the recent developments in quasi-Monte Carlo methods. It promises to be an effective tool in survey practice in complex surveys where the mathematically elegant resampling schemes such as balanced repeated replications break down due to design complexities, while other methods such as the survey bootstrap carry with them a substantial computational burden. The new method is based on quasi-Monte Carlo methods, such as multidimensional Halton sequences, that find a multitude of uses in various areas of mathematics, physics, engineering, econometrics in approximation of multidimensional integrals in complex scientific simulations and in estimation methods that require multidimensional integration.

2 Resampling variance estimation in complex surveys

Complex sample surveys are the major source of information in social science research, as well as in decision making in political and economics processes. In those surveys, information is collected from the units of analysis that are reached through a complex sampling scheme that may involve stratification (i.e., sectioning of the population according to a characteristic known before sampling is taken, e.g., geographical region for samples of individuals, or industry for samples of establishments), clustering (using collections of units to ensure feasibility, as full lists of units may not be available), and probability weights (used sometimes to oversample smaller subpopulations). For general mathematical results in survey statistics, see Cochran (1977), Hansen, Hurwitz & Madow (1953), Chambers & Skinner (2003), Lehtonen & Pahkinen (2004), Chaudhuri & Stenger (2005). For an excellent survey of the history and state of the art in statistical aspects of survey estimation, see Rao (2005).

There are several approaches for statistical inference, i.e., point and interval estimation and estimation of variance, based on complex surveys. The predominant ones are design-based and model-based estimation approaches (Binder & Roberts 2003). The design-based methods assume that the measured characteristics of individuals are fixed, and the randomness is due to the sampling process, and thus the inference is based on randomization paradigm, with expectations taken over discrete probability spaces of distinct samples. In the model-based methods, the observations are assumed to be coming from (an infinite) superpopulation, similarly to the traditional i.i.d. assumption often made in mathematical statistics. Design-based methods

appear to be more robust to both model violations and complexities of sample designs, while the model-based methods are more efficient when the assumed model is correct. There has been some work in merging those two approaches (Särndal, Swensson & Wretman 1992), as well as looking at non-sampling error (Lesser & Kalsbeek 1992, Groves, Couper, Lepkowski, Singer & Tourangeau 2004).

Within the (dominant) design-based approach to inference, the two major directions for variance estimation of parameters in nonlinear models, including regression models, GLMs, structural equation models, multilevel models, and others, can be identified: linearization/Taylor series expansion, and resampling methods. The former aims at analytic derivation of the variances of complex survey estimators using the delta method leading to sandwich estimators of variance (Binder 1983, Skinner 1989). The resampling methods create a series of subsets, or replicates, of the observed data, and estimate the variance of interest through variability in the estimates of the same parameter in those replicates. Those methods are strictly applicable when the first stage units are taken with replacement, and also show good performance in without replacement designs with small sampling fractions. Other methods such as generalized variance functions (Huff, Eltinge & Gershunskaya 2002) are used occasionally as well.

The three major resampling methods used in complex survey inference include balanced repeated replication (BRR), the jackknife and the bootstrap. See Ch. 6 of Shao & Tu (1995), Shao (1996), Ch. 9 of Chaudhuri & Stenger (2005), for description, reviews and discussions. Somewhat less technical recommendations for practitioners are also given in Brick, Morganstein & Valliant (2000) and Phillips (2004). In each of those methods, the PSUs are carefully reshuffled, so that in r-th replication, some PSUs are omitted, and some are included (may be multiple times, as in the bootstrap). The statistic of interest $\hat{\theta}^{(r)}$ for this subsample of data is computed, and the process is repeated R times. The resulting estimator of variance is defined by a standard formula

$$\hat{\mathbb{V}}_R[\hat{\theta}] = \frac{1}{R-1}\sum_{r=1}^{R}(\hat{\theta}^{(r)} - \tilde{\theta})^2 \tag{1}$$

where some possible variations include (i) taking R in the denominator, and (ii) using either the estimate based on the original sample $\hat{\theta}$ for the estimate of location $\tilde{\theta}$, leading to MSE-type estimator, or the average of the resampled values $\tilde{\theta} = 1/R\sum_r \hat{\theta}^{(r)}$, leading to variance-type estimator. The resampling methods of variance estimation then differ in the resampling designs, i.e., the patterns of included and excluded PSUs.

Practical advantages of the resampling methods compared to the linearization estimator include:

- general applicability: the resampling estimators can be applied for an arbitrary estimation procedure, while the linearization estimators require computation of the gradient of the objective function, either analytic or numeric
- lower disclosure risks: no identifiers need to be released in the public data files

It has been observed, however, that the resampling variance estimators are less stable[1] than the linearization estimator.

The balanced repeated replication method was proposed by McCarthy (1969) for designs with $n_h = 2$ primary sampling units (PSU) per stratum. The set of replicates is formed by taking one of the sampling

[1] Stability is formally defined as $\frac{\left(MSE\left[\hat{\mathbb{V}}[\hat{\theta}]\right]\right)^{\frac{1}{2}}}{MSE[\hat{\theta}]}$, and is a measure of variability of the variance estimator itself.

units from each stratum, leading to a total of L units where L is the number of strata. Hadamard matrices (Hedayat, Sloane & Stufken 1999) can then be used to find optimal resampling designs[2], which helps in reducing the number of subsamples R_{BRR} from 2^L to the smallest multiple of four greater than or equal to L where L is the number of strata. Extensions for $n_h > 2$ have also been proposed (Gurney & Jewett 1975, Gupta & Nigam 1987, Wu 1991, Sitter 1993), but their use does not appear to be very wide spread. The difficulties with those extensions is that the design matrices are not easily available, and the number of replicates dictated by the orthogonality constraints is combinatorially large. Gurney & Jewett (1975) analyze the design with L strata each having p PSUs, and in their approach, the number of replicates is p^L. Wu (1991) and Sitter (1993) proposed more efficient designs leading to at most $(p-1)(L+4)$ replicates. In unbalanced situations where the number of PSUs varies across strata (Gupta & Nigam 1987, Wu 1991, Sitter 1993), the number of replicates is given by expressions combinatorial in n_h's. Approximate BRR methods have also been proposed, see discussion in Sec. 6.2.3 of Shao & Tu (1995).

The next resampling method of variance estimation is the jackknife, which is a generalization of the traditional i.i.d. jackknife for complex survey situations. In the jackknife for i.i.d. samples, the statistic of interest $\hat{\theta}_{(r)}$ is computed omitting the r-th observation from the sample, and keeping all other observations. The generalization of (delete-1) jackknife to the complex surveys consists of taking out the whole PSU in a given stratum to preserve the dependence structure within clusters, thus generating the number of replicates equal to the total number of clusters, $R_J = n_1 + \ldots + n_L$. The deficiency of the delete-1 estimator is that it is inconsistent for non-smooth functions such as quantiles. A more general formulation where d PSUs are omitted simultaneously from h-stratum can be considered that remedies the situation. Those methods however involve a need for re-weighting the data, discussed below in the context of the bootstrap.

Consistency of both the jackknife and BRR estimators is established in Krewski & Rao (1981).

The appropriate modifications of the popular bootstrap method (Hall 1992, Efron & Tibshirani 1994, Shao & Tu 1995) that consists of taking samples *with replacement* from the empirical distribution of the data, have been discussed in Rao & Wu (1988) and Rao, Wu & Yue (1992). If m_h PSUs are selected out of n_h available in the sample, then Rao & Wu (1988) showed that the pseudo-values $\theta^{(r)}$ are to be obtained using the rescaled data

$$\tilde{y}_{hi}^{(r)} = \bar{y}_h + \sqrt{\frac{m_h}{n_h - 1}}(y_{hi}^{(r)} - \bar{y}_h)$$

where $\bar{y}_h$ is the mean in h-th stratum in the original sample, and $y_{hi}^{(r)}$ is the estimate of it obtained in r-th subsample, with some of the PSUs included among the m_h, and others excluded. They have also discussed the choice of optimal m_h

$$m_h = \begin{cases} n_h - 1, \\ \dfrac{(n_h-2)^2}{n_h - 1} \approx n_h - 3, \quad n_h > 3 \end{cases} \tag{2}$$

where the first choice eliminates the need for re-weighting the data (and also leads to the random half-sample replication with $n_h = 2$), while the second choice is motivated by tracking the third moment of data leading to the second term of Edgeworth expansion.

[2] In terms of the traditional design of experiments literature, the units are interpreted as levels of a factor, and each stratum is interpreted as a factor. Unlike the design of experiments literature, however, it is possible to have several units (levels of a factor) for a given "observation" (r-th subsample of the data).

Rao et al. (1992) observed that all of the above methods lead to the subsample based estimates that have the form

$$y_{hi}^{(r)} = \sum_{j \in \mathcal{S}^{(r)}} w_{hj}^{(r)} y_{hj} + \sum_{j' \notin \mathcal{S}^{(r)}} w_{hj'}^{(r)} y_{hj'} \tag{3}$$

where the weight $w_{hj}^{(r)}$ increases or decreases relative to the original probability weight of a unit[3] depending on whether the unit is included to or excluded from the r-th subsample. Thus for the balanced repeated replications, the resampling weights is the double of the original weight for the units included in the sample, and zero for the excluded units. For the jackknife, the resampling weights are

$$w_{hj,J}^{(r)} = \begin{cases} w_{hi}, & h \neq g \\ 0, & h = g \text{ and } i = j \\ \frac{n_g}{n_g - 1} w_{gi}, & h = g \text{ and } i \neq j \end{cases} \tag{4}$$

where the r-th subsample omits j-th unit from g-th stratum. Finally, for the bootstrap, the proposed reweighting scheme is (Rao et al. 1992, eq (3.4))

$$w_{hj,B}^{(r)} = \Big[\big(1 - \Big\{ \frac{m_h}{n_h - 1} \Big\}^{1/2} \big) + \big(\Big\{ \frac{m_h}{n_h - 1} \Big\}^{1/2} \frac{n_h}{m_h} m_{hi}^{(r)} \big) \Big] w_{hi} \tag{5}$$

where $m_{hi}^{(r)}$ is the number of times the i-th unit from the h-th stratum appears in r-th subsample. This innovation made the resampling weights very practical for the data collecting agencies, who now can provide their data with the sets of BRR, the jackknife or the bootstrap weights for the benefit of the data user. It should be noted that if any modifications are made to the original probability weights such as post-stratification and non-response adjustments, similar modifications must be made on every set of resampling weights.

Other variants of the complex survey bootstrap are discussed in Sec. 6.2.4 of Shao & Tu (1995). In particular, first-order balance in general and second-order for some balanced sample designs can be achieved by the balanced bootstrap (Nigam & Rao 1996).

Under "normal" circumstances that include consistency of the point estimate of interest, and smoothness of the statistic[4] , the linearization and all of the resampling methods are consistent (Krewski & Rao 1981) and asymptotically equivalent to one another. In fact, linearization, BRR and jackknife estimators coincide for linear statistics. Comparisons of higher order properties and performance in simulations of different resampling methods are given in Rao & Wu (1985), Rao & Wu (1988) and Shao (1996). The latter notes that "...the choice of the method may depend more on nonstatistical considerations, such as the feasibility of their implementation". Unless the sample design is that of two PSUs per stratum, the method of choice is often the bootstrap that does not require construction of complicated mixed orthogonal arrays, and hides the groups of units for the purposes of the disclosure better than the jackknife.

[3] Such a weight accounts for different probabilites of selection, nonresponse, post-stratification and other adjustments performed by the data provider.

[4] The delete-1 jackknife is known to be inconsistent for quantiles.

3 Quasi-Monte Carlo methods

The term "quasi-Monte Carlo" refers to the set of methods of generating deterministic *point sets* and *nets* that achieve highly uniform coverage of the unit cube $[0,1]^s$. Those methods find most applications in numeric integration (Neiderreiter 1992), with some application to stochastic processes modeling (Fox 1999), and they also proliferate in econometrics of discrete choice modeling (Train 2001, Bhat 2001) where the method is used for computation of multivariate CDFs in multinomial probit and logit models. There is a regular conference on Monte Carlo and quasi-Monte Carlo methods with published proceedings (Niederreiter 2004).

The primary measure of performance of an s-dimensional point set $P = (x_1, \ldots, x_N), x_n \in [0,1]^s$ of size N is the *discrepancy* between the counting measure implied by P and the Lebesgue measure of the corresponding set:

$$A(B,P) = \sum_{n=1}^{N} \mathbb{1}_B(x_n), \tag{6}$$

$$D_N(\mathcal{B};P) = \sup_{B\in\mathcal{B}} \Big| \frac{A(B,P)}{N} - \lambda(B) \Big|, \tag{7}$$

$$D_N^*(P) = D_N(\mathcal{I}^*;P), \quad \mathcal{I}^* = \Big\{ \prod_{i=1}^{s} [0,u_i), 0 \le u_i \le 1 \Big\} \tag{8}$$

$$D_N(P) = D_N(\mathcal{I};P), \quad \mathcal{I} = \Big\{ \prod_{i=1}^{s} [u_i,v_i), 0 \le u_i \le v_i \le 1 \Big\} \tag{9}$$

where $\mathbb{1}_B(x)$ is an indicator/charateristic function of a set B.

It can be easily established that if an integral of a function of bounded variation $V(f)$ on $[0,1]$ is approximated by an average over the set P, then the approximation error is given by $V(f)D_N^*(P)$, with an appropriate generalization of the variation concept to multiple dimension. The question then arises of a "good" choice of P. Apparently, the traditional (random) Monte Carlo methods achieve stochastic coverage with discrepancy of order $O_p(N^{-1/2})$.

It has been shown that in the case of single dimension the discrepancies have lower bounds $D_N(S) \ge cN^{-1}\ln N$ for infinitely many N, with some knowledge of explicit constants c (the best value given in Neiderreiter (1992) is $c = 0.12$). The sequences that achieve the asymptotic behavior have been identified, and we shall outline the construction of several of them and state the corresponding discrepancy results following Neiderreiter (1992).

For an integer $b > 2$, the *radical inverse function in base* b is

$$\phi_b(n) = \sum_{j=0}^{\infty} a_j(n) b^{-j-1} \tag{10}$$

where $a_j(n)$ are coefficients of the digit expansion of n,

$$n = \sum_{j=0}^{\infty} a_j(n) b^j \tag{11}$$

Note that $\phi_b(n) \in [0,1)$. For an integer $b > 2$, the *van der Corput sequence in base* b is the sequence $\{\phi_b(n)\}_{n=0}^{\infty}$. It has been established that the discrepancy of this (unidimensional) sequence is

$$\overline{\lim_{N\to\infty}} \frac{ND_N^*(S_b)}{\ln N} = \overline{\lim_{N\to\infty}} \frac{ND_N(S_b)}{\ln N} = \begin{cases} \dfrac{b^2}{4(b+1)\ln b}, & b \in 2\mathbb{N} \\ \dfrac{b-1}{4\ln b}, & b \in 2\mathbb{N}+1 \end{cases} \tag{12}$$

Some improvements in the leading constants are possible with permutations of the digits in base b, and the resulting sequences are referred to as *generalized van der Corput sequences*.

A generalization of van der Corput sequences into multiple dimensions is provided by the *Halton sequences*. For a dimension s, let $b_1, \ldots, b_s$ be integers greater than 1. Then the *Halton sequence in the bases* $b_1, \ldots, b_s$ is the sequence

$$x_n = (\phi_{b_1}(n), \ldots, \phi_{b_s}(n)) \tag{13}$$

If S is the Halton sequence in pairwise relatively prime bases $b_1, \ldots, b_s$, then

$$D_N^*(S) < \frac{s}{N} + \frac{1}{N}\prod_{i=1}^{s}\Big(\frac{b_i-1}{2\ln b_i}\ln N + \frac{b_i+1}{2}\Big) = A(b_1,\ldots,b_s)N^{-1}\ln^s N + O(N^{-1}\ln^{s-1} N) \tag{14}$$

That is a major asymptotic improvement over $O_p(N^{-1/2})$ achieved by the random Monte Carlo methods. The numbers b are usually taken to be the first s primes, as that delivers the minimum to the leading term A. There is also an associated curse of dimensionality, however, as $\ln A \sim s \ln s$, i.e., A grows exponentially fast with s, and the discrepancy may still be large for small to moderate N and large s.

Various modification of the Halton sequences aimed at reducing "autocorrelations" between components of the sequence with high b's, combating the dimensionality curse, and providing for estimation of the (randomization, or design) standard errors have been proposed. Owen (1998) reviews the work on randomized quasi-Monte Carlo, such as random rotations and permutations of the Halton sequences. Bhat (2003) applies those ideas to the empirical research in individual transportation choices.

There is a rising interest in quasi-Monte Carlo methods among statisticians reflected by a recent NSF award DMS #0306612 to Prof. Art Owen. His work in this area includes Latin supercube sampling, Monte Carlo extensions, applications in Markov chain Monte Carlo, dimensionality issues, as well as general reviews aimed at statistics audience (Owen 2003, Owen 2005*b*, Owen 2005*a*, Owen & Tribble 2005)

Computation of Halton sequences is available through commercial Stata software (Drukker & Gates 2006) as well as through free GPL R package (Maechler 2006). Both of those will be used in prospective work to ensure availability of the methodology to the general public, as well as to provide a check on correct implementation between the two computing platforms.

4 The proposed estimator

We are proposing an extension of the balanced repeated replication method for the arbitrary designs, including those that do not necessarily have $n_h = 2$ where the (approximately) balanced designs are constructed with the use of the quasi-Monte Carlo methods. The proposed method relaxes the limitations on the number

of replicates imposed by the availability of mixed orthogonal arrays. Alternatively, the quasi-Monte Carlo based resampling designs can be viewed as a quasi-Monte Carlo bootstrap scheme.

The following research questions need to be addressed.

(i) Proof of consistency of $\hat{\mathbb{V}}_{QMC}$. It is expected to be similar to the proofs in Krewski & Rao (1981), and may also involve non-orthogonality considerations of Wu (1991), with the second-order imbalance characterized by bivariate discrepancy in (14). For other work with imbalanced resampling designs, see Lee (1972), Lee (1973), Sitter (1993).

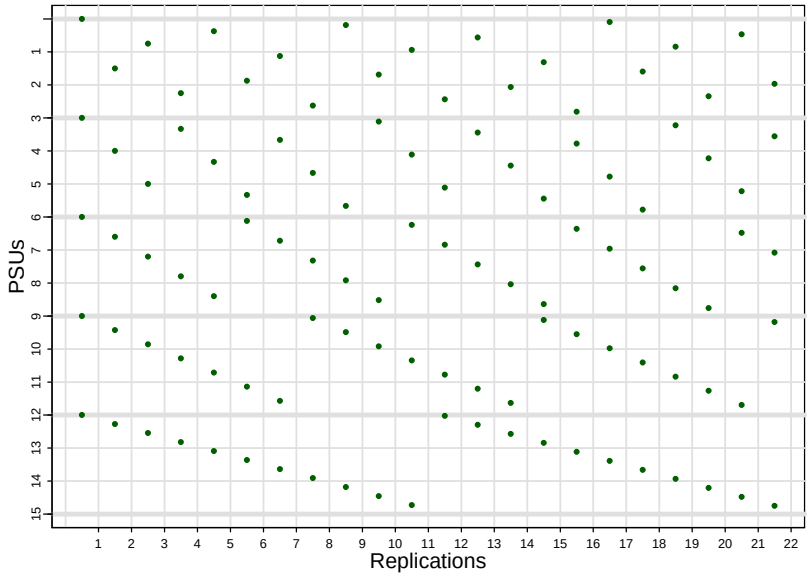
(ii) A specific implementation of QMC. Two possibilities are being entertained so far. One approach might be to use L-dimensional Halton sequence, or a (randomized, permuted, shifted) variation of it, generating the indices of the units to be included in the sample separately for every stratum. This however leads to aforementioned dimensionality curse, and thus may only be reasonable for small L. Fig. 1 (a) gives an example with with $L = 5$ and $n_h = 3$ PSUs per stratum, for a total of $n = 15$ PSUs. The design allocates units 1, 4, 7, 10, 13 to the first replicate; 2, 5, 7, 10, 13 to the second replicate, etc. Note high degree of "autocorrelation" in the large strata numbers, as well as lack of balance (unit #9 is sampled 10 times, while other units in the same stratum are sampled only 6 times). Another alternative is to view the resampling design matrix as a two-dimensional object with say PSUs on the vertical axis, and the replicates on the horizontal axis, as shown on Fig. 1 (b). The Halton sequence in bases (2,3) generates a (both-ways unbalanced) resampling design with $R = 12$ replicates and $\mathcal{N} = 60$ first elements of the Halton sequence. The design picks units 1, 3, 8, 9, 12, 14 for the first replicate; 1, 5, 7, 11, 12 for the second replicate, etc. Note that the resulting designs are first-order unbalanced for PSU. In some replicates, some of the strata contribute zero observations. Apparently, the asymptotic smaller order discrepancy is not kicking in yet for those small $\mathcal{N}$. Thus some additional balancing will be required for either version of QMC implementation. Also, randomized variations of QMC can be employed. An example of a resampling design based on shuffled/permuted Halton sequence is given on Fig. 1 (c).

(iii) An optimal choice of the number of replicates and the number of elements in the Halton sequence needs to be addressed. The number of replicates should at the very least exceed the degrees of freedom of the design, $n-L$. Orthogonality of the resulting design dictates that the number of replicates be proportional to the smallest common multiple of products $n_h \cdot n_{h'}, h \neq h' = 1, \ldots, L$. The number of elements in the Halton sequence that would deliver a good degree of balance would be (the multiple of) the product $b_1 \cdot \ldots \cdot b_L$ for the stratum-by-stratum implementation, and it increases more than exponentially fast in L. On the other hand, the number of elements of the sequences in the two-dimensional array representation of the design matrix needs to be proportional to $2 \cdot 3 = 6$, which is a reasonably mild restriction.

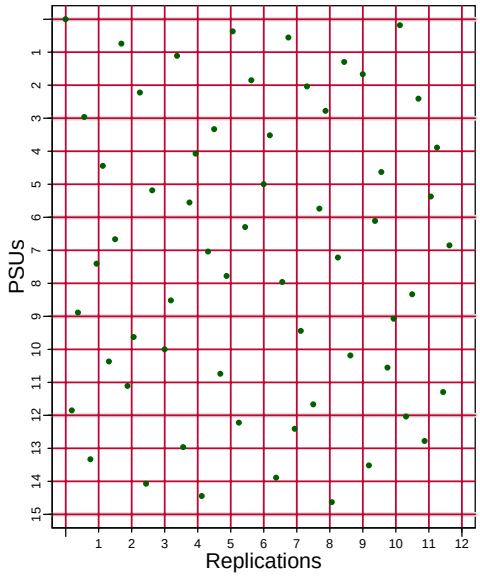
(iv) Related to the previous one is the question of the choice of the number of units to be resampled in each replicate. As mentioned above, Rao & Wu (1988) suggested the choice of m_h such as (2). The fractional average, across replication, number of units to be sampled can be achieved with QMC by taking the nearest integer to $\mathcal{N} = \dfrac{(n_h - 2)^2}{n_h - 1} R$ elements of the Halton sequences. The internal scaling of Wu (1986) and Rao et al. (1992) may then also be necessary. Note that with stratum-by-stratum implementation of QMC, this fractional choice can be taken within each of the strata.

Other research questions are also likely to arise along the way.

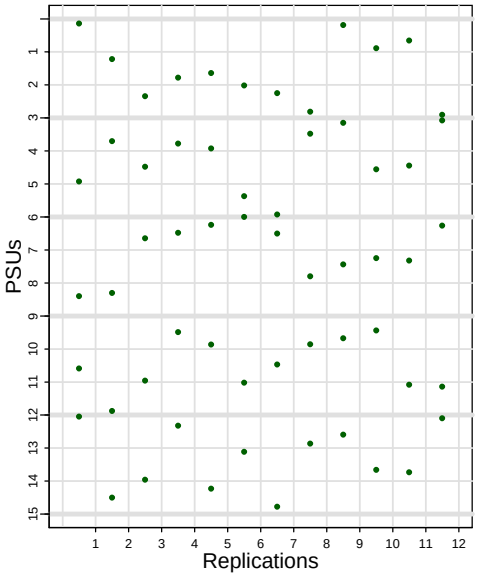
Implementation of the proposed procedures will be made available with the use of existing software sup-



(a)



(b)



(c)

Figure 1: Quasi-Monte Carlo resampling design (a) based on 5-dimensional Halton sequence in bases (2,3,5,7,11), with $n = 15$ PSUs, $R = 22$ replicates corresponding to $\mathcal{N} = 22$ points in the sequence; (b) based on Halton sequence in bases (2,3), with $n = 15$ PSUs, $R = 12$ replicates, and $\mathcal{N} = 60$ points in the sequence; (c) based on Halton sequence in bases (2,3,5,7,11), with $n = 15$ PSUs, $R = 12$ replicates, and $\mathcal{N} = 12$ points in the sequence, permuted from the first $\mathcal{N} = 55$ elements of Halton sequence.

porting Halton sequences and resampling weights. Stata software (Stata Corp. 2005) is an obvious choice, as it has the widest range of design-based estimators for survey data, including the resampling estimators, as well as a set of tools for generating Halton sequences (Drukker & Gates 2006). It is well known in applied areas such as economics and epidemiology, easily programmable, and the third party packages can be easily shared by the users (Hilbe 2005). Another package, more popular among mathematical statisticians, is R, and the modules for Halton sequences are also available for it, as well (Maechler 2006).

A simulation study will be necessary to verify the asymptotic properties of the new method in finite samples, compare different choices of the QMC implementations outlined above, and compare the QMC-based resampling variance estimator with the jackknife and the bootstrap. A high performance computing platform is necessary for this work, and is requested in the proposal budget.

5 Summary of project and its likely contributions

The project is aimed at integrating quasi-Monte Carlo methods into survey inference methodology and practice. The quasi-Monte Carlo algorithms appear to have good promise in generating (approximately) balanced resampling designs, achieving the compromise between designs based on mixed orthogonal arrays, and the bootstrap methods. The former achieve exact balance in relatively simple situations, as well as first-order and second-order balance in more complex situations at the expense of combinatorial complexity in terms of the number of replicates that are needed to achieve this balance. The bootstrap methods, on the other hand, are not aimed at the exact balance, but nevertheless achieve the balance in asymptotic sense due to independent random resampling of the data; on the other hand, they are typically less stable for the number of replicates typically used (around 500).

We plan to have the following outcomes of the project, in approximate time order:

- Literature review; software implementation in Stata and R; preliminary analysis based on this implementation

- A thorough analysis of i.i.d. case, as the simplest case possible, and an accompanying research paper. This will involve proof of consistency, with a double asymptotics in the sample size and the number of elements in Halton sequence or its generalizations, and comparison to the jackknife and the bootstrap in terms of performance (stability of estimators, accuracy of coverage of the confidence intervals) in small ($n \sim 10^1 - -10^2$), moderate ($n \sim 10^2$) and large ($n > 10^3$) samples for linear (means, totals) and nonlinear (ratios, coefficients in linear and logistic regression) statistics. Outlets for such paper might be journals like JASA, JRSS B, JSPI, JCGS.

- A thorough analysis of the complex survey case, using the expertise accumulated during the previous stages of the project, and preparation of the research paper (JASA, JRSS B, Survey Methodology). Again, the proof of consistency lies in the foundation of this work, but, as outlined above, several variations for complex survey samples might be considered, so the proofs are likely to be specific to those methods. As was shown in a simple demonstration in the previous section, additional balancing might be required that might complicate things. In parallel, a simulation study will be conducted to compare

performance of different versions of the QMC-based resampling designs for surveys of different degrees of complexity. E.g., in case of few strata with many PSUs per stratum, an L-dimensional Halton sequence may provide better results due to direct applicability of the designs, while with many strata, the approach may become infeasible, and 2-dimensional array representation may be the main choice.

- Once an understanding of the performance of the QMC-based resampling methods is achieved, the recommendations will be developed for the data providers, for generation of resampling weights, and for the data users, for use of those weights in the software modules developed within this project or in other general software that allows for resampling estimation. The resulting applied paper will be published in the software specific outlet (Stata Journal) and statistical journals for a broader audience (TAS, JRSS A) or journals in the field of survey statistis (Survey Methodology, Journal of Official Statistics).

Let us now summarize the proposed work in the light of NSF review criteria.

This project will assess applicability of quasi-Monte Carlo methods in generating resampling designs for complex survey data, thus establishing a new variance estimation method. It will integrate work in survey statistics with recent development in numerical integration in mathematics, and, to a lesser extent, computational econometrics. Thus we believe there is substantial *intellectual merit* to the proposed research.

The project will also have *broader impacts* provided sufficient funding is available. The important *training and education* component will be addressed through involvement of graduate student(s) in all stages of the project who will also act as co-authors on the papers. This activity is likely to generate a defendable dissertation in the field of survey statistics that is known to have relative lack of new researchers. The proposed work will also strengthen the doctoral level course on complex survey statistics currently taught by the PI at his Department. The *physical infrastructure* generated by the project, although limited to a Linux workstation, will be available for other projects by the PI, his students, and other members of Department of Statistics. The *informal networks* may be created by participation of the researchers involved in the project in national and international statistics conferences, and visits to interact with other researchers in related fields.

As outlined above, the project results will be *broadly disseminated* through presentations, research article publications and software module uploads at publicly accessible archives (SSC for Stata software; CRAN for R software).

Finally, the project will have *impact on the research society* through the survey statistics practices in provision of the public data files, and in the users of the data practice of using resampling methods for survey variance estimation. Involvement of the PI in other projects, both in mathematical statistics, as well as in interdisciplinary collaborative work with researchers in sociology and psychology, is likely to be very effective in generating this broad impact.

References

Bhat, C. R. (2001), 'Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model', *Transportation Research Part B: Methodological* **35**(7), 677–693.

Bhat, C. R. (2003), 'Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences', *Transportation Research Part B: Methodological* **37**(9), 837–855.

Binder, D. A. (1983), 'On the variances of asymptotically normal estimators from complex surveys', *International Statistical Review* **51**, 279–292.

Binder, D. A. & Roberts, G. R. (2003), Design-based and model-based methods for estimating model parameters, *in* R. L. Chambers & C. J. Skinner, eds, 'Analysis of Survey Data', John Wiley & Sons, New York, chapter 3.

Brick, J. M., Morganstein, D. & Valliant, R. (2000), Analysis of complex sample data using replication, technical report, Westat. http://www.westat.com/wesvar/techpapers/ACS-Replication.pdf.

Chambers, R. L. & Skinner, C. J., eds (2003), *Analysis of Survey Data*, Wiley series in survey methodology, Wiley, New York.

Chaudhuri, A. & Stenger, H. (2005), *Survey Sampling: Theory and Methods*, Vol. 181 of *Statistics: Textbooks and Monographs*, 2nd edn, Chapman & Hall/CRC, Boca Raton, FL.

Cochran, W. G. (1977), *Sampling Techniques*, John Wiley and Sons, New York.

Drukker, D. M. & Gates, R. (2006), 'Generating Halton sequences using Mata', *Stata Journal* **6**(2), 214–228.

Efron, B. & Tibshirani, R. J. (1994), *An Introduction to the Bootstrap*, Chapman & Hall/CRC.

Fox, B. L. (1999), *Strategies for Quasi-Monte Carlo*, Springer.

Groves, R. M., Couper, M. P., Lepkowski, J. M., Singer, E. & Tourangeau, R. (2004), *Survey Methodology*, Wiley Series in Survey Methodology, John Wiley and Sons, New York.

Gupta, V. K. & Nigam, A. K. (1987), 'Mixed orthogonal arrays for variance estimation with unequal numbers of primary selections per stratum', *Biometrika* **74**(4), 735–742.

Gurney, M. & Jewett, R. S. (1975), 'Constructing orthogonal replications for variance estimation', *Journal of the American Statistical Association* **70**(352), 819–821.

Hall, P. (1992), 'The bootstrap and edgeworth expansion'.

Hansen, M., Hurwitz, W. N. & Madow, W. G. (1953), *Sample Survey Methods and Theory*, John Wiley and Sons, New York.

Hedayat, A., Sloane, N. J. A. & Stufken, J. (1999), *Orthogonal Arrays: Theory and Applications*, Springer Series in Statistics, Springer-Verlag, New York.

Hilbe, J. M. (2005), 'A review of stata 9.0', *The American Statistician* **59**(4), 335–348.

Huff, L. L., Eltinge, J. L. & Gershunskaya, J. (2002), Exploratory analysis of generalized variance function models for the U.S. Current Employment Survey, technical report 020120, BLS. http://www.bls.gov/ore/abstract/st/st020120.htm.

Krewski, D. & Rao, J. N. K. (1981), 'Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods', *The Annals of Statistics* **9**(5), 1010–1019.

Lee, K.-H. (1972), 'Partially balanced designs for half sample replication method of variance estimation', *Journal of the American Statistical Association* **67**(338), 324–334.

Lee, K.-H. (1973), 'Using partially balanced designs for the half sample replication method of variance estimation', *Journal of the American Statistical Association* **68**(343), 612–614.

Lehtonen, R. & Pahkinen, E. (2004), *Practical Methods for Design and Analysis of Complex Surveys*, Statistics in Practice, John Wiley & Sons, New York.

Lesser, V. M. & Kalsbeek, W. D. (1992), *Non-sampling Error in Surveys*, John Wiley and Sons, New York.

Maechler, M. (2006), *The* `sfsmisc` *package*, R Comprehensive Network Archive (CRAN). http://cran.hu.r-project.org/doc/packages/sfsmisc.pdf.

McCarthy, P. J. (1969), 'Pseudo-replication: Half samples', *Review of the International Statistical Institute* **37**(3), 239–264.

Neiderreiter, H. (1992), *Random Number Generation and Quasi-Monte Carlo Methods*, Vol. 63 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, Society for Industrial and Applied Mathematics, Philadelphia.

Niederreiter, H., ed. (2004), *Monte Carlo and Quasi-Monte Carlo Methods 2002: Proceedings of a Conference held at the National University of Singapore, Republic of Singapore, November 25-28, 2002*, Springer-Verlag, New York.

Nigam, A. K. & Rao, J. N. K. (1996), 'On balanced bootstrap for stratified multistage samples', *Statistica Sinica* **6**(1), 199–214.

Owen, A. (2005*a*), On the Warnock-Halton quasi-standard error, technical report, Stanford University. http://www-stat.stanford.edu/ owen/reports/qse.pdf.

Owen, A. B. (1998), Monto Carlo extension of quasi-Monte Carlo, *in* 'WSC '98: Proceedings of the 30th conference on Winter simulation', IEEE Computer Society Press, Los Alamitos, CA, USA, pp. 571–578.

Owen, A. B. (2003), Quasi-Monte Carlo sampling, *in* H. W. Jensen, ed., 'Monte Carlo Ray Tracing: Siggraph 2003 Course 44', SIGGRAPH, pp. 69–88.

Owen, A. B. (2005*b*), Multidimensional variation for quasi-Monte Carlo, *in* J. Fan & G. Li, eds, 'International Conference on Statistics in honour of Professor Kai-Tai Fang's 65th birthday', pp. 49–74.

Owen, A. B. & Tribble, S. D. (2005), 'A quasi-Monte Carlo Metropolis algorithm', *Proceedings of the National Academy of Sciences of the USA* **102**, 8844–8849. doi:10.1073/pnas.0409596102.

Phillips, O. (2004), Using bootstrap weights with WesVar and SUDAAN, Technical Report 2, Statistics Canada.

Rao, J. N. K. (2005), 'Interplay between sample survey theory and practice: An appraisal', *Survey Methodology* **31**(2), 117–138.

Rao, J. N. K. & Wu, C. F. J. (1985), 'Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics', *Journal of the American Statistical Association* **80**(391), 620–630.

Rao, J. N. K. & Wu, C. F. J. (1988), 'Resampling inference with complex survey data', *Journal of the American Statistical Association* **83**(401), 231–241.

Rao, J. N. K., Wu, C. F. J. & Yue, K. (1992), 'Some recent work on resampling methods for complex surveys', *Survey Methodology* **18**(2), 209–217.

Särndal, C.-E., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer, New York.

Shao, J. (1996), 'Resampling methods in sample surveys', *Statistics* **27**, 203–254. with discussion.

Shao, J. & Tu, D. (1995), *The Jackknife and Bootstrap*, Springer, New York.

Sitter, R. R. (1993), 'Balanced repeated replications based on orthogonal multi-arrays', *Biometrika* **80**(1), 211–221.

Skinner, C. J. (1989), Domain means, regression and multivariate analysis, *in* C. J. Skinner, D. Holt & T. M. Smith, eds, 'Analysis of Complex Surveys', Wiley, New York, chapter 3, pp. 59–88.

Stata Corp. (2005), *Stata Statistical Software: Release 9*, College Station, TX, USA.

Train, K. (2001), Halton sequences for mixed logit, Econometrics Working Paper 0012002, EconWPA. available at http://ideas.repec.org/p/wpa/wuwpem/0012002.html.

Wu, C. F. J. (1986), 'Jackknife, bootstrap and other resampling methods in regression analysis', *The Annals of Statistics* **14**(4), 1261–1295.

Wu, C. F. J. (1991), 'Balanced repeated replications based on mixed orthogonal arrays', *Biometrika* **78**(1), 181–188.