Statistics and Probability Letters 113 (2016) 41–48

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Nonparametric Bayes modeling with sample survey weights

T. Kunihama [a,*], A.H. Herring [c], C.T. Halpern [d], D.B. Dunson [b]

[a] Department of Statistics, University of Washington, Seattle, WA 98195, USA

[b] Department of Statistical Science, Duke University, Durham, NC 27708, USA

[c] Department of Biostatistics and Carolina Population Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[d] Department of Maternal and Child Health and Carolina Population Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

ARTICLE INFO

Article history:
Received 23 January 2016
Accepted 17 February 2016
Available online 4 March 2016

Keywords:
Biased sampling
Dirichlet process
Mixture model
Stratified sampling
Survey data

ABSTRACT

In population studies, it is standard to sample data via designs in which the population is divided into strata, with the different strata assigned different probabilities of inclusion. Although there have been some proposals for including sample survey weights into Bayesian analyses, existing methods require complex models or ignore the stratified design underlying the survey weights. We propose a simple approach based on modeling the distribution of the selected sample as a mixture, with the mixture weights appropriately adjusted, while accounting for uncertainty in the adjustment. We focus for simplicity on Dirichlet process mixtures but the proposed approach can be applied more broadly. We sketch a simple Markov chain Monte Carlo algorithm for computation, and assess the approach via simulations and an application.

1. Introduction

In sample surveys, it is routine to conduct stratified sampling designs to ensure that a broad variety of groups are adequately represented in the sample. In particular, the population is divided into mutually exclusive strata having different probabilities of inclusion. Data from stratified probability designs cannot be analyzed as if they are a random sample from the super-population without potentially large amounts of bias. To correct for discrepancies in the statistical analysis, survey weights are constructed. There is a rich literature on including adjustments for stratified sampling designs in estimation. However, the vast majority of such methods are not appropriate in model-based inferences, particularly under nonparametric Bayes frameworks.

Little (2004) and Gelman (2007) clarify the importance of including survey weights into model-based analyses. Zheng and Little (2003, 2005) propose a nonparametric spline model and Chen et al. (2010) extend the framework for binary variables. Although these approaches can flexibly connect the survey weights with the response, they rely on the assumption of survey weights being known for all population units. Zangeneh and Little (2012) propose a modification to allow the number of non-sampled units to be unknown. Si et al. (2015) instead propose a nonparametric model in which the survey weights are linked with a response through a Gaussian process regression. However, additional modeling of survey weights for non-sampled subjects in the population can lead to highly complex models.

* Corresponding author.
E-mail addresses: tsuyoshi.kunihama@uw.edu (T. Kunihama), aherring@bios.unc.edu (A.H. Herring), carolyn_halpern@unc.edu (C.T. Halpern), dunson@duke.edu (D.B. Dunson).

http://dx.doi.org/10.1016/j.spl.2016.02.009
0167-7152/

In this article, we propose a simple approach in which we apply standard mixture models, such as Dirichlet process mixtures, for the selected sample, and then adjust the mixture weights based on the survey weights. We allow probabilistic uncertainty in this adjustment in a Bayesian manner. Posterior computation relies on a simple modification to add an additional step to Markov chain Monte Carlo algorithms for mixture models.

2. Mixture models with survey weights

2.1. Adjusted density estimates

Let $y_1, \ldots, y_N$ denote independently and identically distributed observations from a superpopulation density f_0 with $y_i \in \mathcal{R}$ for $i \in D = \{1, \ldots, N\}$. From the finite population D, n subjects are sampled, with $w_i = c/\pi_i$ the survey weight for subject i, c a positive constant, and π_i the inclusion probability for $i \in D$. We assume D can be divided into mutually exclusive subpopulations $D_1, \ldots, D_M$, with $\{y_i,\ i \in D_m\}$ independently and identically distributed from density f_m, for $m = 1, \ldots, M$. Then, f_0 can be expressed as

$$f_0(y) = \sum_{m=1}^{M} \nu_m f_m(y), \tag{1}$$

where $\nu_m \geq 0$ and $\sum_{m=1}^{M} \nu_m = 1$. By applying kernel density estimation to each f_m in (1), Buskirk (1998) and Bellhouse and Stafford (1999) propose an adjusted density estimate,

$$\hat{f}_0(y) = \sum_{i \in S} \frac{\tilde{w}_i}{b} \mathcal{K}\left(\frac{y - y_i}{b}\right), \tag{2}$$

where $S \subset D$ are the selected subjects in the survey, $\tilde{w}_i = w_i / \sum_{j \in S} w_j$, $\mathcal{K}$ is a kernel function and $b > 0$. Estimator (2) adjusts for bias in the usual kernel estimator applied to sample S by modifying the weight for the ith subject from $1/n$ to $\tilde{w}_i$. This adjustment leads to consistency under some conditions (Buskirk and Lohr, 2005).

2.2. Bayesian adjustments with uncertainty

Section 2.1 focuses on univariate continuous variables, while our goal is to develop a general approach for adjusting posterior distributions to take into account sample survey weights. Let $y \in \mathcal{Y}$ denote a random variable, with $\mathcal{Y}$ a Polish space that may correspond to a p-dimensional Euclidean space, a discrete space, a mixed continuous and discrete space, a non-Euclidean Riemannian manifold, such as a sphere, and other cases. Extending (1) to general spaces, we let $f_0(\cdot)$ and $f_m(\cdot)$, for $m = 1, \ldots, M$, denote densities on $\mathcal{Y}$ with respect to a dominating measure μ. The density in the mth subpopulation is expressed as a mixture,

$$f_m(y) = \sum_{h=1}^{H} \nu_{mh} f(y \mid \theta_h), \tag{3}$$

where $\nu_{mh} \geq 0$, $\sum_{h=1}^{H} \nu_{mh} = 1$ and θ_h are parameters characterizing the hth mixture component. Then, f_0 can be approximately expressed as a mixture having the same kernels as in (3) but with adjusted weights as in (2).

Theorem 1. *Let $s_i \in \{1, \ldots, H\}$ denote the mixture index for subject i for $i \in S$. Let $S_h = \{i : s_i = h, i \in S\}$, for $h = 1, \ldots, H$. Then, for large N and n,*

$$f_0(y) \approx \sum_{h=1}^{H} \frac{\sum_{i \in S_h} w_i/c}{N} f(y \mid \theta_h) \approx \sum_{i \in S} \tilde{w}_i f(y \mid \theta_{s_i}). \tag{4}$$

Proof. Letting N_m be the number of subjects in D_m, $N_m/N \to \nu_m$ as $N \to \infty$ by the law of large numbers. Letting w_m^* and π_m^* denote the survey weight and inclusion probability for the mth subpopulation, $w_i = w_m^*$ and $\pi_i = \pi_m^*$ for $i \in D_m$. From (1) and (3), f_0 can be expressed as

$$\begin{aligned} f_0(y) &= \sum_{m=1}^{M} \nu_m f_m(y) \approx \sum_{m=1}^{M} \frac{N_m}{N} f_m(y) = \sum_{h=1}^{H} \sum_{m=1}^{M} \frac{N_m \nu_{mh}}{N} f(y \mid \theta_h) \\ &\approx \sum_{h=1}^{H} \frac{\sum_{i \in S_h} w_i/c}{N} f(y \mid \theta_h) \approx \sum_{i \in S} \tilde{w}_i f(y \mid \theta_{s_i}). \end{aligned} \tag{5}$$