Journal of Multivariate Analysis 119 (2013) 16–31

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

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

Fisher information in different types of perfect and imperfect ranked set samples from finite mixture models

Armin Hatefi, Mohammad Jafari Jozani *

Department of Statistics, University of Manitoba, Winnipeg, Manitoba, Canada, R3T 2N2

ARTICLE INFO

Article history:
Received 10 August 2012
Available online 12 April 2013

AMS subject classifications:
62D05
62B10
62G30
62F07

Keywords:
Complete data likelihood
Fisher information matrix
Finite mixture models
Imperfect ranked set sampling
Missing information principle
Perfect ranked set sampling

ABSTRACT

We derive some general results on the Fisher information (FI) contained in the data obtained from the ranked set sampling (RSS) design relative to its counterpart under the simple random sampling (SRS) for a finite mixture model. We propose different variations of RSS data and show how to calculate the FI matrix for each variation under both perfect and imperfect ranking assumptions. Also, a comparison is made among the proposed variations of RSS data using the missing information criterion. We discuss some interesting cases where the ratio of the determinant of the FI matrices for the RSS and SRS data is independent of the component densities and the number of components of the model and it is always equal to the set size used through the RSS procedure. Theoretical results are augmented by numerical studies for a mixture of two exponential distributions.

1. Introduction

The Fisher information (FI) matrix is a measure to quantify the amount of information that observations from the statistical experiment designed to investigate a parametric model carry about the unknown parameters of interest. The FI is also useful, for example, to obtain a matrix lower bound for the covariance matrix of unbiased estimators of the vector of unknown parameters $\boldsymbol{\Psi}$, or to study the asymptotic properties of the maximum likelihood estimators of $\boldsymbol{\Psi}$. In this paper, we obtain the Fisher information matrices of different types of ranked set samples from finite mixture models and compare them with their counterparts under simple random samples for both perfect and imperfect ranking situations. Ranked set sampling (RSS) is a sampling procedure which can be used in situations where a small number of sampling units can be ordered fairly accurately with respect to a variable of interest without actual measurements on them and this can also be achieved at low cost. This means some form of ranking of units is possible. This is a useful property since quite often, exact measurements of these units can be very tedious and/or expensive. For example, for fisheries studies (such as age structure or the length distribution of an age or sex class of fish, especially in the case of short-lived species) or environmental risks such as radiation (soil contamination and disease clusters) or pollution (water contamination and root disease of crops), exact measurements would require substantial scientific processing of materials or sampling units and a high cost as a result, while the variable of interest from a small number of experimental (sampling) units may easily be ranked. The RSS, as proposed by McIntyre [8] in estimating the mean of the pasture yields, is based on this premise and it provides an interesting alternative to simple random sampling in these situations. The process of RSS is considered by many to be a more efficient method of

* Corresponding author.
E-mail address: m_jafari_jozani@umanitoba.ca (M. Jafari Jozani).

0047-259X/$ – see front matter
http://dx.doi.org/10.1016/j.jmva.2013.03.018

data collection. In addition, in some ways, the sample so collected may be considered more representative of the population. To evaluate the efficiency of ranked set sampling, it is commonly compared to simple random sampling.

Suppose X is a random variable distributed according to a finite mixture of M component densities with the probability density function (pdf)

$$f(x;\boldsymbol{\Psi}) = \pi_1 f_1(x;\theta_1) + \cdots + \pi_M f_M(x;\theta_M), \tag{1}$$

where $\pi_j > 0$ are mixing proportions with $\sum_{j=1}^{M}\pi_j = 1$, and $f_j(\cdot;\theta_j)$ is the pdf of the j-th component of the mixture which is specified up to a vector θ_j of unknown parameters, known a priori to be distinct. The vector of all unknown parameters is denoted by $\boldsymbol{\Psi} = (\boldsymbol{\pi}, \boldsymbol{\xi})^t$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{M-1})$, $\boldsymbol{\xi} = (\theta_1^t, \ldots, \theta_M^t)^t$ and the superscript t refers to the vector transpose. For more information about the theory and application of finite mixture model we refer to McLachlan and Peel [9]. Let $f^{(r)}(x;\boldsymbol{\Psi})$ denote the pdf of the r-th order statistic $X_{r:k}$ of a simple random sample of size k from (1), where

$$f^{(r)}(x;\boldsymbol{\Psi}) = k\binom{k-1}{r-1} f(x;\boldsymbol{\Psi})\,[F(x;\boldsymbol{\Psi})]^{r-1}\,[\bar{F}(x;\boldsymbol{\Psi})]^{k-r}, \tag{2}$$

$\bar{F}(x,\boldsymbol{\Psi}) = 1-F(x,\boldsymbol{\Psi})$, and $X_{1:k} \le \cdots \le X_{k:k}$. To obtain a ranked set sample of size mk, an initial simple random sample of size k is taken (this is called a set of size k). These units are ordered, but without actually being measured; we call this judgement ranking, which may be perfect or imperfect. Upon ranking, only the smallest unit is measured. Following this, a second SRS of size k is taken, ranked and the second smallest is measured. This process is repeated until k units have been measured in which case we would have k independent order statistics $\{X_{(r)1}, r = 1, \ldots, k\}$. The whole process can be repeated m times to obtain a ranked set sample of size $n = mk$ denoted by $\mathbf{X}_{\mathrm{RSS}} = \{X_{(r)i},\ r = 1, \ldots, k;\ i = 1, \ldots, m\}$. Although RSS requires identification of mk^2 units from the underlying population but only mk of them are actually measured. The measurements $X_{(1)i}, \ldots, X_{(k)i}$ are independent order statistics (obtained from independent sets) and each of them provides information about different aspects of the distribution. Note that the essence of RSS is conceptually similar to the stratified sampling technique. In particular, RSS technique uses inherent heterogeneity among the sampling units through a ranking process to create artificial strata and so it can be considered as a stratification of the units during the sampling process based on their ranks in the set.

Recently, Hatefi et al. [4] have considered the problem of the maximum likelihood estimation of the vector of unknown parameters $\boldsymbol{\Psi}$ for the finite mixture model (1) using RSS data. They developed suitable EM algorithms to calculate the maximum likelihood estimators of $\boldsymbol{\Psi}$ and showed that RSS based estimators of $\boldsymbol{\Psi}$ are more efficient than their SRS counterparts. They also explained the superiority of RSS estimators over their SRS counterparts using the structural differences between SRS and RSS. In simple random sampling (SRS), observations are independent and identically distributed and each of them represents a typical value from the underlying distribution. However, there is no additional structure imposed on their relationship to one another. But in RSS, additional information and structure have been provided through the ranking process. Indeed, it is this extra information provided by the ranking and the independence of the resulting order statistics that makes procedures based on RSS to be more efficient than their counterparts based on SRS data with the same number of measurements. So, it is natural to quantify the amount of information that RSS data from the mixture model (1) carries about $\boldsymbol{\Psi}$ and compare it with that of SRS data. Chen [2] and Barabesi and El-Sharaawi [1] studied the FI contained in the usual RSS data for multi parameter family of distributions and showed that it is always larger than the FI of the SRS data.

In this paper, we consider three types of ranked set samples from the finite mixture model (1). We calculate the FI contained in each type and compare them with their SRS counterparts. The case where RSS data are obtained from the whole model and no more information about the component of origin of each observation is available is referred to as Type-M0 or uncategorized RSS data. For Type-M1 or fully categorized RSS data we assume that the component of origin of each observation $x_{(r)i}$ is known. In addition, we assume that in the set where $x_{(r)i}$ is observed from, the number of observations smaller than (larger than) $x_{(r)i}$ used from the j-th component of the mixture in the ranking process is known. Finally, we consider Type-M2 or partially categorized RSS data where each $x_{(r)i}$ is known to be obtained from a set consisting of units selected from only one of the components of the mixture model. Note that for Type-M2, the data contains no more information than its SRS counterpart about the mixing proportions $\boldsymbol{\pi}$. Examples of these types of data under simple random sampling can be found in [6]. Also, since there is no ranking involved in the SRS scheme, the SRS counterparts of both Type-M1 and Type-M2 RSS data will be the same (see Remark 1). Adapting the notation of Titterington et al. [11], we refer to the SRS counterpart of Type-M1 and Type-M2 RSS data as Type-C SRS data.

The outline of the paper is as follows. Section 2 develops the FI matrices for three types of RSS data under the perfect ranking assumption. We show that the FI contained in each variation of the RSS data is larger than the FI contained in its SRS counterpart. The results for Type-M0 RSS relative to its SRS counterpart closely resemble the results of Chen [2], but for the finite mixture models, and they are reproduced here for the sake of completeness. We obtain some interesting results concerning the FI of RSS data about the mixing proportions relative to its corresponding one with SRS data. In particular, we find cases where the ratio of the determinant of the RSS and SRS Fisher information matrices neither depends on the component densities nor to the number of components of (1) and it is always equal to the set size k. In Section 3, we pursue with the FI matrices of the imperfect ranked set samples and explore the effect of ranking error on the amount of information contained in each RSS data type. In Section 4, we make a comparison among different types of perfect RSS using the missing information criterion. Section 5 is devoted to several numerical studies using a mixture of two exponential distributions. Finally, in the Appendix, we present some of the proofs.