Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 170 (2011) 1–21

FUZZY sets and systems

www.elsevier.com/locate/fss

Parameter estimation from small biased samples: Fuzzy sets vs statistics

Leonid I. Piterbarg

Department of Mathematics, University of Southern California, Kaprielian Hall, Room 108, 3620 Vermont Avenue, Los Angeles, CA 90089-2532, United States

Received 4 May 2010; received in revised form 20 January 2011; accepted 24 January 2011
Available online 1 February 2011

Abstract

We consider a problem of estimating an unknown location parameter from several small biased samples. The biases and scale parameters of the samples are not known as well. For the case of two samples a fuzzy estimator based on a triangular membership function is introduced and studied. In particular, it is shown that its asymptotic bias is less than that of the weighted mean for the majority of key parameters in the problem. For small samples the fuzzy estimator is compared with the weighted mean and weighted median for a bunch of distributions. The main conclusion is that the fuzzy estimator performs better in most of the scenarii, however its advantage is subtle except for a few cases. Similar conclusions are obtained for the case of three information sources. The theoretical and simulation results for two samples might serve as a guidance for choosing a particular estimation method from the discussed ones based on preliminary information on relations between unknown parameters.

Keywords: Small samples; Biasness; Fuzzy estimate; Location parameter estimation

1. Introduction

This work was inspired by data assimilation problems arising in the physical oceanography and meteorology, e.g. [12]. A particular formulation discussed here is as follows. Let us suppose that information about an unknown scalar or vector parameter θ (for example temperature or velocity at fixed point in the ocean) comes from several different sources such as in situ observations, a circulation model, historical data, etc. As a result, one has several samples which typically are small and biased due to high cost and uncertainties in modeling and observing the ocean. Moreover, in many situations the biases cannot be properly evaluated and are assumed to be unknown as well. Thus, a new class of problems arises where the parameter of interest θ is simply not identifiable from the classical statistical viewpoint [4]. In such a case one of the reasonable ways to make progress is to assume that the biases are not too large or, say, have opposite signs, and then apply standard statistical tools ignoring the biases and hoping that the estimate still gives a right idea about the true parameter value. An alternative approach we suggest here is based on ideas of the fuzzy sets (possibility theory) e.g. [7,14], where the biasness and sample size are not critical issues. The main goal of this work is to compare these two approaches by considering a simple formulation.

E-mail address: piter@usc.edu.

0165-0114/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.fss.2011.01.010

Namely, we mostly focus on the problem of estimating an unknown scalar parameter θ from two different independent samples of size n_1 and n_2 respectively

$$x_{1i} = \theta + b_1 + \sigma_1 \xi_i, \; i = 1, 2, \ldots, n_1, \quad x_{2i} = \theta + b_2 + \sigma_2 \eta_i, \; i = 1, 2, \ldots, n_2 \tag{1}$$

where b_j is the bias of the j-th sample, σ_j the scale parameter, $j = 1, 2$, and ξ_i, η_i are random noises centered at zero. Both b_j and σ_j are assumed to be unknown but fixed. They are not of immediate interest and for this reason are considered as nuisance parameters. Thus, one can think about two devices measuring the same quantity with different unknown precisions.

Certainly, properties of any reasonable estimator for the parameter of interest θ would depend on values of the nuisance parameters as well as on the noise distribution. Here one of the goals is to compare that dependence for a suggested fuzzy estimator with traditional statistics such as weighted mean and weighted median. Conditions imposed on the noises will be specified below.

A similar problem with more than two samples is addressed in this paper as well, but to this end the reported progress is quite limited comparing to the case of two samples.

In the above set up one should forget about constructing unbiased or consistent estimators. Now this paper purpose is much more modest: assuming that the biases do not dominate the signal (say $|b_1|, |b_2| \leq |\theta|/2$), compare different sensible estimators in terms of both, the bias and efficiency. Presumably, the whole formulation may look a little bit strange or even awkward for main stream statisticians, but this is exactly the problem assimilation community people encounter regularly: sparse biased observations and a biased model output. Henceforth a reader can think about the first sample as observations and about the second sample as a model output. The problem is to combine them in an optimal way to reach an appropriate "assimilated" value of the parameter of interest.

If ξ_i, η_i are independent standard normal random variables, $b_1 = b_2 = 0$ and σ_i are known, then the Maximum Likelihood estimator of θ is given by e.g. [4]

$$\widehat{\theta} = \frac{\bar{x}_1 n_1/\sigma_1^2 + \bar{x}_2 n_2/\sigma_2^2}{n_1/\sigma_1^2 + n_2/\sigma_2^2} \tag{2}$$

where $\bar{x}_j$ is the sample mean of the j-th sample, $j = 1, 2$.

Apparently, if the biases and scales in model (1) are unknown, then there is nothing better as to assume that the biases are zeros and replace variances by their estimates. Thus, one come up with the weighted mean (WM)

$$\widehat{\theta} = \frac{\bar{x}_1 n_1/s_1^2 + \bar{x}_2 n_2/s_2^2}{n_1/s_1^2 + n_2/s_2^2} \tag{3}$$

where

$$s_j^2 = \frac{1}{n_j - 1} \sum (x_{ji} - \bar{x}_j)^2$$

is the sample variance.

The estimator (3) is a ML estimator only when $\sigma_1 = \sigma_2$ [4]. For different σ's any explicit expression for the ML estimator is not feasible and for this reason it is not practical.

Formula (3) is widely used in the physical oceanography and meteorology for combining data and model output [9]. Unfortunately when it is applied in reality, the principal assumptions under its derivation turn out to be forgotten time to time. That can lead to a nonsense like the following. Velocity vectors of opposite direction with whatever large but close magnitudes, coming from a model and observations, result in the assimilated velocity which turns out to be close to zero. Such a lapse happens because of ignoring the assumption of unbiasness. In such a case it is much more reasonable and honest to claim the model output and observations are incompatible and look for errors in the model or data. This is precisely what the fuzzy sets methodology prescribes to do.

Even for unbiased samples, the estimator (3) is not good for noises with heavy tails like Cauchy noise. In that case a more appropriate estimator is a weighted median (WMED)

$$\widehat{\theta}_{med} = \frac{\mu_1 n_1/s_1^2 + \mu_2 n_2/s_2^2}{n_1/s_1^2 + n_2/s_2^2} \tag{4}$$