Journal of Statistical Planning and Inference 139 (2009) 2147 – 2161

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Asymptotic properties of the MAMSE adaptive likelihood weights

Jean-François Plante*

University of Toronto, Canada

ARTICLE INFO

Article history:
Received 18 January 2008
Received in revised form
3 October 2008
Accepted 8 October 2008
Available online 19 October 2008

MSC:
primary 62F12, 62G20
secondary 62F10, 62G30

Keywords:
Statistical inference
Weighted likelihood
Maximum weighted likelihood estimate
Nonparametrics
Asymptotics
Strong consistency
Borrowing strength
Bootstrap

ABSTRACT

The weighted likelihood is a generalization of the likelihood designed to borrow strength from similar populations while making minimal assumptions. If the weights are properly chosen, the maximum weighted likelihood estimate may perform better than the maximum likelihood estimate (MLE). In a previous article, the minimum averaged mean squared error (MAMSE) weights are proposed and simulations show that they allow to outperform the MLE in many cases. In this paper, we study the asymptotic properties of the MAMSE weights. In particular, we prove that the MAMSE-weighted mixture of empirical distribution functions converges uniformly to the target distribution and that the maximum weighted likelihood estimate is strongly consistent. A short simulation illustrates the use of bootstrap in this context.

1. Introduction

The weighted likelihood is a frequentist method that allows to borrow strength from datasets that do not follow the target distribution exactly. This work is a sequel to that of Hu (1994), later published as Hu and Zidek (2002), who designed the weighted likelihood in order to take advantage of the relevant information contained in such samples. In the formulation of the weighted likelihood, an exponential weight discounts the contribution of each datum based on the discrepancy of its distribution with that of the target population.

In the context of dependent data (e.g. smoothing), Hu et al. (2000) use covariates to determine likelihood weights, but not the response variables themselves. In a different setting where the distribution of data stabilizes through time, Hu and Rosenberg (2000) use weights that are determined by a function whose parameter is set by minimizing the mean squared error (MSE) of the resulting estimate.

Although the initial paradigm of the weighted likelihood allows each datum to come from a different population, we rather adopt the same framework as Wang (2001), Wang and Zidek (2005) and Wang et al. (2004) where data come as samples from m populations. In this context, one could hope to set the weights based on scientific information, but it is more pragmatic and less arbitrary to determine them based on the data.

Under this paradigm, neither an ad hoc method suggested by Hu and Zidek (2002) nor the cross-validation method explored by Wang and Zidek (2005) provide a satisfactory recipe for finding likelihood weights. The cross-validation weights, for instance, lack

* Tel.: +1 416 978 3452; fax: +1 416 978 5133.
E-mail address: plante@utstat.toronto.edu.

0378-3758/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2008.10.001

numerical stability. Recently, Plante (2008) suggested nonparametric adaptive weights whose formulation is based on heuristics showing that the weighted likelihood is a special case of the entropy maximization principle. Simulations show that the so-called MAMSE (minimum averaged mean squared error) weights allow to outperform the likelihood under many scenarios.

Competing methods that borrow strength from a fixed number of samples typically rely on a hierarchical model. By opposition, the MAMSE-weighted likelihood does not require to model the extra populations and hence cannot be negatively affected by model misspecification on a population of secondary interest. In situations where no hierarchical model arises naturally, this may constitute a major advantage.

The asymptotic properties of the weighted likelihood are studied by Hu (1997) for weights that do not depend on the data. Asymptotics for adaptive weights are developed by Wang et al. (2004) under the assumption that the weights asymptotically shift towards Population 1 at a certain rate. As Plante (2008) points out, the MAMSE weights do not follow this behavior and hence require a special treatment.

In this paper, we study the asymptotic properties of the MAMSE weights, the MAMSE-weighted mixture of empirical distribution functions (EDFs) and of the corresponding maximum weighted likelihood estimate (MWLE). In Section 2, we introduce the weighted likelihood and the MAMSE weights formally. A sequence of lemmas is presented in Section 3 to show that a MAMSE-weighted mixture of EDFs converges uniformly to the target distribution. In Section 4, we prove that the MWLE is a strongly consistent estimate by generalizing the proof of Wald (1949) for the likelihood. Section 5 discusses the asymptotic behavior of the MAMSE weights themselves. The use of bootstrap methods is illustrated through simulations in Section 6. The MAMSE-weighted MWLE offers better performances than the maximum likelihood estimate (MLE) in many cases, yielding good coverage for shorter bootstrap confidence intervals (CIs).

2. The weighted likelihood and the MAMSE weights

We introduce a notation that allows for increasing sample sizes as it will be useful for the remaining of this manuscript. Let $(\Omega, \mathscr{B}(\Omega), P)$ be the sample space on which the random variables

$$X_{ij}(\omega) : \Omega \to \mathbb{R}, \quad i = 1, \ldots, m, \ j \in \mathbb{N}$$

are defined. The X_{ij} are assumed to be independent with continuous distribution F_i.

We consider samples of nondecreasing sizes: for any positive integer k, the random variables $\{X_{ij} : i = 1, \ldots, m, \ j = 1, \ldots, n_{ik}\}$ are observed. Moreover, the sequences of sample sizes are such that $n_{1k} \to \infty$ as $k \to \infty$. We do not require that the sample sizes of the other populations tend to ∞, nor do we restrict the rate at which they increase.

Suppose that Population 1 is of inferential interest. If we denote by $f(x \mid \theta)$ the family of distributions used to model Population 1, the weighted likelihood and the weighted log-likelihood are written as

$$L(\theta) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} f(X_{ij}|\theta)^{\lambda_i/n_i} \quad \text{and} \quad \ell(\theta) = \sum_{i=1}^{m} \frac{\lambda_i}{n_i} \sum_{j=1}^{n_i} \log f(X_{ij}|\theta)$$

where the $\lambda_i \geqslant 0$ are likelihood weights such that $\sum_{i=1}^{m} \lambda_i = 1$.

Let $\hat{F}_{ik}(x) = (1/n_{ik}) \sum_{j=1}^{n_{ik}} \mathbb{1}(X_{ij} \leqslant x)$ be the EDF based on the sample $X_{ij}, j = 1, \ldots, n_{ik}$. The empirical measure associated with $\hat{F}_{ik}(x)$ allocates a weight $1/n_{ik}$ to each of the observations $X_{ij}, j = 1, \ldots, n_{ik}$.

Plante (2008) shows heuristically that maximizing the weighted likelihood is comparable to maximizing the proximity between the model $f(x|\theta)$ and a mixture of the m EDFs obtained from the samples at hand. Such a mixture was considered before by Hu and Zidek (1993, 2002) and Hu (1994) who called it relevance weighted empirical distribution function (REWED). By comparison, the usual likelihood is akin to maximizing the entropy between $f(x|\theta)$ and $\hat{F}_{1k}(x)$.

Inspired by the heuristics briefly described above, Plante (2008) tries to find weights that make the mixture of EDFs $\sum_{i=1}^{m} \lambda_i \hat{F}_{ik}(x)$ close to $\hat{F}_{1k}(x)$, but less variable. He proposes the MAMSE objective function.

Some preprocessing steps first discard any sample whose range of values does not overlap with that of Population 1. For the remaining m samples, we write $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^{\mathsf{T}}$ and minimize

$$P_k(\boldsymbol{\lambda}) = \int \left[\left\{ \hat{F}_{1k}(x) - \sum_{i=1}^{m} \lambda_i \hat{F}_{ik}(x) \right\}^2 + \sum_{i=1}^{m} \lambda_i^2 \widehat{\mathrm{var}}\{\hat{F}_i(x)\} \right] \mathrm{d}\hat{F}_{1k}(x) \tag{1}$$

as a function of $\boldsymbol{\lambda}$ under the constraints $\lambda_i \geqslant 0$ and $\sum_{i=1}^{m} \lambda_i = 1$. We proceed to the substitution

$$\widehat{\mathrm{var}}\{\hat{F}_i(x)\} = \frac{1}{n_{ik}} \hat{F}_{ik}(x)\{1 - \hat{F}_{ik}(x)\}$$

in Eq. (1) based on the variance of the Binomial variable $n_{ik}\hat{F}_i(x)$ for fixed x. The choice of $\mathrm{d}\hat{F}_{1k}(x)$ allows to integrate where the target distribution $F_1(x)$ has most of its mass.