Journal of Multivariate Analysis 158 (2017) 103–116

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Nonparametric tests for multi-parameter M-estimators

John E. Kolassa [a,*], John Robinson [b]

[a] Department of Statistics and Biostatistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854, USA

[b] School of Mathematics and Statistics, University of Sydney, Carslaw Building F07, Eastern Avenue, Camperdown NSW 2006, Australia

ARTICLE INFO

Article history:
Received 26 November 2016
Available online 25 April 2017

AMS subject classifications:
62G09
62G10
62G20

Keywords:
Empirical saddlepoint
Tilted bootstrap
Regression
Non-linear regression
Generalized linear models

ABSTRACT

We consider likelihood ratio like test statistics based on M-estimators for multi-parameter hypotheses for some commonly used parametric models where the assumptions on which the standard test statistics are based are not justified. The nonparametric test statistics are based on empirical exponential families and permit us to give bootstrap methods for the tests. We further consider saddlepoint approximations to the tail probabilities used in these tests. This generalizes earlier work of Robinson et al. (2003) in two ways. First, we generalize from bootstraps based on resampling vectors of both response and explanatory variables to include bootstrapping residuals for fixed explanatory variables, resulting in a surprising result for the weighted resampling. Second, we obtain a theorem for tail probabilities under weak conditions providing essential justification for the approximation to bootstrap results for both cases. We use as examples linear regression, non-linear regression and generalized linear models under models with independent and identically distributed residuals or vectors of observations, giving numerical illustrations of the results.

1. Introduction

Let $Y_1(\theta), \ldots, Y_n(\theta)$ be a sample of independent and identically distributed random vectors, with $Y_j(\theta)$ from a distribution F on the sample space $\mathcal{Y}$. Suppose that θ satisfies

$$\mathrm{E}\Big[\sum_{j=1}^{n} \psi_j\{Y_j(\theta), \theta\}\Big] = 0 \tag{1}$$

and consider test statistics based on T, the M-estimate of θ, defined by the solution of

$$\sum_{j=1}^{n} \psi_j\{Y_j(\theta), T\} = 0, \tag{2}$$

where ψ_j are assumed to be smooth functions from $\mathcal{Y} \times \mathbb{R}^p$ to $\mathbb{R}^p$. The functions ψ_j are often chosen to make an analysis more robust.

We have, in particular, two cases in mind, where, for example, in linear regression with response variables Z_j and explanatory variables X_j, $Y_j(\theta) = (Z_j, X_j^\top)^\top$ and

$$\psi_j\{Y_j(\theta), t\} = X_j(Z_j - t^\top X_j),$$

* Corresponding author.
E-mail addresses: kolassa@stat.rutgers.edu (J.E. Kolassa), john.robinson@sydney.edu.au (J. Robinson).

http://dx.doi.org/10.1016/j.jmva.2017.04.004
0047-259X/

or $Y_j(\theta) = Z_j - \theta^\top x_j$ and

$$\psi_j\{Y_j(\theta), t\} = x_j\{Y_j(\theta) + (\theta - t)^\top x_j\} = x_j(Z_j - t^\top x_j),$$

for fixed $X_j = x_j$. Note that it is $Y_j(\theta)$ that are identically distributed allowing resampling.

Let $\theta = (\theta_1^\top, \theta_2^\top)^\top$, where $\theta_1 \in \mathbb{R}^{p_1}$ and $\theta_2 \in \mathbb{R}^{p_2}$, $p_1 + p_2 = p$, and suppose we wish to test the null hypothesis

$$\mathcal{H}_0 : \theta_2 = \theta_{20}.$$

If the common distribution of $Y_j(\theta)$ belongs to some parametric model, then F belongs to a class of distributions such that (1) holds with $\theta_2 = \theta_{20}$, and standard likelihood theory for estimation and inference is available. However, when the sample size is moderate to small or when the model is incorrectly specified, the p-values obtained from the asymptotic theory can be very inaccurate. [10] proposed a new likelihood like statistic based on an empirical exponentially tilted distribution considering only the case $\psi_j = \psi$. Assuming that the density of $\sum_{j=1}^n \psi\{Y_j(\theta), \theta\}$ exists, they gave a saddlepoint approximation with *relative* error of order $O(n^{-1})$. This method can only be used when F is known. Further, they considered a formal approach to empirical likelihood ratio tests using bootstrap tilting. The saddlepoint approximation to the distribution of the bootstrap statistic requires a proof of the result without the restrictive condition that a density exists. This proof, given in Section 6, is of an entirely different character from that of [10].

The two purposes of this paper are to justify the formal approach for saddlepoint approximations of [10] for empirical likelihood tests and to consider score functions ψ_j which change with each observation. We note that [4] obtained tests in the case of one-dimensional parameters for identically distributed score functions but their methods could not be extended to the case of multi-dimensional parameters. In Section 2, a test statistic related to that from exponential families is derived from the cumulant generating function of the left hand side of the estimating Eq. (2) when the distribution of $Y_j(\theta)$ is known under the null hypothesis. If the distribution is not known, a tilted empirical distribution satisfying the null hypothesis is obtained as an approximation and its cumulant generating function is used to obtain a natural test statistic. We use weighted bootstrap sampling from this tilted empirical distribution to obtain p-values for the test. The theorem of Section 3 gives a saddlepoint approximation of this bootstrap p-value and could be used instead of resampling. Bootstrap sampling requires a double optimization for each bootstrap replicate and so is extremely computationally intensive, so the saddlepoint approximation may be useful as an alternative. Note that the nonparametric approach depends only on $\psi_j\{Y_j(\theta), t\}$ for all $j \in \{1, \dots, n\}$. These functions may have been derived from some parametric model, but this model is not used except to give these estimating functions. In Section 4 we provide applications to three special cases, linear regression, robust non linear regression and robust generalized linear models. In Section 5 we give numerical results to illustrate the accuracy of the approximations for some important cases of tests and compare the power of the tests to the power of the standard tests in two cases.

2. A nonparametric test

First consider the simpler case in which the distribution F of $Y_j(\theta)$ is known. Denote the cumulant generating function of $\sum_{j=1}^n \psi_j\{Y_j(\theta), t\}$ by

$$nK(\tau, t) = \sum_{j=1}^n K_j(\tau, t) = \sum_{j=1}^n \ln\left\{\mathrm{E}\left(\exp[\tau^\top \psi_j\{Y_j(\theta), t\}]\right)\right\}. \tag{3}$$

Let $T = (T_1^\top, T_2^\top)^\top$ be the M-estimator, the solution to

$$\sum_{j=1}^n \psi_j\{Y_j(\theta), T\} = 0.$$

Consider a test statistic based on the function h defined by

$$h(t_2) = \inf_{t_1} \sup_{\tau}\{-K(\tau, t)\} = -K[\tau\{t(t_2)\}, t(t_2)], \tag{4}$$

where $t(t_2) = (t_1(t_2)^\top, t_2^\top)^\top$ for

$$\tau(t) = \arg\sup_{\tau}\{-K(\tau, t)\} \quad \text{and} \quad t_1(t_2) = \arg\inf_{t_1}[-K\{\tau(t), t\}].$$

Note that $h(\theta_{20}) = 0$. So a test can be based on $h(T_2)$. This is the statistic considered in [10]. In Section 3.2 of [7] it is shown that, in the case of generalized linear models with the classical score statistic when $t = t_2$, the test based on $h(t_2)$ reduces to the likelihood ratio statistic.

In practice, the distribution underlying the data sample $Y_1(\theta), \dots, Y_n(\theta)$ is often unknown, and hence K is unknown, and a nonparametric approach is required. An empirical exponential likelihood, equivalent to a tilted bootstrap, provides empirical versions of the test of $\mathcal{H}_0 : \theta_2 = \theta_{20}$. We consider weighted empirical distributions

$$\hat{F}(x) = \sum_{k=1}^n w_k \mathbf{1}\{Y_k(\theta) \le x\},$$