Computational Statistics and Data Analysis 55 (2011) 973–982

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Stepwise local influence analysis

Lei Shi [a,*], Mei Huang [b,c]

[a] Statistics and Mathematics School, Yunnan University of Finance and Economics, Kunming, 650221, PR China
[b] Mathematics and Statistics School, Yunnan University, Kunming, 650091, PR China
[c] Management School, Yunnan University of Nationalities, Kunming, 650031, PR China

ARTICLE INFO

Article history:
Received 29 July 2009
Received in revised form 7 June 2010
Accepted 2 August 2010
Available online 13 August 2010

Keywords:
Local influence analysis
Influential observations
Subset perturbation scheme
Masking effects

ABSTRACT

A new method called stepwise local influence analysis is proposed to detect influential observations and to identify masking effects in a dataset. Influential observations are detected step-by-step such that any highly influential observations identified in a previous step are removed from the perturbation in the next step. The process iterates until no further influential observations can be found. It is shown that this new method is very effective to identify the influential observations and has the power to uncover the masking effects. Additionally, the issues of constraints on perturbation vectors and bench-mark determination are discussed. Several examples with regression models and linear mixed models are illustrated for the proposed methodology.

1. Introduction

Local influence analysis introduced by Cook (1986) employs a simultaneous perturbation scheme and maximizes the local change of the normal curvature of an influence graph that is based on likelihood displacement to find a local diagnostic. This method is flexible since one can perturb different parts of models and study the local changes caused by these perturbations. Since then, many applications and extensions of this method have been studied (e.g. Beckman et al., 1987; Lawrance, 1988; Thomas and Cook, 1990; Wu and Luo, 1993; St. Laurent and Cook, 1993; Shi, 1997; Poon and Poon, 1999; Zhu and Lee, 2001; Zhu et al., 2007).

The identification of masking effects is an important issue and has been extensively discussed in the field of influence analysis (Cook and Weisberg, 1982; Atkinson, 1985; Rousseeuw and Leory, 1987; Chatterjee and Hadi, 1988). Lawrance (1995) provided a detailed classification for masking effects. The current approaches for dealing with the problem of masking effects include the use of multiple case deletions (Chatterjee and Hadi, 1988; Lawrance, 1995) or the use of robust detection (Atkinson, 1985, 1986; Rousseeuw and Leory, 1987). Bruce and Martin (1989) suggested an iterative leave-k-out diagnostic procedure to deal with the masking and smearing effects in time series data. Obviously deletion diagnostics need extensive computations.

Local influence analysis employs a simultaneous perturbation scheme and has an advantage in that it can detect joint influence of observations and thus can identify some masking effects (Lawrance, 1988). However, when there exist strong masking effects in the data, such as those caused by a group of outliers or influential observations, the local influence analysis will fail to identify such influential patterns, as noted by Bruce and Martin (1989, p. 420) in time series models and our studies in this paper. For this reason, some authors have used the first two eigenvectors of the key matrix in local diagnostics to extract more information about influential observations (Lu et al., 1997; Shi, 1997; Shi and Ojeda, 2004), or they have defined an aggregate measure that includes several influential eigenvectors (Poon and Poon, 1999; Zhu and Lee, 2001) to detect

* Corresponding author. Tel.: +86 871 5113808; fax: +86 871 5113808.
E-mail address: shi_lei65@hotmail.com (L. Shi).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2010.08.001

the influential observations. However these approaches suffer from the problem about how many eigenvectors should be inspected or included. In addition, when several eigenvectors are used to detect influential observations, it is difficult to compare the magnitudes of the influential observations identified from different eigenvectors.

In local influence analysis, although the observations are jointly detected, when one observation is highly influential, the associated local diagnostic for this point will have a high value, which will mask the effects of other influential observations if there is a mutual influence among influential observations. We propose here to uncover masking effects more generally by removing the effects of any highly influential observations identified by perturbing all observations in the first time step, and then performing a local influence analysis that is based on a subset perturbation scheme in which the influential observations identified in the first step are excluded. In this way, the influential observations masked by the first step can be identified in this new step. The procedure continues iteratively until no further influential observations can be found. In this paper we use this idea to suggest a new local influence method, which we call stepwise local influence analysis. We also discuss the issues of constraints on the perturbation vector and bench-mark determinations on local influence analysis. The analysis of three examples shows that this technique is very effective for identifying outliers or influential observations and uncovering masking effects.

The rest of the paper is organized as follows. Section 2 presents the local influence analysis method and discusses the constraints of the perturbation vector and the bench-mark determination problem of the local diagnostic. Section 3 proposes the stepwise local influence method. Section 4 examines the proposed technique using two examples in a regression model. Section 5 illustrates the method using an example in a linear mixed model. Section 6 gives some concluding remarks and discussions.

2. Local influence analysis and some notes

2.1. Local influence analysis

Let $\mathbf{y} = (y_1, \ldots, y_n)'$ denote an $n \times 1$ vector of observations with probability density function $p(\mathbf{y}|\theta)$, where θ is an $r \times 1$ unknown parameter vector. The log-likelihood function of the postulated models is denoted by $\ell(\theta)$. A perturbation scheme is introduced through a vector $\omega \in \Omega \in R^m$, where m is the dimension of ω and a commonly used case is $m = n$. Let $\ell(\theta|\omega)$ denote the perturbed log-likelihood function, and assume there is a ω_0 such that $\ell(\theta|\omega_0) = \ell(\theta)$. Following the Cook' methodology the likelihood displacement is defined as $LD(\omega) = 2[\ell(\hat{\theta}) - \ell(\hat{\theta}_\omega)]$, where $\hat{\theta}$ and $\hat{\theta}_\omega$ are, respectively, the MLE of θ in postulated model ($\omega = \omega_0$) and perturbed model. Let $\omega = \omega_0 + \epsilon h$, where ϵ is a small scalar and h is an unit-length vector in R^m. Cook (1986) used the normal curvature C_h of the influence graph $\alpha(\omega) = (\omega', LD(\omega))'$ to measure the local change caused by the perturbation, which has the following form

$$C_h = 2h'\Delta'(-\ddot{L})^{-1}\Delta h \tag{2.1}$$

where $\ddot{L} = \partial^2\ell(\theta)/\partial\theta\partial\theta'$ and $\Delta = \partial^2\ell(\theta|\omega)/\partial\theta\partial\omega'$ evaluated at $\omega = \omega_0$ and $\theta = \hat{\theta}$. The main diagnostic of local influence is obtained by maximizing C_h with respect to h under the condition that $h'h = 1$. It is known this maximum direction, denoted by $h_{\max}$, is the eigenvector associated with the largest absolute eigenvalue of matrix $\Delta'(-\ddot{L})^{-1}\Delta$. $h_{\max}$ is the main diagnostic in local influence analysis, and the maximum curvature $C_{\max} = C_{h_{\max}}$ can be used to check the magnitude of influence in this direction.

Poon and Poon (1999) noted that C_h is not invariant under the scale change of perturbation vector, and suggested the conformal normal curvature. At the critical point ω_0, the local diagnostic can be found by maximizing

$$C_h = h'Ah, \quad A = \frac{\Delta'(-\ddot{L})^{-1}\Delta}{\sqrt{\mathrm{tr}(\Delta'(-\ddot{L})^{-1}\Delta)^2}} \tag{2.2}$$

with respect to h with $h'h = 1$. Let (λ_i, α_i), $i = 1, \ldots, n$, denote the eigen-pair of matrix A, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, and $\sum_i \lambda_i^2 = 1$. Then $h_{\max} = \alpha_1$, which has nothing changed with that based on Cook's normal curvature, however $C_{\max} = \lambda_1$ in this case has a range that varies from 0 to 1, which is convenient for measuring the magnitude of influence in direction $h_{\max}$. It is noted that some authors have also used the first two eigenvectors to study the influential patterns (Lu et al., 1997; Shi, 1997; Shi and Ojeda, 2004).

2.2. Constraint on the perturbation vector

Critchley and Marriott (2004) noted that it is sensible to have a perturbation vector ω with some constraints. In case weights of likelihood function, the perturbed likelihood can be written as (Lu et al., 1997)

$$\ell(\theta|\omega) = \sum_i \omega_i \ell(\theta|y_i) = nE_{F_n(\omega)} l(\theta|y),$$