Available online at www.sciencedirect.com
ScienceDirect
Comput. Methods Appl. Mech. Engrg. 313 (2017) 632–646
Computer methods in applied mechanics and engineering
www.elsevier.com/locate/cma

Outlier detection and robust regression for correlated data

Ka-Veng Yuen*, Gilberto A. Ortiz

Faculty of Science and Technology, University of Macau, 999078, Macao, China

Received 6 June 2016; received in revised form 2 September 2016; accepted 3 October 2016
Available online 14 October 2016

Highlights

- The proposed probabilistic approach is applicable for correlated and uncorrelated data.
- The proposed method requires no information on the outlier distribution model.
- The proposed outlier probability is a holistic measure of the entire dataset.
- The methodology is demonstrated with simulated and real seismic data.

Abstract

Outlier detection has attracted considerable interest in various areas. Existing outlier detection methods usually assume independence of the modeling errors among the data points but this assumption does not hold in a number of applications. In this paper we propose a probabilistic method for outlier detection and robust updating of linear regression problems involving correlated data. First, suspicious data points will be identified using the minimum volume ellipsoid method and the maximum trimmed likelihood method. Then, the outlierness of each suspicious data point will be determined according to the proposed outlier probability in consideration of possible correlation among the data points. The proposed method is assessed and validated through simulated and real data.

Keywords: Bayesian inference; Correlated noise; Maximum likelihood; Model updating; Outlier detection

1. Introduction

Outlier detection has attracted considerable interest because misuse of outliers deteriorates substantially the outcome of model updating, leading to erroneous and drastically misleading predictions. The occurrence of outliers is due to model weaknesses [1], measurement errors (human errors or environmental disturbances) [2], or other mechanisms not taken into account by the underlying model [3]. A number of techniques have been developed to cope with outliers and applications can be found in various research areas, such as network-wide traffic [4], network intrusion detection [5], damage detection [6], surveillance systems [7], and structural health monitoring [8], etc.

* Corresponding author. Fax: +853 8822 2426.
E-mail address: kvyuen@umac.mo (K.-V. Yuen).

http://dx.doi.org/10.1016/j.cma.2016.10.004
0045-7825/

It has been shown that ordinary least squares-based outlier detection methods perform poorly in the presence of leverage points [2]. This motivates the development of robust regression techniques to minimize the adverse effects due to anomalous data. Examples include the M-estimator [9], L-estimator [10], trimmed elemental estimators [11], least median of squares [12], and least trimmed squares (LTS) [13], etc. Although robust regression methods have been shown effective in many applications, their success relies on the proper choice of some threshold parameters, which are usually assigned according to experience.

Bayesian model updating is a rigorous statistical framework which uses the Bayes' rule to update the probability of a hypothesis or the probability distribution of uncertain parameters. The application of Bayesian inference covers essentially all scientific disciplines, including subsurface flow modeling [14], structural dynamics [15–20], reliability analysis [21–23], damage detection [24–27], and geotechnical engineering [28–30], etc. In terms of outlier detection, Bayesian inference algorithms have been proposed by use of the posterior probability distribution of the error terms [31,32]. Some researchers utilized heavy-tailed distributions to model the outliers [33,34] while others assumed the outlier generation models [35,36] to take into account atypical data in the likelihood function. However, it is both conceptually and practically subtle to construct the outlier distribution model because outliers are data points with large error due to unknown/unmodeled mechanisms. If the distribution model is available, they will no longer be outliers. Recently, Yuen and Mu [37] proposed a probabilistic outlier detection approach to resolve these problems. This method incorporates the number of data points, the noise level and the posterior uncertainty for outlier detection and it does not require any information about the outlier characteristics, e.g., the amount and/or distribution of outliers. However, its assumption on independent fitting errors is invalid for some applications.

In this paper, from the foundation of the work in [37], we propose a probabilistic outlier detection and robust updating method for linear regression problems involving correlated data. First, we generalize the definition of outlier probability in [37] to account explicitly for the possible correlation among the data points. Second, this generalized definition is a holistic measure of all data points in contrast to the original definition in [37] that assesses each data point individually. Third, a robust estimator is utilized to estimate the standard deviation of the modeling errors. The structure of this paper is outlined as follows. Section 2 introduces the Bayesian framework for linear regression problems with correlated data. Then, Section 3 presents the proposed probabilistic method for outlier detection and robust identification to handle possibly correlated data. Section 4 summarizes the procedure of the proposed algorithm. Finally, Section 5 presents applications using simulated and real data to validate the effectiveness and robustness of the proposed method.

2. Bayesian inference for linear regression

Consider the linear regression problem:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{y} = [y_1, \ldots, y_N]^T \in \mathbb{R}^N$ is the output/response vector; $\mathbf{X} \in \mathbb{R}^{N \times N_\beta}$ is the design matrix containing the input/design variables; and $\boldsymbol{\beta} = \left[\beta_1, \ldots, \beta_{N_\beta}\right]^T \in \mathbb{R}^{N_\beta}$ is the unknown regression parameter vector to be identified. In most applications, the probabilistic model describing the residual vector $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_N]^T$ is unknown and uncorrelated Gaussian model is assumed due to its convenience for analysis. Even though this supposition is often reasonable, it is invalid for a number of applications. For instance, for a network of seismometers at different locations to record the ground motion induced by an earthquake, it is common practice in the seismology society to take into explicit account the correlation among the measurements. Otherwise, the uncertainty estimation of the regression parameters will be underestimated [38].

In the ideal situation with no outliers, the residual vector $\boldsymbol{\varepsilon}$ follows the N-variate Gaussian distribution of zero mean and covariance matrix $\sigma^2 \boldsymbol{\Gamma}(\boldsymbol{\rho})$ with unknown parameters σ and $\boldsymbol{\rho}$, i.e., $\boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{\Gamma}(\boldsymbol{\rho})\right)$, where $\boldsymbol{\Gamma}(\boldsymbol{\rho})$ represents the correlation structure of the prediction-error covariance matrix with unknown parameter vector $\boldsymbol{\rho}$. Given this model, the likelihood function is given by [39,40]:

$$p(\mathbf{y}|\boldsymbol{\beta}, \sigma, \boldsymbol{\rho}, \mathbf{X}) = \left(2\pi\sigma^2\right)^{-N/2} |\boldsymbol{\Gamma}(\boldsymbol{\rho})|^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Gamma}(\boldsymbol{\rho})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]. \tag{2}$$