Outlier Detection Using the Outlier Probability for Robust Linear Regression

Ka-Veng Yuen*
University of Macau, Macao, China kvyuen@umac.mo

He-Qing Mu
University of Macau, Macao, China maxathere@gmail.com

Abstract

Outlier detection and its treatment is one of the important problems in statistics and it has attracted tremendous attention over the last decades. In real measurements, outliers exist due to extraordinarily large measurement error and/or modeling error. Since outliers may lead to undesirable identification results, detection and special treatment of outliers becomes an important task in system identification. In this paper, we introduce a novel concept of outlier probability for outlier detection and robust linear regression. First, the Mahalanobis distance is utilized to identify the leverage points. By excluding the leverage points, the least trimmed squares method will be used to obtain an initial set of regular data points while the remaining (including the leverage points) are included in the initial suspicious data set. Then, each suspicious data point and their combinations are evaluated by a novel outlier probability that depends not only on residual but also on the size of the entire data set. Incorporating the data size is important as it controls the probability of the existence of a data point with a given value of the residual. This outlier probability is robust because it incorporates also the posterior uncertainty quantified using the Bayesian approach. Then, the data points with outlier probability below 0.5 will be transferred to the set of regular data points. This iteration is continued until all suspicious data points are associated with outlier probability over 0.5. Finally, linear regression will be conducted by considering only the set of regular points. Therefore, it is expected that the identification results are robust to outliers. In contrast to other existing outlier detection criteria that require some subjective bound (e.g., normalized residual larger than 2.5), the outlier probability threshold of 0.5 is intuitive. Finally, a challenging application will be presented and comparison with other well-known methods will be given.

Key Words: Bayesian inference, leverage, inverse problem, robust analysis, system identification