Outlier Detection via Feature Selection Algorithms in Covariance Estimation

Menjoge, Rajiv S.
M.I.T., Operations Research Center
77 Massachusetts Avenue, E40-130
Cambridge, MA 02139-4307, USA
E-mail: menjoge@alum.mit.edu

Welsch, Roy E.
M.I.T., Sloan School of Management and Engineering Systems
77 Massachusetts Avenue, E62-564
Cambridge, MA 02139-4307, USA
E-mail: rwelsch@mit.edu

1. Introduction

Robust covariance estimation is a central problem in robust statistics and has an extensive literature base (see for instance [1]). Finding an algorithm with desirable properties that works in every situation has proven to be very difficult, but several good algorithms that are useful in practice exist (for example [2], [3], [4], [5]).

In this paper, we present an alternative way of viewing the robust covariance estimation problem by posing it as a feature selection problem for a multivariate linear model. In particular, we set our data matrix to be Y, construct X to be a column of 1's with an identity matrix appended to the right, perform linear regression with feature selection in the multivariate linear model of Y onto X, and output as a robust covariance matrix the resulting conditional covariance matrix of the regression. This leads to the development of a class of algorithms which can be used to construct robust covariance matrices. We use backwards selection as a candidate feature selection algorithm and discuss the results.

The rest of this paper is organized as follows: The remainder of this section provides a brief literature review on the mean shift outlier model and related ideas. Section 2 develops the link between feature selection and outlier detection and describes our implementation of the backwards selection algorithm for multivariate regression. Section 3 describes results on the data sets we used, and section 4 concludes.

1.1 Literature Review

The mean-shift outlier model, which relates outliers to dummy features was originally developed to compute efficient diagnostics for linear regression. Several extensions have been made since then. For instance, [6] relates outliers to dummy features for the case of generalized linear models. Meanwhile, Morgenthaler et al [7] and McCann et al [8] used the relationship to establish a connection between outlier detection in linear regression and feature selection, and Kim et al [9] and McCann et al [10] applied this connection to developing new outlier detection techniques in linear regression. In particular, they computed a robust coefficient vector by appending an identity matrix to the design matrix and then performing linear regression with feature selection. In this paper, we establish this connection for the case of outlier detection in multivariate analysis, where we are interested in the estimation of a robust covariance matrix, rather than a robust coefficient vector. This connection makes it possible to use feature selection algorithms for multivariate linear models in order to detect and control for outliers in multivariate data.

2. Methodology

We pose the robust covariance estimation problem as a feature selection problem as follows: We first set our data matrix to be Y and construct X by creating one column of 1's, and then appending a set of dummy variable columns, corresponding to observations which are candidate outliers, to the right (in the case where we have no knowledge about which observations are outliers, we would append an entire identity matrix to the right). We then perform linear regression with feature selection, or dimension reduction, on the multivariate linear model of Y onto X, where Y is considered to be the matrix of realizations of the response variables, and X is the design matrix. Our robust estimate of the covariance matrix of the original data is then the covariance matrix Σ_ε of the error term for the regression.

The justification for our algorithm lies in the following two key relationships: 1. The estimated classical covariance matrix for multivariate data is the same as the estimated conditional covariance matrix of the multivariate linear model of Y onto X, where Y is the data matrix, and X is a column of 1's. 2. The mean-shift outlier model [11] establishes that performing a regression with deleted observations yields the same results (estimated coefficient vector and estimated conditional covariance matrix) as performing a regression onto an augmented X matrix, where the augmented columns are dummy columns corresponding to the observations deleted in the first model.

One could, in principal, use this methodology to apply any feature selection algorithm to estimate a robust covariance matrix, though in this paper, we focus on using backwards selection to demonstrate our methodology. In the next subsection, we provide the details of our implementation of backwards selection and, in the following subsection, we describe other potential extensions of our methodology, the implementation of which will be future work.

2.1 Backwards Selection

In our implementation of the generic backwards selection algorithm, we first scale the data ($n \times r$ Y matrix) by subtracting the column medians and dividing by the column mads and construct the $n \times (n+1)$ X matrix with the entire identity matrix appended to the right of the column of 1's. We then eliminate the least 'relevant' feature of X at each step. In particular, in a given iteration where there are q features, we find the $q-1$ features such that the determinant of the sample conditional covariance matrix, $\hat{\Sigma}_\varepsilon$, is minimized and keep those features. If there are ties, as in the case where $q \geq n-r$, we instead compare the product of nonzero diagonal elements of S, where USV^T is the SVD of $\hat{\Sigma}_\varepsilon$ (when all diagonal elements are nonzero, this is the determinant). When there are ties in this nonzero product, as in the case where $q = n$, we instead compare $\left\|\hat{\beta}_{\text{MLE}}\right\|_1$, where the norm is the $L1$ vector norm rather than the matrix norm. Note that a high breakdown robust start is not required to initiate this process.

2.2 Other Implications of the Methodology

The ideas in this paper lead to immediate extensions in the case where we have some prior knowledge or a good heuristic. If, for instance, we had some candidate outliers, we could append only dummy columns corresponding to them to the column of 1's in the design matrix, rather than appending the entire identity matrix. Alternatively, one could also add artificial rows to the bottom of X and Y to specify prior knowedge in the sense of mixed estimation [12]. In addition, if we could specify a reasonable guess at what Σ could be, say Σ_{guess}, then one could use alternative feature selection algorithms and dimension reduction algorithms. A full exploration of these other implications will be future work.

3. Results on Real and Simulated Data

3.1 Description of Data Sets

We evaluate backward elimination on two data sets: The Hertzsprung-Russell Star Data [13] and the Wine Data [14].

Our first data set, the Hertzsprung-Russell Star Data set is an example of a Hertzsprung-Russell star

data set used to make a Hertzsprung-Russell star diagram, which is a scatter plot of a star's luminosity against its temperature on a log scale. In the original data set, there are 47 observations and one explanatory variable. Figure 1 shows a plot of the data set. There are four gross outliers at the upper left and two moderate outliers (in Figures 1 and 2, these are the six observations with the highest number, although in the original data sets, these are observations 11, 20, 30, and 34, and observations 7 and 9 respectively). The data set is typically used to demonstrate robust linear regression algorithms, but we treat the response and explanatory data together as multivariate data in this paper, because it is nonetheless a good demonstration of influential, masked outliers, which can easily be visualized.

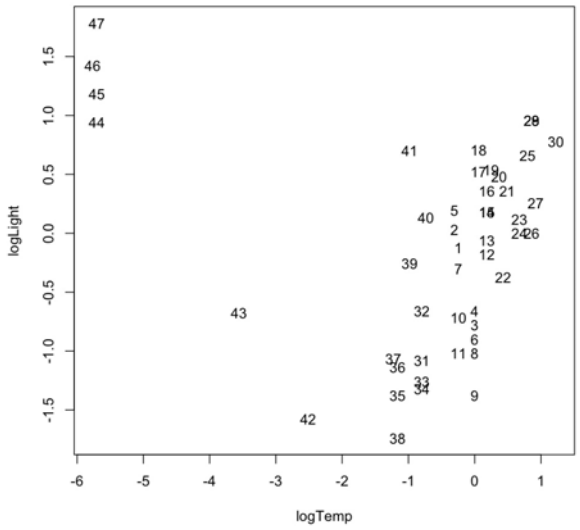

Figure 1: Plot of Hertzsprung-Russell Star Data with Sequence Numbers

Our second data set, the Wine Data, is explored in [1]. It contains, for each of 59 wines grown in the same region in Italy, the quantities of 13 constituents. There are known to be at least 7 masked outliers in this data set, the corresponding observation numbers being: {20, 22, 47, 40, 44, 46, 42}.

3.2 Evaluation of our method

For each of the two data sets, the specified outliers are among the last features eliminated in our algorithm, indicating that our algorithm correctly identifies them as the most outlying points. Figure 1 shows the plot of the Hertzsprung-Russell Star Data with the point number representing when in the algorithm the dummy variable corresponding to the point was eliminated thus adding that observation back into the data set. This figure demonstrates that our algorithm does well from the start (when the entire identity matrix is appended).

Figures 2 and 3 show the plot of the log of the determinant of the covariance matrices vs. the step of the algorithm for the respective data sets. In Figure 3, the outliers {20, 22, 47, 40, 44, 46, 42} are the seven observations added back at the right (53 to 59). Based on the descriptions of the data sets, the indicated time to stop is clear for the Hertzsprung-Russell Star Data and Wine Data.

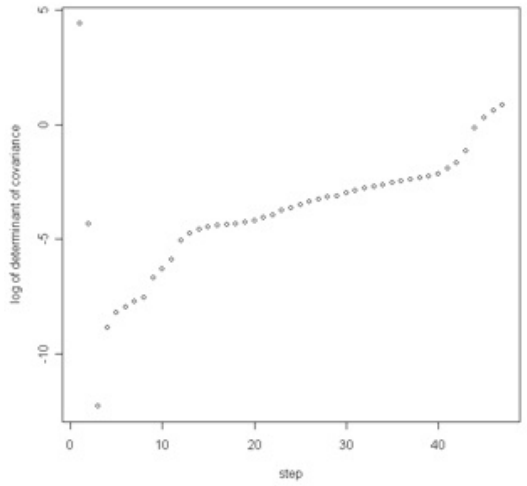

Figure 2: Plot for Hertzsprung-Russell Star Data Outlier Selection

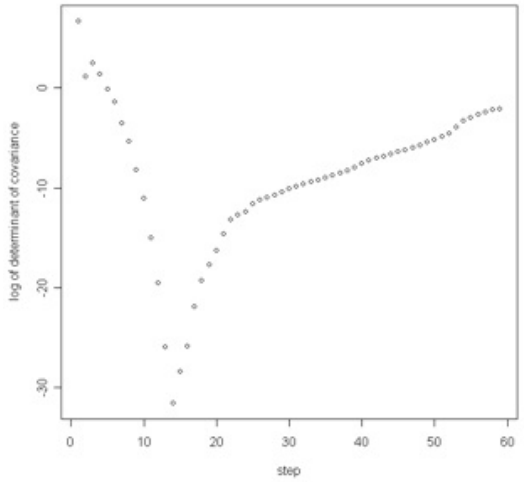

Figure 3: Plot for Wine Data Outlier Selection

It should also be noted that although our algorithm identified the seven prominent masked outliers in the wine data, the description of the wine data indicated the presence of some other minor outliers, some of which the backwards selection algorithm eliminated early in its search. Hence, our algorithm is not perfect, but we tested it in the harder case with the assistance of no prior knowledge or heuristic. When we provided our algorithm with a hint by appending an extra row corresponding to adding prior knowledge that the data should be centered near the coordinate-wise median, the algorithm worked satisfactorily for the data set.

4. Conclusion

We have proposed an alternative way of viewing the problem of outlier detection in multivariate data, by posing it as a feature selection problem for a multivariate linear model. We have also implemented this idea, using backward selection as a feature selection method, producing good results on the data sets attempted. Backward selection in itself is not guaranteed to work on any data set, but we hope that it provides one useful method for the computation of robust covariance matrices, and that the idea of viewing robust covariance estimation as a feature selection problem in a multivariate linear model leads to the development of other effective methods for detecting outliers in multivariate data.

REFERENCES (RÉFERENCES)

[1] Ricardo A Maronna, R. Douglas Martin, and Victor J. Yohai, 2006. *Robust Statistics: Theory and Methods*, John Wiley & Sons Ltd., West Sussex, England.

[2] Khan J., Van Aelst, S., and Zamar, R.H., 2007. Robust Linear Model Selection based on Least Angle Regression, *Journal of the American Statistical Association*, 102(480), 1289-1299.

[3] Hubert, M., Rousseeuw, P.J., Aelst, S.V., 2008. High-breakdown Robust Multivariate Methods. *Statistical Science*, 23(1), 92-119.

[4] Nguyen, T. and Welsch, R.E., 2010. Robust Regression Using Semi-Definite Programming, *Computational Statistics and Data Analysis*, 54, 3212-3226.

[5] Atkinson, A.C., Riani, M., 2004. *Exploring Multivariate Data with the Forward Search*, New York, Springer-Verlag.

[6] Wei, W.H, Fung, W.K., 1999. The Mean-shift Outlier Model in General Weighted Regression and its Applications, *Computational Statistics and Data Analysis*, 30(4), 429-441.

[7] Morgenthaler, S.,Welsch, R. E., and Zenide, A., 2003. Algorithms for Robust Model Selection in *Linear Regression. Theory and Applications of Recent Robust Methods*, eds. M. Hubert, G. Pison, A. Struyf, and S. Van Aelst, Basel (Switzerland), Birkhauser-Verlag.

[8] McCann, L. and Welsch, R. E., 2007. Robust Variable Selection Using Least Angle Regression and Elemental Set Sampling. *Computational Statistics and Data Analysis*, 52, 249-257.

[9] Kim, S., Park, S.H., and Krzanowski, W.J., 2008. Simultaneous Variable Selection and Outlier Identification in Linear Regression Using the Mean-shift Outlier Model. *Journal of Applied Statistics*, 35(3), 283-291.

[10] McCann, L., Welsch, R., 2004. Diagnostic Data Traces Using Penalty Methods in *Proceedings in Computational Statistics: COMPSTAT 2004* edited by J. Antoch, Physica-Verlag, Heidelberg, 1481-1488.

[11] McCann, L., 2005. Robust Model Selection and Outlier Detection in Linear Regression. PhD Thesis. MIT.

[12] Theil, H. and Goldberger, A.S., 1961. On Pure and Mixed Estimation in Economics, *International Economic Review*, 2, 65-78.

[13] Rousseeuw, P.J., and Leroy, A.M., 1987. *Robust Regression and Outlier Detection*, John Wiley & Sons, Inc. NY.

[14] Hettich, S. and Bay, S.D., 1999. The UCI KDD Archive, kdd.ics.uci.edu, Irvine, CA: University of California, Department of Information and Computer Science.

[15] Menjoge, R. S., 2010. New Procedures for Visualizing Data and Diagnosing Regression Models, PhD

Thesis, M.I.T.

RÉSUMÉ (ABSTRACT)

Recent literature has established a connection between outlier detection in linear regression, and feature selection on an augmented design matrix. In this paper, we establish this connection for the case of outlier detection in multivariate analysis, where we are interested in the estimation of a robust covariance matrix. This connection makes it possible to use feature selection and dimension reduction algorithms for multivariate linear models in order to detect and control for outliers in multivariate data, thus opening up a new class of algorithms to deal with multivariate outliers. We explore one such algorithm, using backwards selection as the feature selection method of choice. We test this algorithm on real data and discuss the results.