Computational Issues in Nonlinear Optimization for Robust Estimation and Outlier Detection Using a Multivariate Minimum Distance Criterion

David W. Scott
Rice University, Department of Statistics
MS-138, 6100 Main
Houston, TX 77005
scottdw@rice.edu

A minimum distance criterion is automatically robust when fitting a parametric model. Such an approach may be used to identify multivariate outliers and to fit multivariate parameters, including regression.

However, if a minimum distance criterion is chosen that is not dimensionless, then care must be exercised in scaling the criterion to be used in the nonlinear optimization code. In this paper, we illustrate the algorithms and describe how dimension limits the application.

Introduction

The study of spatial data has an important role in the mission of governmental entities. The topic has drawn the interest of statisticians (Cressie, 1991) and geographers (MacEachren, 1995), among others. One of the many challenges in mapping is a better understanding of how many variables are related spatially to variables of interest. One classical dataset much studied is the Boston housing data (Harrison et al, 1978); the goal is to predict median housing prices, but much of the data does not fit a single multivariate model. Robust methods are often applied to such data, but are iterative and sometimes difficult to interpret.

We have introduced an alternative multivariate regression fitting algorithm, called L2E (Scott, 2001). In place of a maximum likelihood or a least-squares criterion, L2E attempts to estimate parameters so the *shape* of the residuals is as close to Normal as possible. Why does this alternative criterion help? If the data contain a cluster of "bad" data fitted by least-squares, the residuals will not be Normal and may be skewed, for example. Regression diagnostic plots must be carefully screened and interpreted in order to identify and then correct these practical problems. Making the residuals look Normal will avoid such difficult diagnostic steps. Instead, 90% of the residuals will look Normal (with a mean of 0), and the 10% of residuals corresponding to the bad data will be clearly shown.

In the case of spatial data, another diagnostic is available. Namely, the values of the residuals may be plotted spatially in order to better understand the nature of the "bad" data. Of course, the bad data may in fact be the most interesting data. A simple estimation example is presented.

Fitting Multivariate Regressions by L2E

Given a variable of interest, y, and a number of covariates, $(x_1, x_2, \ldots, x_p)$, collected over a number of spatial units (census tracks, for example), we seek to model y as a simple function of the predictors: $\hat{y}(x) = \beta^T x$. Given estimates of the $\beta_j' s$, a residual for the i-th case would be computed as $\epsilon_i = y_i - \beta^T x_i$. If there was a cluster of bad data, then the residuals would not look Normal and be centered at 0.

With the L2E, least-squares is replaced by another criterion, which can be optimized using standard software, for example, *nlmin* in the Splus package. The criterion to minimize is simply

$$\frac{1}{2\sqrt{\pi}\sigma_\epsilon} - \frac{2}{n}\sum_{i=1}^{n}\phi(\epsilon_i|0,\sigma_\epsilon^2),$$

where σ_ϵ^2 is the variance of the residuals and $\phi(x|\mu,\sigma^2)$ is the Normal density.

The new exciting extension which we will illustrate was introduced by Scott and Szewcyzk (2003). Specifically, instead of modeling the residuals as Normal, $N(0,\sigma_\epsilon^2)$, we add an additional weight parameter, w, and use the residual model $w \cdot N(0,\sigma_\epsilon^2)$. What this unusual model is really doing can be made clearer by the following observation. If we must deal with a "bad" data cluster, then the residual plot will have a separate component for that cluster. In other words, we believe the residual density is actually a mixture of two components, one centered at 0 (the good data), and a second elsewhere (and of unknown shape). The "magic" of the L2E algorithm, as described in Scott and Szewczyk (2001), is that we can use L2E to estimate only the major mixture component. The criterion given above is modified only slightly to:

$$\frac{w^2}{2\sqrt{\pi}\sigma_\epsilon} - \frac{2\,w}{n}\sum_{i=1}^{n}\phi(\epsilon_i|0,\sigma_\epsilon^2).$$

The optimization is over the parameters w and σ_ϵ, as well as the parameters $\{\beta_0,\beta_1,\ldots,\beta_p\}$, which are used implicitly to recompute the estimated residuals, $\{\epsilon_i\}$.

The estimated value of the weight, w, indicates the amount of the data that the L2E algorithm believes is being fitted by the multivariate regression model. Thus, when all works well, the investigator simultaneously obtains a model fit that is very good as though only the "good" data were fitted, as well as an explicit estimate of the fraction of "bad" data. By sorting on the magnitude of the fitted residuals, the "suspect" data can clearly be identified.

Application to Boston Housing Data

The Boston housing data were originally modeled to determine if levels of air pollution appeared to be correlated with housing prices. Data were assembled for the 506 census tracts in greater Boston in the early 1970's. The 13 predictor variables included information on per capita crime rates, proportion of residential land zoned for large lots, proportion of non-retail business acres, average age and number of rooms per dwelling, full-value property-tax rates, and pupil-teacher ratios, in addition to the nitric oxides concentration (the pollution variable). Dummy variables were added to account for spatial correlation, for example, adjacency to the Charles River, distances to five Boston employment centers, and accessibility to radial highways.

In the talk, we will examine the residuals and their display on maps, using the ArcView program. The interesting fact is that almost 15% of the census tracks are estimated to be outliers by L2E.

Discussion and Future Directions

The digital government work of our research team has focused on a number of statistical tools and techniques which can handle many of challenging real situations. In this paper, we have examined the all-too-common problem of statistical models which fit a large fraction of the data, but not all of the data. The common practice of iteratively trying to figure out which data points are the problematic ones is very difficult and, in our experience, often leads to unnecessary failure. This is not too surprising as the number of possible remedies with 13 predictor variables is almost unlimited, and both a novice and experienced investigator may not be able to find the correct combination of fixes.

In contrast, our new fitting technique goes straight to the heart of the issue, finding the "best" for the largest fraction of "good" data that can be identified. Diagnostics are much easier with this approach. The generality of our approach will become clearer, as regression is the fundamental model for many if not most data analyses.

Acknowledgments:

Research was supported in part by the National Science Foundation grants NSF EIA-9983459 (digital government) and DMS 02-04723 (non-parametric methodology). The authors would like to thank our Digital Government collaborators Drs. Carr, MacEachren, and Brewer.

REFERENCES

Brewer, Cynthia A. (1999), "Color Use Guidelines for Data Representation," Proceedings of the Section on Statistical Graphics, American Statistical Association, Baltimore, pp. 55-60.

Cressie, Noel A. C. (1991), *Statistics for spatial data,* Wiley-Interscience, New York.

Harrison, D. and Rubinfeld, D.L. (1978), "Hedonic Housing Prices and the Demand for Clean Air," Journal of Environmental Economics and Management, 5, 81-102.

MacEachren, Alan M. (1995), *How maps work,* Guilford Publications.

Scott, D.W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization,* John Wiley, New York.

Scott, D.W. (2000), "Multidimensional Smoothing and Visualization," In *Smoothing and Regression. Approaches, Computation and Application,* M. G. Schimek, Ed., John Wiley, New York, pp. 451-470 (with 5 color plates).

Scott, D.W. (2001), "Parametric Statistical Modeling by Minimum Integrated Square Error," Technometrics, 43, pp. 274–285.

Scott, D.W. and Whittaker, G. (1996), "Multivariate Applications of the ASH in Regression," Communications in Statistics, 25, pp. 2521-2530.

Scott, D.W. and Szewczyk, W.F. (2003), "The Stochastic Mode Tree and Clustering," J. Comp. Graph. Stat., to appear.

RÉSUMÉ

Nous etudions un critère pour l'estimation des fonctions de densité et régression. Ils sont robuste. Quelques examples sont presenté.