行政院國家科學委員會專題研究計畫成果報告
資料之變數轉換的穩健估計(2/2)
Robust Estimation in Data Transformations (2/2)

計畫編號：NSC 90－2118－M－004－013
執行期限：90 年 8 月 1 日至 91 年 7 月 31 日
主持人：鄭宗記　執行機構及單位名稱: 國立政治大學統計系

中文摘要

常態分配之假設在迴歸與多變量分析中提供了一個方便且有力的途徑，當資料不是常態時，一個適當的變數轉換可使問題簡單化，亦即藉由變數轉換的過程使資料符合常態的假設，例如 Box-Cox 轉換(1964)。另一方面有時僅是因資料中所存在的一個觀測值或數個觀測值，則必須考慮對變數做轉換。也就是說，變數轉換的過程極易受到離群值的影響。

本研究之主要目的，在於利用穩健統計估計方法，使多變量資料在轉換的過程不受到離群值的影響。因此，我們提出一個穩健檢定統計量，其結果將使資料之轉換結果符合常態假設，並提供轉換後之穩健估計，並判斷觀測值在轉換後是否為離群值。

關鍵詞：Box-Cox 轉換、離群值偵測、前進搜尋演算法、高破壞點估計式、穩健估計

Abstract

The assumption of normality provides the customary powerful and convenient way of analyzing linear regression problem and multivariate data. The problem of non-normality may often be simplified by an appropriate transformation, e.g. the parametric family of power transformations of Box and Cox (1964). The evidence for transformations may sometimes depend crucially on e one or a few observations. Therefore, multivariate data transformations are very sensitive to outliers.

The purpose of the paper is to develop methods that would not be influenced by potential outliers during the process of data transformations. They essentially need robust statistics. We propose a robust likelihood ratio test for the transformation parameters. The resulting methods will be able to verify the role of every observation playing in the data transformation as well as to provide the robust estimation after transformation.

Keywords: Boc-Cox transformation; Detection of multiple outliers; High breakdown point; Robust estimation.

1. Multivariate transformations to normality

For multivariate data, let y_i be the $p\times1$ vector of responses at observation i with y_{ij} the observation on response j. In the extension of the Box and Cox (1964) family to multivariate responses the normalized transformation of y_{ij} is

$$z_{ij}(\alpha)=\begin{cases}\dfrac{y_{ij}^{\alpha_j}-1}{\alpha_j \dot{y}_j^{\alpha_j-1}} & \alpha\neq 0\\ \dot{y}_j \log(y_{ij}) & \alpha=0\end{cases}$$

where $\dot{y}_j$ is the geometric mean of the jth response. The value $\alpha_j=1$ ($j=1,\ldots,p$) corresponds to no transformation of any of the response. If the transformed observations are normally distributed with mean μ_j for the ith observation and covariance matrix Σ, twice the profile log likelihood of the observations is given by

$$2L_{\max}(\alpha) = const - n\log|\hat{\Sigma}(\alpha)| -$$
$$\sum_{i=1}^{n}\{(z_i(\alpha) - \hat{\kappa}_i(\alpha))^T\hat{\Sigma}(\alpha)^{-1}(z_i(\alpha) - \hat{\kappa}_i(\alpha)\}$$

$$= const - n\log|\hat{\Sigma}(\alpha)| -$$
$$\sum_{i=1}^{n}\{e_i(\alpha)^T\hat{\Sigma}(\alpha)^{-1}e_i(\alpha)\} \qquad (1)$$

In (1) $\hat{\kappa}_i(\alpha)$ and $\hat{\Sigma}(\alpha)$ are derived from the least squares for fixed α and $e_i(\alpha)$ is the $n\times 1$ vector of residuals. The least squares estimates can be applied to find the $n\times p$ matrix of parameter estimates

$$\hat{}(\alpha) = (X^TX)^{-1}X^Tz(\alpha).$$

Then, in the usual way,

$$\hat{\Sigma}(\alpha) = \sum_{i=1}^{n} e_i e_i^T$$
$$= \sum_{i=1}^{n}\{(z_i(\alpha) - \hat{\kappa}_i(\alpha))^T(z_i(\alpha) - \hat{\kappa}_i(\alpha)\}$$

When these estimates are substitute in (1), the profile likelihood reduces to

$$2L_{\max}(\alpha) = const - n\log|\hat{\Sigma}(\alpha)|.$$

Therefore, to test the hypothesis $\alpha = \alpha_0$, the statistic

$$T_{LR} = n\log\{|\hat{\Sigma}(\alpha_0)|/|\hat{\Sigma}(\hat{\alpha})| \qquad (2)$$

is compared with the χ^2 distribution on p degree of freedom. In (2) $\hat{\alpha}$ is the vector of p parameter estimates maximize (2), which is found by numerical search. The details for above can be found in Atkinson (1995).

2. Robust estimation

Both the minimum volume ellipsoid (MVE) and the minimum covariance determinant (MCD) estimators provide high breakdown robust estimation of multivariate location and shape (Rousseeuw and Leroy 1987). Moreover, Butler *et al.* (1993) show that the MCD estimator has better theoretical properties than the MVE. Woodruff and Rocke (1996) also give empirical results which show that the MCD is preferred over the MVE in their applications. The role of the MCD for multivariate data is similar to that of the least trimmed squares (LTS) estimator for linear regression (Atkinson and Cheng 1999).

Hadi and Luceno (1997) propose a trimmed likelihood principle based on trimming the likelihood function rather than directly trimming the data. They show that this trimming likelihood principle produces many existing estimators, such as MLE, LMS, LTS and MVE. It is always possible to order and trim observations according to their contributions to the likelihood function, because the likelihood is scalar-valued. They refer to the method as the *maximum trimmed likelihood* (MTL) method and to the corresponding estimator as the *maximum trimmed likelihood estimators* (MTLE).

In this project, we first showe that the MCD is also the MTLE of mean κ and covariance matrix Σ.. The details are not shown here.

3. Robust transformation to normality

3.1 Robust likelihood ratio test

It is known that the statistic (2) will be affected by multiple outliers. Therefore, based on Section 2, we propose the robust likelihood ratio test

$$T^*_{LR} = n\log\{|\hat{\Sigma}_q(\alpha_0)|/|\hat{\Sigma}_q(\hat{\alpha})|, \quad (3)$$

where $\hat{\Sigma}_q$ is the robust estimation of covariance matrix based on MCD criterion. Here the value of q is the number of observations are used to obtain the MCD.

3.2 The forward search algorithm

The forward search algorithm is proposed by Atkinson (1994). The advantages of the algorithm are that it not only rapidly finds estimates satisfying the criterion but it also leads to the detection of multiple outliers.

On the other hand, it is easily to obtain the result (3) with no need of extra computation during the process of the forward search for the MCD. The algorithm starts with a elemental subset (usually p+1 observations). Then the forward step is to add one or few observations based on the smallest squared Mahalanobis distances, which are also based on the MCD. The algorithm ends at all observations included. Those potential outliers intend to be excluded at the earlier stages of the forward procedure. We can identify outliers as well as obtain the robust estimates at the same time. Moreover, the

robust likelihood ratio test for transformation parameter can be monitored. The resulting statistics can be visualized by the stalactite plot and fan plot.

5. References

[1] Atkinson, A. C. (1994) “Fast Very Robust Methods for the Detection of Multiple Outliers”, *Journal of the American Statistical Association*, **89**, 1329-1339.

[2] Atkinson, A. C. (1995) “Multivariate Transformations, Regression Diagnostic and Seemingly Unrelated Regression”, Kitsos, C. P. and Muller, W. G. eds., *Proceedings of MODA4*, Physica Verlag, Heidelberg.

[3] Atkinson, A. C. and Cheng, T.-C. (1999) ``Computing the Least Trimmed Squares Regression with the Forward Search'', *Statistics and Computing*, **9**, 251-263.

[4] Box, G. E. P., and Cox, D. R. (1964), “An Analysis Of Transformations”(with discussion), *Journal of the Royal Statistical Society*, Ser. B, **26**, 211-246.

[5] Butler, R. W., Davies, P. L. and Jhun, M. (1993) “Asymptotics for the Minimum Covariance Determinant Estimator”, *The Annals of Statistics*, 21, 1385-1400.

[6] Hadi, A. S. and Luceno, A. (1997) ``Maximum Trimmed Likelihood Estimators: a Unified Approach, Examples, and Algorithms'', *Computational Statistics & Data Analysis*, **25**, 251-272.

[7] Rocke, D. M. and Woodruff, D. L. (1996) ``Identification of Outliers in Multivariate Data'', *Journal of the American Statistical Association*, **91**, 1047-1061.

[8] Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: John Wiley.