Robust Transformations and Outlier Detection with Autocorrelated Data

Andrea Cerioli and Marco Riani

Department of Economics – Section of Statistics,
University of Parma, 43100 Parma, Italy

Abstract. The analysis of regression data is often improved by using a transformation of the response rather than the original response itself. However, finding a suitable transformation can be strongly affected by the influence of a few individual observations. Outliers can have an enormous impact on the fitting of statistical models and can be hard to detect due to masking and swamping. These difficulties are enhanced in the case of models for dependent observations, since any anomalies are with respect to the specific autocorrelation structure of the model. In this paper we develop a forward search approach which is able to robustly estimate the Box-Cox transformation parameter under a first-order spatial autoregression model.

1 Introduction

The development of robust high-breakdown methods for spatially autocorrelated data is an important research topic. Models for such data are usually fitted through maximum likelihood under a Gaussian assumption. It is notorious that maximum likelihood estimation is not robust to the presence of outliers. Furthermore, spatial autocorrelation can be the reason for additional troubles in the outlier detection process, since any anomalies have to be checked with respect to the assumed spatial model and neighbourhood structure. Cressie (1993) provides a wide description of exploratory tools that can be applied to uncover spatial outliers and that make use of neighbourhood information. However, these methods are based on case-deletion diagnostics and are prone to masking and swamping with a cluster of spatial outliers.

Most high-breakdown methods for regression and multivariate estimation, such as least median of squares regression and minimum volume ellipsoid estimation (Rousseeuw and van Zomeren (1990)), are difficult to extend to autocorrelated observations, both conceptually and computationally. Cerioli and Riani (2002) and Atkinson et al. (2004) suggest a forward search approach to robustly fit spatial models. Their technique rests upon a computationally simple and statistically efficient forward algorithm, where at each step observations are added to the fitted subset in such a way that outliers and influential observations enter at the end.

In this paper we show how the forward search approach can be extended to robustly improve normality of spatially autocorrelated data, a topic that

has been somewhat neglected in the statistical literature (Griffith and Layne (1995) and Pace et al. (2004) are two non-robust exceptions). Specifically, in §2 we focus on the popular first-order Simultaneous Autoregressive (SAR) model. Transformation of the response using the Box-Cox family of power transformations is considered in §3, where we introduce the notion of a transformed SAR model. §4 gives an overview of the forward search algorithm used for fitting the transformed model. The usefulness of our method is shown in §5 through a number of examples.

2 The Simultaneous Autoregressive (SAR) Model

Let $S \equiv \{s_1, \ldots, s_n\}$ be a collection of n spatial locations and y_i be a random variable observed at site s_i, $i = 1, \ldots, n$. Spatial relationships between pairs of locations are represented through the simple weighting scheme:

$$\begin{aligned} w_{ij} &= 1 \qquad \text{if sites } s_i \text{ and } s_j \text{ are neighbours,} \\ w_{ij} &= 0 \qquad \text{otherwise,} \end{aligned}$$

and $w_{ii} = 0$. For a regular grid the most common definition of a neighbourhood structure is that for which $w_{ij} = 1$ if s_j is immediately to the north, south, east or west of s_i. We write $y = (y_1, \ldots, y_n)'$ and $W = (w_{ij})$ for $i, j = 1, \ldots, n$. Edge points typically raise problems in the statistical analysis of spatial systems. The basic difficulty is that they have fewer neighbours than interior points. For this reason we assume that, whenever possible, W has been suitably modified to account for edge effects. A simple but widely adopted technique is toroidal correction, which wraps a rectangular region onto a torus. Edge points on opposite borders are thus considered to be close, and all sites have the same number of neighbours.

At each location we might have additional non-stochastic information about $p-1$ spatial covariates. Let X denote the corresponding design matrix of dimension $n \times p$, allowing also for the mean effect. The first-order Simultaneous Autoregressive (SAR) model is defined as (Cressie, 1993)

$$(I_n - \rho W)(y - X\beta) = \varepsilon, \tag{1}$$

where $\beta = (\beta_0, \ldots, \beta_{p-1})'$ is a p-dimensional parameter vector, I_n is the $n \times n$ identity matrix, ρ is a measure of spatial interaction between neighbouring sites, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ is an n-dimensional vector of disturbances. Errors ε_i are defined to be independent and normally distributed with mean 0 and common variance σ^2. It is assumed that $(I_n - \rho W)^{-1}$ exists. It is not essential for W to be symmetric, although in practice this is often the case.

Estimation of parameters in (1) is by maximization of the likelihood

$$l(\beta, \sigma^2, \rho) = (2\pi\sigma^2)^{-n/2} |I_n - \rho W| \exp\{-\frac{1}{2\sigma^2}(y - X\beta)'\Sigma(y - X\beta)\}, \tag{2}$$

where

$$\Sigma = (I_n - \rho W)'(I_n - \rho W).$$

3 The Transformed SAR Model

One crucial assumption underlying model (1) is normality of the additive errors ε_i. If this requirement is not satisfied in the original scale of measurement of the response, it may be that there is a nonlinear transformation of y which will yield normality, at least approximately. In this paper we adopt the popular class of nonlinear transformations suggested by Box and Cox (1964).

Let $y_* = (y_{*1}, \ldots, y_{*n})' = (I_n - \rho W)y$. Under the first-order SAR model (1), y_* has mean $(I_n - \rho W)X\beta$ and scalar covariance matrix. For this modified data vector, the Box-Cox normalized power transformation to normality is

$$z(\lambda) = \begin{cases} \frac{y_*^{\lambda} - 1}{\lambda \dot{y}_*^{\lambda - 1}} & \lambda \neq 0 \\ \dot{y}_* \log y_* & \lambda = 0 \end{cases} \tag{3}$$

where $\dot{y}_* = \exp(\Sigma_i \log y_{*i}/n)$ is the geometric mean of $y_{*1}, \ldots, y_{*n}$. We define the transformed SAR model to be a linear regression model with response $z(\lambda)$, design matrix $(I_n - \rho W)X$ and Gaussian disturbance ε, as in model (1). That is,

$$z(\lambda) = (I_n - \rho W)X\beta + \varepsilon. \tag{4}$$

When $\lambda = 1$, there is no transformation in the standard SAR model; $\lambda = 1/2$ is the square root transformation, $\lambda = 0$ gives the log transformation and $\lambda = -1$ the reciprocal. These are the most widely used transformations in practical applications.

Maximum likelihood estimation of λ could be performed by suitable modification of equation (2). However, likelihood analysis for spatial Gaussian processes can encounter numerical difficulties, such as convexity or multimodality of the resulting profile likelihood function (see Ripley (1988), §2.1), and adding a further parameter to (2) might result in unpredicted consequences. In addition, repeated evaluation of model (4) at subsequent steps of the forward algorithm requires a fast computational procedure. For these reasons, we do not resort to numerical maximization of the likelihood function with respect to the extended parameter set $(\beta, \sigma^2, \rho, \lambda)$. Following Atkinson and Riani (2000, §4.2), instead we derive an approximate score statistic by Taylor series expansion of (3) about a known value λ_0. The score statistic does not require computation of the maximum likelihood estimate of λ.

As a result, the transformed SAR model (4) is approximated as

$$z(\lambda_0) \doteq (I_n - \rho W)X\beta + \gamma w(\lambda_0) + \varepsilon, \tag{5}$$

where $\gamma = -(\lambda - \lambda_0)$ and $w(\lambda_0) = \left.\frac{\partial z(\lambda)}{\partial \lambda}\right|_{\lambda = \lambda_0}$ is known as a constructed variable. The t test for $\gamma = 0$ in model (5) is then the approximate score statistic for testing

$$H_0 : \lambda = \lambda_0 \tag{6}$$

in the transformed SAR model (4). This statistic makes proper allowance for spatial autocorrelation in the process of finding the best transformation, thus achieving the desirable goal of a joint spatial and transformation analysis.

4 Robust Fitting of the Transformed SAR Model and Diagnostic Monitoring

The transformed SAR model is repeatedly fitted through a block forward search (BFS) algorithm similar to the one suggested by Cerioli and Riani (2002). This algorithm is both efficient and robust. It is efficient because it makes use of the Gaussian likelihood machinery underlying models (1) and (4). It is robust because the outliers enter in the last steps of the procedure and their effect on the parameter estimates is clearly depicted. More generally, our approach allows evaluation of the inferential effect each location, either outlying or not, exerts on the fitted model. The key features of the BFS for finding the best transformation under model (4) are summarized as follows.

Choice of the initial subset. We take blocks of contiguous spatial locations as the basic elemental sets of our algorithm. Blocks are intended to retain the spatial dependence properties of the whole study region and are defined to resemble as closely as possible the shape of that region. Confining attention to subsets of neighbouring locations ensures that spatial relationships are preserved by the BFS algorithm, so that ρ can be consistently estimated within each block. Atkinson et al. (2004) provide details about practical selection of blocks and empirical evidence of the effects produced by different choices. The initial subset for the BFS algorithm is then obtained without loss of generality through a least median of squares criterion applied to blocks.

Progressing in the search. The transformed SAR model is repeatedly fitted to subsets of observations of increasing sizes, selected in such a way that outliers are included only at the end of the search. For this reason, progression in the BFS algorithm is performed by looking at the smallest squared standardized regression residuals from the fit at the preceding step. At each step, model (4) can be fitted either by exact maximum likelihood given the available data subset, or by a faster approximation to it. The weight matrix W is usually corrected for edge effects for the reason sketched in §2.

Diagnostic monitoring. One major advantage of the forward search over other high-breakdown techniques is that a number of diagnostic measures can be computed and monitored as the algorithm progresses. Under model (4), we are particularly interested in producing forward plots of regression parameter estimates and transformation statistics. In the latter instance, we produce forward plots of the approximate score statistic for testing (6) under different values λ_0, using a separate search for each λ_0. These plots are then combined into a single picture which is named a "fan plot" after Atkinson and Riani (2000, p. 89). In most applications five values of λ_0 are sufficient for selecting the appropriate transformation: 1, 0.5, 0, -0.5, -1, thus running from no transformation to the reciprocal.

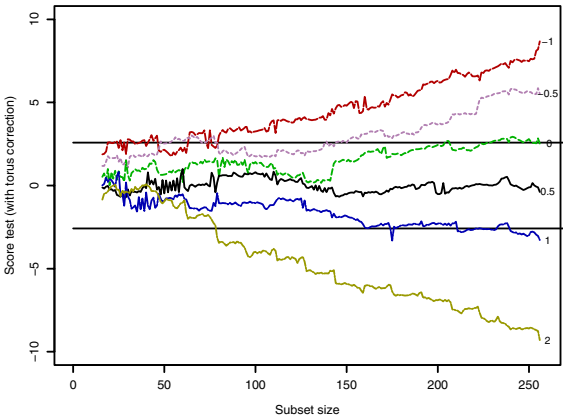

Fig. 1. Fan plot for the clean dataset of Example 1.

5 Examples

5.1 Example 1: Clean Data

In our first example we analyze the behaviour of the BFS algorithm for robustly fitting the transformed SAR model (4) in a dataset without outliers, to check that it does not produce spurious information. We first simulate $n = 256$ observations from model (1), with S a 16×16 regular grid, $\rho = 0.1$, $p = 4$, $w_{ij} = 1$ if s_j is immediately to the north, south, east or west of s_i, and toroidal edge correction. Then we square the response values. The dataset is available at `http://www.riani.it/gfkl2005`. Any sensible transformation analysis should point to the square root transformation of y, i.e. to $\lambda = 0.5$.

Figure 1 is the fan plot showing the forward plots of the approximate score statistic for testing hypothesis (6) under six values λ_0, ranging from -1 to 2, when the BFS is run with blocks of size 4×4 and toroidal edge correction. The central horizontal bands are at ± 2.58, the 99% percentage points of the reference asymptotic normal distribution. The fan plot clearly depicts the correct transformation $\lambda = 0.5$, as the corresponding score statistic varies around zero along the search. Evidence against the other values of λ increases as the fitting subset grows. There is no effect of outlying observations at the end of the search. We conclude that our method provides the appropriate transformation, as well as the effect on the choice of λ exerted by each spatial location, in this "clean" example.

We complement our transformation analysis by seeing how the forward plot of the maximum likelihood estimate of ρ changes under different values of λ. The corresponding plots are in Figure 2. Apart from the initial steps, where results from the search may be unstable, it is seen that estimation of ρ is not much affected by the specific transformation parameter. This indicates lack of appreciable interaction between the strength of spatial autocorrelation and the scale on which y is represented.

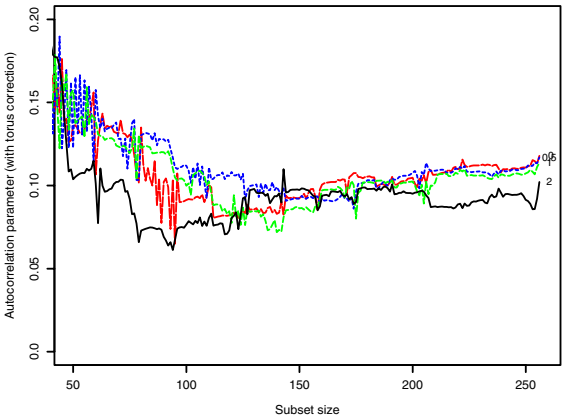

Fig. 2. Example 1. Forward plots of the maximum likelihood estimate of ρ under different transformations.

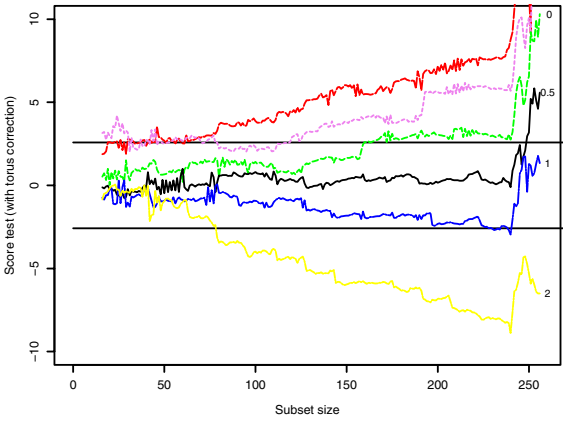

Fig. 3. Fan plot for the contaminated dataset of Example 2.

5.2 Example 2: Contaminated Data

In our second example we evaluate the robustness properties of the BFS approach for transformation to normality with correlated data. For this purpose, we introduce a cluster of 16 spatial outliers in the simulated dataset of Example 1, by modifying the response values in the 4×4 area in the left-hand corner of S. Also this dataset is available at `http://www.riani.it/gfkl2005`.

The outliers are masked and hard to detect by standard exploratory methods, such as visual inspection of the scatterplot matrix and diagnosis of the regression residuals. On the contrary, Figure 3 is extremely clear in picturing the influence that the outliers have on the selection of the transformation parameter. The true $\lambda = 0.5$ is correctly supported by all the non-contaminated data: the forward plot of the score statistic for testing $\lambda = 0.5$ varies around zero until the first spatial outlier is included in the fitted subset, at step

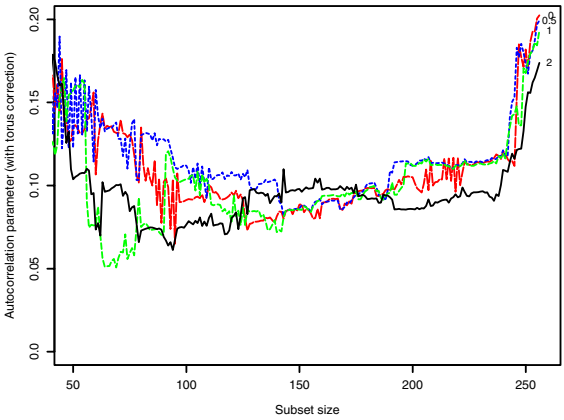

Fig. 4. Example 2. Forward plots of the maximum likelihood estimate of ρ under different transformations.

241. Even allowing for spatial autocorrelation, progressive inclusion of the outliers renders the correct transformation increasingly less plausible. Non-robust transformation analysis based on all the data would then wrongly suggest that this dataset does not need to be transformed ($\lambda = 1$). Furthermore, the outliers now have a disproportionate effect on estimation of ρ, again irrespective of the value of λ (Figure 4).

5.3 Example 3: Simulation Envelopes

In §3 we stressed the point that the approximate score statistic for testing (6) in the transformed SAR model makes proper allowance for spatial autocorrelation. However, it is not known how the asymptotic normal distribution approximates the true null distribution of the score statistic in small or moderate spatial samples. Therefore, it is useful to provide simulation evidence of the finite sample accuracy of approximation (5) and of the effect of spatial autocorrelation on the actual significance level of the score statistic.

Figure 5 reports 90%, 95% and 99% envelopes of the distribution of the score statistic obtained from 200 independent simulations of the transformed SAR model (4) under the null hypothesis, in the setting of Example 2 with ρ estimated at the step before the inclusion of the first outlier. These envelopes are compared with the corresponding percentage points of the normal distribution (the horizontal lines in the figure). After the first steps, it is seen that there is good agreement between asymptotic and simulated percentage points. This result strenghtens our confidence in pointwise inference based on simple displays such as the fan plots of Figures 1 and 3 with spatially autocorrelated data, at least when the sample size is moderately large and the transformed SAR model fits well the data.

Some preliminary simulation results (not reported here) seem to show that, for a fixed sample size, the accuracy of the normal approximation dete-

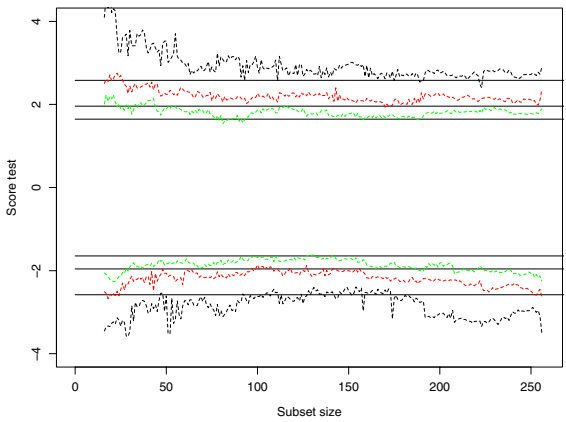

Fig. 5. Simulation envelopes (dashed bands) and asymptotic percentage points (solid lines) of the score statistic for testing $\lambda = \lambda_0$, in the setting of Example 2.

riorates as the model fit worsens. The development of a general approach for calibrating asymptotic confidence bands of the score statistic under different SAR model fits is currently under investigation.

References

ATKINSON, A.C. and RIANI, M. (2000): *Robust Diagnostic Regression Analysis*. Springer, New York.

ATKINSON, A.C., RIANI, M. and CERIOLI, A. (2004): *Exploring Multivariate Data with the Forward Search*. Springer, New York.

BOX, G.E.P. and COX, D.R. (1964): An Analysis of Transformations (with discussion). *Journal of the Royal Statistical Society B, 26, 211–246*.

CERIOLI, A. and RIANI, M. (2002): Robust Methods for the Analysis of Spatially Autocorrelated Data. *Statistical Methods and Applications - Journal of the Italian Statistical Society, 11, 335–358*.

CRESSIE, N.A.C. (1993): *Statistics for Spatial Data*. Wiley, New York.

GRIFFITH, D.A. and LAYNE, L.J. (1999): *A Casebook for Spatial Statistical Data Analysis*. Oxford University Press, New York.

PACE, R.K., BARRY, R., SLAWSON, V.C. Jr. and SIRMANS, C.F. (2004): Simultaneous Spatial and Functional Form Transformations. In: L. Anselin, R.J.G.M. Florax and S.J. Rey (Eds.): *Advances in Spatial Econometrics*. Springer, New York.

RIPLEY, B.D. (1988): *Statistical Inference for Spatial Processes*. Cambridge University Press, Cambridge.

ROUSSEEUW, P.J. and van ZOMEREN, B.C. (1990): Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association, 85, 633-639*.