DETECTION OF BOUNDARIES IN REGRESSION DATA IN THE PRESENCE OF SPATIAL CORRELATION

by

L. Xie
Shire Pharma
Rockville, MD

I. B. MacNeill
Department of Statistical and Actuarial Sciences
The University of Western Ontario
London, Canada N6A 5B9

Key Words: changepoint; change-boundary; spatial data; spatial correlation.

ABSTRACT

A regression model defined by a single set of parameters may be suitable for fitting a set of spatial data. However the data may be divided into subsets each of which is modelled using a different set of regression parameters. Statistics for detecting the presence of boundaries of such subsets of spatial data are available provided the data do not exhibit spatial correlation. Since the distribution of a change-boundary statistic is modified by the presence of spatial correlation, distributional results are presented to deal with the problem. Detecting boundaries in higher dimensional spatially correlated data is discussed.

1. Regression Models and Error Process Structure

We first define the basic 2-dimensional model. Let $X(n,m)$ $(n,m = 0,\pm 1,\dots)$ be a zero mean, stationary spatial series defined on a lattice with covariance function

$$R(u,v) = E\{X(t,s)X(t+u,s+v)\}, \quad |u|,\ |v| < \infty \ .$$

If the covariance function is absolutely summable, i.e., $\sum_{|u|<\infty}\sum_{|v|<\infty}|R(u,v)| < \infty$, then the spectral density function,

$$f(\lambda_1,\lambda_2) = \frac{1}{4\pi^2}\sum_{|u|<\infty}\sum_{|v|<\infty} e^{-i\lambda_1 u - i\lambda_2 v}R(u,v) \ , \quad \lambda_1,\lambda_2 \in [-\pi,\pi] \ ,$$

exists. Then for $t,s \in [0,1]$ and under regularity assumptions,

$$S_X([nt],[ns]) = \frac{1}{n}\sum_{i=1}^{[nt]}\sum_{j=1}^{[ns]} X(i,j) \ ,$$

converges in distribution to the normal with zero mean and variance $\{4\pi^2 f(0,0)ts\}$.

We now consider the regression model

$$Y_n(i,j) = \sum_{k=0}^{p}\beta_k g_k(i/n, j/n) + X(i,j) \ ,$$

where $\{g_k(\cdot,\cdot),\ 0 \le k \le p\}$ is a collection of non-stochastic regressor functions defined on the unit square. If we denote the vector of regression coeffients by $\beta = (\beta_0, \ldots, \beta_p)'$, the design matrix by $\mathbf{A}_n$, the stacked vector of observations by $\mathbf{Y}_n$ and the stacked vector of stationary spatial series by $\mathbf{X}_n$, then the model may be written in matrix form as $\mathbf{Y}_n = \mathbf{A}_n\beta + \mathbf{X}_n$. The regression parameter estimators are $\hat{\beta} = (\mathbf{A}_n'\mathbf{A}_n)^{-1}\mathbf{A}_n'\mathbf{Y}_n$. The matrix array of partial sums of regression residuals are defined as

$$S_{g_n}(k,l) = \sum_{i=1}^{k}\sum_{j=1}^{l}\{Y_n(i,j) - \hat{Y}_n(i,j)\}\ ,\quad 1 \le k,l \le n,$$

where $\hat{Y}(i,j) = \beta'\mathbf{g}(i/n, j/n)$ and $\mathbf{g}(i/n,j/n)' = (g_0(i/n,j/n), \ldots, g_p(i/n,j/n))$. If we define a sequence of stochastic processes $\{Z_{g_n}(t,s), t,s \in [0,1]\}$ $(n \ge 1)$ by:

$$Z_{g_n}([nt],[ns]) = \frac{1}{n}S_{g_n}([nt],[ns])$$

and if we let $\mathbf{e}_{nt,ns}$ denote the n^2-dimensional vector that has 1 for components where $\mathbf{X}_n$ has as its component $X_n(i,j)$ with $i \le [nt]$ and $j \le [ns]$ and zero otherwise, then we can write

$$nZ_{g_n}(t,s) = \mathbf{e}'_{nt,ns}\{\mathbf{I} - \mathbf{A}_n(\mathbf{A}_n'\mathbf{A}_n)^{-1}\mathbf{A}_n'\}\mathbf{X}_n\ .$$

To establish the limit process for $\{Z_{g_n}(t,s),\ t,s \in [0,1]\}$ we need first to examine the properties of the matrix array of partial sums of the error process $X(n,m)$ $(n,m = 0,\pm1,\ldots)$. Hence, we let $S_{X_n}(k,l) = \sum_{i=1}^{k}\sum_{j=1}^{l}X(i,j)$ and define another sequence of stochastic processes $\{Z_{X_n}(t,s),\ \ t,s \in [0,1]\}$ $(n \ge 1)$ by $nZ_{X_n}(t,s) = S_{X_n}([nt],[ns])$. It can be shown that the covariance kernels of the processes, $K_n(t_1,s_1;t_2,s_2)$ converge as follows:

$$\frac{1}{n^2}K_n(t_1,s_1;t_2,s_2) \to 4\pi^2 f(0,0)(t_1 \wedge t_2)(s_1 \wedge s_2)\ ,$$

where $t_1 \wedge t_2 = \min(t_1,t_2)$. Thus, if the process $\{Z_X(t,s),\ t,s \in [0,1]\}$ is defined by

$$Z_X(t,s) = \{4\pi^2 f(0,0)\}^{\frac{1}{2}}Z(t,s)\ ,$$

where $Z(t,s)$ is a Brownian sheet, then it can be shown that $Z_{X_n}(\cdot,\cdot)$ converges weakly to $Z_X(\cdot,\cdot)$.

We now consider the matrix array of partial sums of regression residuals when the error process is a stationary spatial series. The vector of regressor functions evaluated at (t,s) is denoted by $\mathbf{f}(t,s) = (f_0(t,s), \ldots, f_p(t,s))'$. It may be seen that the matrix

$$\lim_{n\to\infty}\frac{1}{n^2}(\mathbf{A}_n'\mathbf{A}_n) = G$$

has as its (i,j) th component

$$\int_0^1\int_0^1 f_i(t,s)f_j(t,s)dtds\ .$$

The inverse of G exists provided the regressor functions are linearly independent and square integrable; with this proviso, we define a multilinear form, $g(t_1, s_1; t_2, s_2)$, as follows:

$$g(t_1, s_1; t_2, s_2) = \mathbf{f}'(t_1, s_1) G^{-1} \mathbf{f}(t_2, s_2) .$$

Then we define a limit process $\{Z_{X_g}(t,s),\ t,s \in [0,1]\}$ by

$$Z_{X_g}(t,s) = Z_X(t,s) - \int_0^t \int_0^s \int_0^1 \int_0^1 g(t_1, s_1; t_2, s_2) dZ_X(t_2, s_2) dt_1 ds_1 ,$$

where $Z_X(t,s) = \sqrt{4\pi^2 f(0,0)} Z(t,s)$ and $Z(t,s)$ is the Brownian sheet. The partial sum process of regression residuals is given by

$$nZ_{X_{g_n}}(t,s) = \mathbf{e}'_{nt,ns}\{\mathbf{I} - \mathbf{A}_n(\mathbf{A}'_n\mathbf{A}_n)^{-1}\mathbf{A}'_n\}\mathbf{X}_n .$$

Then, under regularity conditions, $Z_{X_{g_n}}(t,s) \Longrightarrow Z_{X_g}(t,s)$. It can be shown that $E\{Z_{X_g}(t,s)\} = Z_{X_g}(0,0) = 0$, $t,s \in [0,1]$ and that the covariance kernel for any $t_1, s_1, t_2, s_2 \in [0,1]$ is

$$\begin{aligned} K(t_1, s_1; t_2, s_2) =& E\{Z_g(t_1, s_1) Z_g(t_2, s_2)\} \\ =& 4\pi^2 f(0,0)\{(t_1 \wedge t_2)(s_1 \wedge s_2) \\ & - \int_0^{t_1} \int_0^{s_1} \int_0^{t_2} \int_0^{s_2} g(t_1, s_1; t_2, s_2) dt_1 ds_1 dt_2 ds_2\} . \end{aligned}$$

2. Effect of Spatial Autocorrelation on Change Detection Statistics

For the case of i.i.d error structure with the set of boundaries on the unit square consisting of rectangles with sides parallel to the square and with one vertex $(0,0)$, a statistic for detecting change at unknown boundary in regression parameters is shown (see Xie and MacNeill 2000) to be

$$Q_{g_n} = \frac{1}{n^4\sigma^2} \sum_{l=1}^{n} \sum_{k=1}^{n} \left\{ \sum_{i=1}^{l} \sum_{j=1}^{k} [Y_n(i,j) - \hat{Y}_n(i,j)] \right\}^2 .$$

To make the statistic both operational and effective it is necessary to estimate σ^2 with an estimator that is consistent under both null and alternative hypotheses. Now assume the spatial error process is not i.i.d and $R(0,0)$ is used in place of σ^2. Note that

$$R(0,0) = \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} f(\lambda_1, \lambda_2) d\lambda_1 d\lambda_2 .$$

Then, if $Z_g(t,s) = \{4\pi^2 f(0,0)\}^{-1/2} Z_{X_g}(t,s)$,

$$Q_{g_n} \rightarrow \frac{4\pi^2 f(0,0)}{\int_{-\pi}^{\pi} \int_{-\pi}^{\pi} f(\lambda_1, \lambda_2) d\lambda_1 d\lambda_2} \int_0^1 \int_0^1 Z_g^2(t,s) dt ds \qquad (1).$$

The above results indicate that the large sample effects of spatial correlation on Q_{g_n} can be adjusted for precisely by multiplying the quantiles of distributions for the i.i.d case by

$$\frac{4\pi^2 f(0,0)}{\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(\lambda_1,\lambda_2)d\lambda_1 d\lambda_2} .$$

The same adjustment may be applied to any change-boundary statistic that is defined in terms of the squares of the partial sums of the residuals. These results have time series analogues (see Tang and MacNeill 1993).

3. Extension to Higher Dimensional Spaces

We first define the basic model for d-dimensional spaces. Let $X_{n_1,\ldots,n_d}(n_1,\ldots,n_d = 0,\pm1,\ldots)$ be a zero mean, stationary spatial series defined on a lattice with covariance function

$$R(u_1,\ldots,u_d) = E\{X(t_1,\ldots,t_d)X(t_1+u_1,\ldots,t_d+u_d)\},\quad |u_i| < \infty \ .$$

If the covariance function is absolutely summable, i.e., $\sum_{u_1=-\infty}^{\infty}\cdots\sum_{u_d=-\infty}^{\infty}|R(u_1,\ldots,u_d)| < \infty$, then the spectral density function,

$$f(\lambda_1,\ldots,\lambda_d) = \frac{1}{(2\pi)^d}\sum_{|u_1|<\infty}\cdots\sum_{|u_d|<\infty} e^{-i\sum_{i=1}^{d}\lambda_i u_i}R(u_1,\ldots,u_d)\ ,\quad \lambda_i\in[-\pi,\pi]\ ,$$

exists. Also

$$R(0,\ldots,0) = \int_{-\pi}^{\pi}\cdots\int_{-\pi}^{\pi} f(\lambda_1,\ldots,\lambda_d)\prod_{i=1}^{d}d\lambda_d \ .$$

Then, if we have, analogous to (1) except instead of the 2-dimensional case we have the d-dimensional case with

$$Z_{gd}(t_1,\ldots,t_d) = \{(2\pi)^d f(0,\ldots,0)\}^{-1/2}Z_{X_{gd}}(t_1,\ldots,t_d)\ ,$$

then the change-boundary quadratic form converges as follows,

$$Q_{gd_n} \to \frac{(2\pi)^d f(0,\ldots,0)}{\int_{-\pi}^{\pi}\cdots\int_{-\pi}^{\pi} f(\lambda_1,\ldots,\lambda_d)\prod_{i-1}^{d}d\lambda_i}\int_0^1\cdots\int_0^1 Z_{gd}^2(t_1,\ldots,t_d)\prod_{i=1}^{d}dt_i$$

where Z_{yd} and $Z_{X_{gd}}$ are d-dimensional Brownian sheets. Thus, the distributional results for the stationary case can be found from the i.i.d. case by a simple precise adjustment.

REFERENCES

- *Tang, S.M. and MacNeill, I.B. 1993,* The effect of serial correlation on tests for parameter change at unknown time, Annals of Statistics, pp. 552-575.
- *Xei, L. and MacNeill, I.B., 2000,* Detection and estimation of boundaries in spatial data for regression models, Accuracy 2000 pp. 743 746.