1 Simulation study

1.1 Rationale

Our objective is to identify outliers given a fitted simultaneous autoregressive (SAR) model. Consider n spatial units and p parameters, we derived the analytic expression to case influence analysis under the Gaussian assumption.

We then run through the Bayesian estimation for p parameters. It can be shown that $\boldsymbol{\beta}$ and σ^2 follow standard conjugacy result, but ρ needs extra care. The sampling algorithm draws $\boldsymbol{\beta}$ and σ^2 via Gibbs and ρ via rejection.

Coupling the analytic expression with r retained Markov chain Monte Carlo (MCMC) draws, we end up getting an $r \times n$ matrix. Each element calculates the case influence statistic given a set of MCMC parameters draw with single spatial observation removed.

1.2 Simulation design

We now perform a simulation exercise to examine the efficacy in identifying influential SAR observations. Thomas et al. (2018) includes artificial and empirical examples for models exhibiting conditional independence and Markovian dependence. The similar procedure is applied here, but under a SAR dependence structure. Results are compared and contrasted.

Again, our goal is to detect abnormal data points which do not agreed with the assumed model. Therefore, we consider the following:

1. Single outlier and conditional independence
2. Single outlier and SAR dependence
   (a) Examine the effect of changing SAR parameter
   (b) Examine the effect of changing neighbourhood structure

Note that the first case is a straightforward iteration of Thomas et al. (2018), which serves as the baseline model for comparison purpose. Since the degree of SAR dependence is governed by the SAR parameter ρ (which is the regression coefficient on the heterogenous spatial lag $\mathbf{Wy}$) and the spatial weight matrix $\mathbf{W}$, changing one of the SAR parameter and neighbourhood structure examines the performance of our methodology.

1.3 Bayesian SAR(1) model

$$\begin{aligned} \mathbf{y} &= \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad -1 < \rho < 1 \\ \boldsymbol{\epsilon} &\sim MVN\left(\mathbf{0}, \sigma^2\mathbf{I}\right) \quad \sigma^2 > 0 \end{aligned}$$

$$\begin{aligned} \boldsymbol{\beta} &\sim MVN\left(\mathbf{0}, 1000\mathbf{I}\right) \\ \sigma^2 &\sim IG\left(0.001, 0.001\right) \end{aligned}$$

Table 1: Larger δ and ρ produce higher leverage data and stronger SAR dependence, respectively

$\delta \backslash \rho$	0	0.2	0.4	0.6	0.8
1	★				
3					
5	★		☆		☆
7					
9	★				

1.4 Case influence statistic

$$\mathbf{W}_{i,j} = \log p\left(\mathbf{y}_{-j} \mid \boldsymbol{\theta}^{(i)}\right) - \log p\left(\mathbf{y} \mid \boldsymbol{\theta}^{(i)}\right) \quad \boldsymbol{\theta} = \left(\rho, \boldsymbol{\beta}, \sigma^2\right)^T$$

1.5 Cases 1, 2(a) and 2(b)

See Table 1.

$$\mathbf{y} = (y_1, y_2, \ldots, y_{50})^T$$

$$y_{23} = y_{23} + \delta\sigma \quad \sigma = \sqrt{45}$$

Blue ones represent spatial inflow from $\mathbf{y}_{-23}$ to y_{23}, and purple ones represent spatial outflow from y_{23} to $\mathbf{y}_{-23}$. Note that, we row standardised $\mathbf{B}_1$ and $\mathbf{B}_2$ to get the respective spatial weight matrices $\mathbf{W}_1$ and $\mathbf{W}_2$.

$$\mathbf{B}_1 = \begin{bmatrix} & 1 & & \\ 1 & & \ddots & \\ & \ddots & & 1 \\ & & 1 & \end{bmatrix}$$

$$\mathbf{B}_2 = \begin{bmatrix} & & & 1 & & & \\ \ddots & \ddots & \ddots & 1 & & & \\ & 1 & 0 & 1 & & & \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ & & & 1 & 0 & 1 & \\ & & & 1 & \ddots & \ddots & \ddots \\ & & & 1 & & & \end{bmatrix}$$

Math background

PCA

Data

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}$$

Covariance

$$s_{j,k} = \frac{1}{n} \sum_{i=1}^{n} \left(x_{i,j} - \bar{x}_j \right) \left(x_{i,k} - \bar{x}_k \right) \quad j,k = 1,2,\ldots,p$$

Correlation

$$r_{j,k} = \frac{\sum_{i=1}^{n} \left(x_{i,j} - \bar{x}_j \right) \left(x_{i,k} - \bar{x}_k \right)}{n} \Big/ s_j s_k$$

Eigendecomposition

$$\mathbf{A} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^{-1}$$

Eigenvalue

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_p \end{bmatrix}$$

Eigenvector (`rotation` in `prcomp` and `loadings` in `princomp`)

$$\mathbf{V} = \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,p} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ v_{p,1} & v_{p,2} & \cdots & v_{p,p} \end{bmatrix}$$

Reduced data (`x` in `prcomp` and `scores` in `princomp`)

$$\widetilde{\mathbf{X}} = \mathbf{X} \mathbf{V}$$

sPCA

Spatial data

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Spatial weight

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & w_{n,2} & \cdots & w_{n,n} \end{bmatrix}$$

Spatial correlation (I and *LISA*)

$$\begin{aligned} I &= \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \left(x_i - \bar{x}\right) \left(x_j - \bar{x}\right)}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}} \Bigg/ s^2 \\ &= \sum_{i=1}^{n} \frac{I_i}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}} \Bigg/ s^2 \end{aligned}$$

Spatial filtering

References

Thomas, Z. M., MacEachern, S. N. & Peruggia, M. (2018), 'Reconciling curvature and importance sampling based procedures for summarizing case influence in bayesian models', *Journal of the American Statistical Association* **113**(524), 1669–1683.