Inference based on the pivotal quantity and Monte Carlo simulation for spatial regression models

Young Eun Jeon[a] • Suk-Bok Kang[a] • Joon Jin Song[b] • Jung-In Seo[c]

[a] Department of Statistics, Yeungnam University, Gyeongsan, Korea
[b] Department of Statistical Science, Baylor University, Texas, USA
[c] Department of Information Statistics, Andong National University, Andong, Korea

1. Abstract

Backgrounds

- The maximum likelihood method is the most widely used estimation method, but it has a fatal drawback in that it provides the approximate confidence intervals (CIs) which often fail to satisfy nominal levels for a small sample size.

Purpose

- This study aims to improve the performance of the interval inference method for model parameters, scale and regression parameters, in the simultaneous autoregressive (SAR) and conditional autoregressive (CAR) models when a sample size is not large enough.
- We propose novel interval inference approaches using the pivotal quantity and Monte Carlo simulation in the SAR and CAR models, which appropriately treat nuisance parameters.

Results and Conclusions

- The simulation study demonstrates that the proposed interval estimating methods construct more reasonable CIs than the interval inference method based on the maximum likelihood estimators (MLEs) under the considered nominal level when the number of areal units is small.
- Real data analysis reveals the applicability of the proposed method.

2. Model

❖ Let $\{Y(A_i): A_i \in (A_1, \dots, A_n)\}$ be a Gaussian random process where $\{A_1, \dots, A_n\}$ forms a lattice of D if $A_1 \cup A_2 \cup \cdots \cup A_n = D$ and $A_i \cap A_j = 0$ for all $i \neq j$.

▪ SAR Model

$$Y(A_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \rho \sum_{j=1}^{n} w_{ij}\left(Y(A_j) - \boldsymbol{x}_j'\boldsymbol{\beta}\right) + e_i,\ i = 1, \dots, n, \tag{1}$$

where $\boldsymbol{e} = (e_1, \dots, e_n)' \sim N_n(\boldsymbol{0}, \sigma^2)$, σ^2 is variance, $\boldsymbol{x}_i = (1, x_{i1}, \dots, x_{ip})'$ is a $(p+1)\times 1$ vector of explanatory variables, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ is a vector of unknown regression parameters to be estimated. In addition, ρ is a parameter associated with spatial dependence, and w_{ij} is a spatial weight given by

$$w_{ij} = \begin{cases} 1, & \text{if an areal unit } A_i \text{ shares a common edge or border with an areal unit } A_j, \\ 0, & \text{otherwise.} \end{cases}$$

▪ CAR Model

$$Y(A_i) \mid \boldsymbol{Y}(A_{(-i)}) \sim N\left(\boldsymbol{x}_i'\boldsymbol{\beta} + \rho \sum_{j=1}^{n} w_{ij}\left(Y(A_j) - \boldsymbol{x}_j'\boldsymbol{\beta}\right),\ \sigma^2\right),\ i = 1, \dots, n, \tag{2}$$

where $\boldsymbol{Y}(A_{(-i)}) = \{Y(A_j): j \neq i\}$ and σ^2 is condition variance.

▪ Unified Model

A unified model including joint distributions of both models (1) and (2) is

$$\boldsymbol{Y} \sim N_n\left(\boldsymbol{X\beta},\ \sigma^2\boldsymbol{\Sigma}_\rho\right), \tag{3}$$

where $\boldsymbol{Y} = (Y(A_1), \dots, Y(A_n))'$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)'$. In addition, $\boldsymbol{\Sigma}_\rho = (I_n - \rho\boldsymbol{W})^{-h}$, $\boldsymbol{W} = (w_{ij})_{n\times n}$, and h is a fixed positive constant such as $h = 1$ (CAR) and $h = 2$ (SAR).

3. Inference

▪ Classical Approach

The likelihood function of unknown parameters $(\sigma^2, \boldsymbol{\beta}, \rho)$ in the unified model (3) is

$$L(\sigma^2, \boldsymbol{\beta}, \rho; \boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}|\boldsymbol{\Sigma}_\rho|^{1/2}} \exp\left(-\frac{S_\rho^2}{2\sigma^2}\right),$$

where $S_\rho^2 = (\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{\Sigma}_\rho^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})$.

1) MLEs of σ^2 and $\boldsymbol{\beta}$

$$\hat{\sigma}^2(\rho) = \frac{\hat{S}_\rho^2}{n} \sim Gamma\left(\frac{n-p-1}{2}, \frac{n}{2\sigma^2}\right) \text{ and } \hat{\boldsymbol{\beta}}(\rho) = \left(\boldsymbol{X}'\boldsymbol{\Sigma}_\rho^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_\rho^{-1}\boldsymbol{y} \sim N_{p+1}\left(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}'\boldsymbol{\Sigma}_\rho^{-1}\boldsymbol{X}\right)^{-1}\right),$$

where $\hat{S}_\rho^2 = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\rho)\right)'\boldsymbol{\Sigma}_\rho^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\rho)\right)$.

2) Approximate CIs based on the MLEs (M-ACIs)

$$\hat{\sigma}^2(\rho) \pm z_{\alpha/2}\frac{\sigma^2}{n}\sqrt{2(n-p-1)} \text{ and } \hat{\boldsymbol{\beta}}(\rho) \pm z_{\alpha/2}\sigma\sqrt{\text{diag}\left\{\left(\boldsymbol{X}'\boldsymbol{\Sigma}_\rho^{-1}\boldsymbol{X}\right)^{-1}\right\}},$$

where z_α denotes the upper α percentile of $N(0,1)$.

▪ Novel Inference Approach

We propose novel interval inference approaches for σ^2 and $\boldsymbol{\beta}$ using the pivotal quantity and Monte Carlo simulation in the unified model (3).

1) Pivotal-based Exact CI (P-ECI) for σ^2

$$\left(\frac{\hat{S}_\rho^2}{\chi^2_{\alpha/2, n-p-1}}, \frac{\hat{S}_\rho^2}{\chi^2_{1-\alpha/2, n-p-1}}\right),$$

where $\chi^2_{\alpha, n-p-1}$ denotes the upper α percentile of χ^2 distribution with $n-p-1$ degrees of freedom.

2) P-ECI for $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}(\rho) \pm t_{\alpha/2, n-p-1}\sqrt{\hat{\sigma}_p^2(\rho)\text{diag}\left\{\left(\boldsymbol{X}'\boldsymbol{\Sigma}_\rho^{-1}\boldsymbol{X}\right)^{-1}\right\}},$$

where $t_{\alpha, n-p-1}$ denotes the upper α percentile of t distribution with $n-p-1$ degrees of freedom and $\hat{\sigma}_p^2(\rho) = \hat{S}_\rho^2/(n-p-1)$.

3) CI based on Pivotal and Mote Carlo Simulation (PMC-CI) for $\boldsymbol{\beta}$

(1) Generate t from the χ^2_{n-p-1} distribution, and compute $\sigma^{2*} = \hat{S}_\rho^2/t$.

(2) Generate $\boldsymbol{v} = (v_1, \dots, v_{p+1})'$ from $N_{p+1}\left(\boldsymbol{0}, \sigma^{2*}\left(\boldsymbol{X}'\boldsymbol{\Sigma}_\rho^{-1}\boldsymbol{X}\right)^{-1}\right)$.

(3) Repeat $N(\geq 10{,}000)$ times (1) and (2). Then, we have the PMC-CI for $\boldsymbol{\beta}(\rho)$, given by

$$\left(\hat{\boldsymbol{\beta}}(\rho) - \boldsymbol{v}_{([(1-\alpha/2)N])}, \hat{\boldsymbol{\beta}}(\rho) - \boldsymbol{v}_{([(\alpha/2)N])}\right), \tag{4}$$

where $\boldsymbol{v}_{([\alpha N])}$ is the $[\alpha N]$th smallest of $\{\boldsymbol{v}_i\}$.

To treat the nuisance parameter ρ, integrated MLE $\hat{\rho}_I$ is obtained by maximizing the integrated likelihood function of ρ

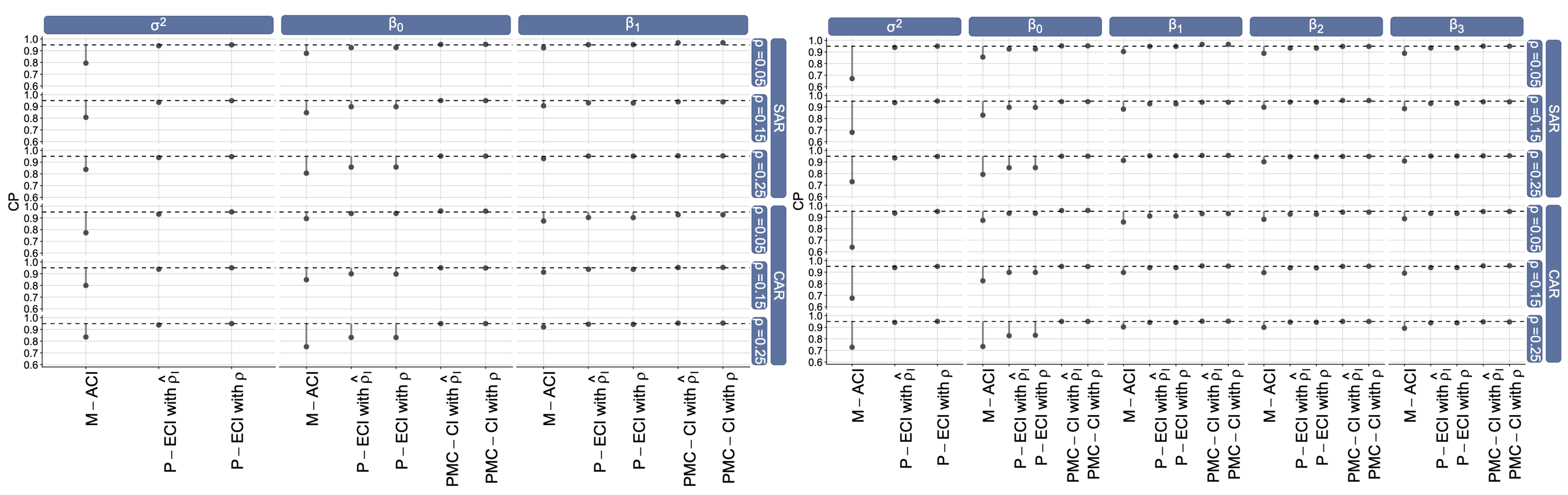
$$L^I(\rho; \boldsymbol{y}) = \int_{\mathbb{R}^{p+1}}\int_0^\infty L(\sigma^2, \boldsymbol{\beta}, \rho; \boldsymbol{y})\,\mathrm{d}\sigma^2\mathrm{d}\boldsymbol{\beta}$$

$$\propto |\boldsymbol{\Sigma}_\rho^{-1}|^{1/2}|\boldsymbol{X}'\boldsymbol{\Sigma}_\rho^{-1}\boldsymbol{X}|^{-1/2}\left(\hat{S}_\rho^2\right)^{-(n-p-1)/2+1}.$$

4. Simulation and Case Study

▪ Simulation Study

- We consider a 5×5 regular lattice with a first order neighborhood system to demonstrate the excellence of the proposed methods in a situation where the sample size is not large enough.
- The true values of unknown parameters $(\sigma^2, \boldsymbol{\beta}, \rho)$ are assigned as follows : $\sigma^2 = 1, \boldsymbol{\beta} = \{10, 1\}$ or $\{10, 1, 2, 4\}$, and $\rho = 0.05(0.1)0.25$.
- The coverage probabilities (CPs) are computed using 5,000 simulated datasets.
- The PMC-CI (4) is obtained based on $N = 15{,}000$.



▪ COVID-19 in Seoul

- For practical illustration, the district-level COVID-19 data in Seoul between 02 April 2022 and 08 April 2022 is employed (Data Source : https://data.seoul.go.kr/).
- This analysis aims to discern potential risk factors that contribute to the transmission of COVID-19.

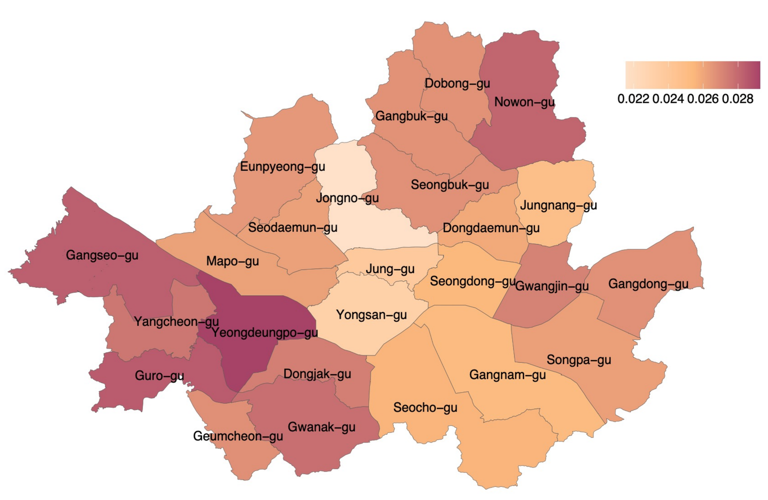


Fig.1. Map for district-level confirmed rates of COVID-19 in Seoul between 02 April 2022 and 08 April 2022

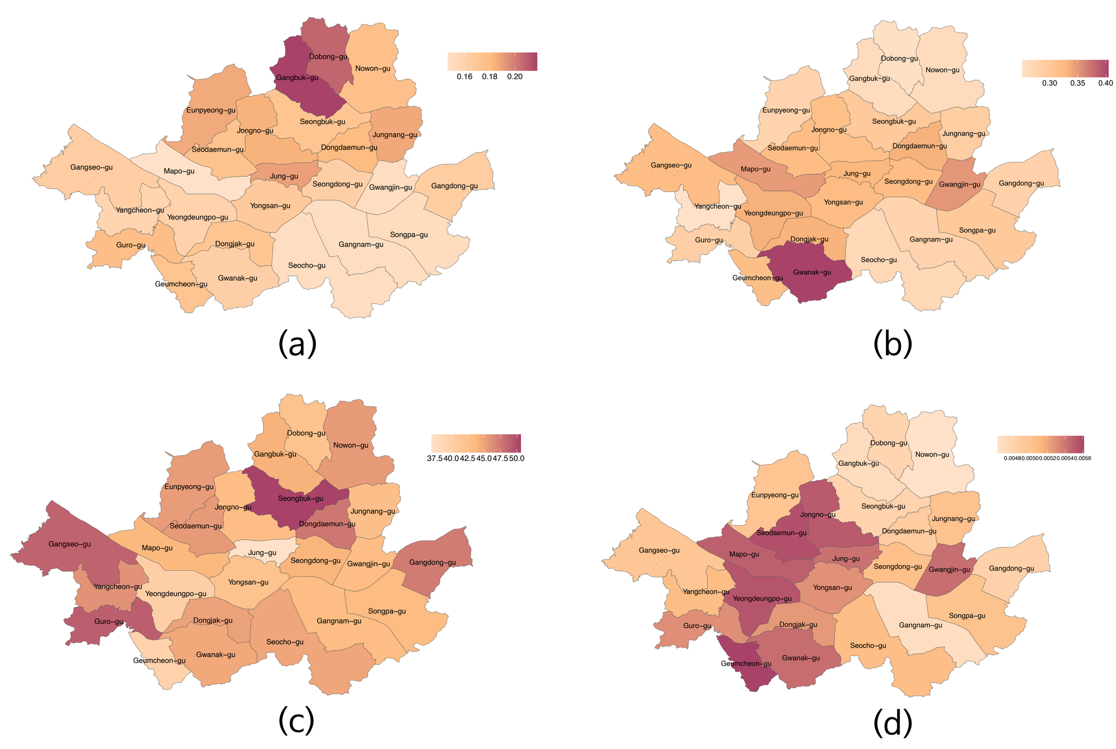


Fig.2. Map for district-level explanatory variables considered
(a) Elder.rate (b) Younger.rate (c) PM10 (d) Vaccine3.rate

Table 1. Inference results for real data

	$\hat{\rho}$	$\hat{\rho}_I$
SAR	0.130	0.135
CAR	0.194	0.193

	Estimate		95% CI	
	$\hat{\sigma}^2(\hat{\rho})$	$\hat{\sigma}_p^2(\hat{\rho}_I)$	M-ACI	P-ECI with $\hat{\rho}_I$
SAR	0.003	0.004	(0.001, 0.004)	(0.002, 0.008)
CAR	0.003	0.003	(0.001, 0.004)	(0.002, 0.007)

		Estimate		95% CI		
		With $\hat{\rho}$	With $\hat{\rho}_I$	M-ACI	P-ECI with $\hat{\rho}_I$	PMC-CI with $\hat{\rho}_I$
β_0	SAR	−3.596	−3.594	(−3.643, −3.548)	(−3.653, −3.536)	(−3.652, −3.536)
	CAR	−3.571	−3.572	(−3.628, −3.514)	(−3.639, −3.505)	(−3.638, −3.505)
Elder.rate	SAR	−0.012	−0.012	(−0.039, 0.015)	(−0.045, 0.020)	(−0.044, 0.020)
	CAR	−0.013	−0.013	(−0.040, 0.013)	(−0.045, 0.018)	(−0.044, 0.018)
Younger.rate	SAR	0.013	0.014	(−0.016, 0.043)	(−0.021, 0.049)	(−0.021, 0.049)
	CAR	0.018	0.018	(−0.011, 0.048)	(−0.017, 0.053)	(−0.017, 0.054)
PM10	SAR	0.011	0.011	(−0.013, 0.036)	(−0.018, 0.040)	(−0.018, 0.040)
	CAR	0.010	0.011	(−0.014, 0.035)	(−0.018, 0.039)	(−0.018, 0.039)
Vaccine3.rate	SAR	−0.017	−0.017	(−0.052, 0.019)	(−0.059, 0.025)	(−0.060, 0.026)
	CAR	−0.016	−0.016	(−0.051, 0.018)	(−0.057, 0.025)	(−0.057, 0.025)

5. Conclusions

- In the simulation study, the P-ECI and PMC-CI with $\hat{\rho}_I$ have CPs much closer to 0.95 than M-ACI.
  ⇒ **The proposed interval estimating methods construct more reasonable CIs than the interval inference method based on the MLEs under the considered nominal level when the number of areal units is small.**
- In the real data analysis, we analyzed the confirmed rates of COVID-19 in 25 districts of Seoul, and the analysis results revealed the validity and applicability of the proposed methods for real data.
- The proposed methods not only are applicable to other spatial models such as spatial lag and spatial durbin models, but also extend to spatio-temporal models.