Supplementary to “Spatial Homogeneity Pursuit of Regression Coefficients for Large Datasets ”

Furong Li
School of Mathematical Sciences, Ocean University of China
and
Huiyan Sang
Department of Statistics, Texas A&M University

1 Supplementary Proof of Theorem 1

To prove Theorem 1, we first derive the following two lemmas.

<u>Lemma</u> 1 *Define* $\Lambda_n = \{\underset{\ell=1,\ldots np}{max} |V_\ell| \leqslant \lambda_n/4\}$ *where* $V_\ell = n^{-1}\sum_{i=1}^{n} \widetilde{X}_{i,\ell}\varepsilon_i$*. Then*

$$P(\Lambda_n) \geqslant 1 - 2p \cdot n^{-C_2}$$

<u>Lemma</u> 2 *On the event* Λ_n*, we have*

$$\frac{1}{n}\|\widetilde{\mathbf{X}}\boldsymbol{\theta} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}\|_2^2 + \frac{\lambda_n}{2}\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1 \leq 2\lambda_n\|\boldsymbol{\theta}_A - \widehat{\boldsymbol{\theta}}_A\|_1$$

Proof 1 *According to Assumption 1a,* V_ℓ *is a sub-Gaussian random variable with a zero mean and a sub-Gaussian parameter* $C_1\sigma/\sqrt{n}$*. Using the upper and lower deviation inequalities, we have:*

$$P(|V_\ell| \leq \lambda_n/4) \geq 1 - 2\exp(-\frac{\lambda_n^2/16}{2C_1^2\sigma^2/n}) \geq 1 - 2n^{-(1+C_2)}$$

and

$$P(\max_{\ell=1,\dots np} |V_\ell| \le \lambda_n/4) \ge (1 - 2n^{-(1+C_2)})^{np} \ge 1 - 2p \cdot n^{-C_2}$$

This proves Lemma 1.

Proof 2 *As the estimator is the minimizer of penalized least square (5), we have*

$$\frac{1}{n}\|\mathbf{y} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}\|_2^2 + \lambda_n \|\widehat{\boldsymbol{\theta}}_B\|_1 \le \frac{1}{n}\|\mathbf{y} - \widetilde{\mathbf{X}}\boldsymbol{\theta}\|_2^2 + \lambda_n\|\boldsymbol{\theta}_B\|_1$$

After some manipulations, we have:

$$\frac{1}{n}\|\widetilde{\mathbf{X}}\boldsymbol{\theta} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}\|_2^2 \le \lambda_n\|\boldsymbol{\theta}_B\|_1 - \lambda_n\|\widehat{\boldsymbol{\theta}}_B\|_1 + \frac{2}{n}\boldsymbol{\epsilon}^\top\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

Then on the event Λ_n*, we have*

$$\frac{1}{n}\|\widetilde{\mathbf{X}}\boldsymbol{\theta} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}\|_2^2 \le \lambda_n\|\boldsymbol{\theta}_B\|_1 - \lambda_n\|\widehat{\boldsymbol{\theta}}_B\|_1 + \frac{\lambda_n}{2}\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1$$

It follows that

$$\begin{aligned}\frac{1}{n}\|\widetilde{\mathbf{X}}\boldsymbol{\theta} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}\|_2^2 + \frac{\lambda_n}{2}\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1 &\le \lambda_n\|\boldsymbol{\theta}_B\|_1 - \lambda_n\|\widehat{\boldsymbol{\theta}}_B\|_1 + \lambda_n\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1 \\ &= \lambda_n\|\boldsymbol{\theta}_B\|_1 - \lambda_n\|\widehat{\boldsymbol{\theta}}_B\|_1 + \lambda_n\|\boldsymbol{\theta}_B - \widehat{\boldsymbol{\theta}}_B\|_1 \\ &+ \lambda_n\|\boldsymbol{\theta}_{B^c} - \widehat{\boldsymbol{\theta}}_{B^c}\|_1 \\ &\le 2\lambda_n\|\boldsymbol{\theta}_{A-B^c} - \widehat{\boldsymbol{\theta}}_{A-B^c}\|_1 + \lambda_n\|\boldsymbol{\theta}_{B^c} - \widehat{\boldsymbol{\theta}}_{B^c}\|_1 \\ &\le 2\lambda_n\|\boldsymbol{\theta}_A - \widehat{\boldsymbol{\theta}}_A\|_1\end{aligned}$$

This proves Lemma 2.

Proof 3 *Return to the proof of Theorem 1. From Lemma 2, we have:*

$$\|\boldsymbol{\theta}_{A^c} - \widehat{\boldsymbol{\theta}}_{A^c}\|_1 \le 3\|\boldsymbol{\theta}_A - \widehat{\boldsymbol{\theta}}_A\|_1$$

indicating that $\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}$ *is a vector satisfying the condition in Assumption 1b. Therefore, we have*

$$\frac{1}{n}\|\widetilde{\mathbf{X}}\boldsymbol{\theta} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}\|_2^2 \geq \Phi\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2^2 \tag{S.1}$$

From Lemma 2, we also have

$$\frac{1}{n}\|\widetilde{\mathbf{X}}\boldsymbol{\theta} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}\|_2^2 \leq \frac{3}{2}\lambda_n\|\boldsymbol{\theta}_A - \widehat{\boldsymbol{\theta}}_A\|_1 \leq \frac{3}{2}\lambda_n\sqrt{|A|}\|\boldsymbol{\theta}_A - \widehat{\boldsymbol{\theta}}_A\|_2 \leq \frac{3}{2}\lambda_n\sqrt{|A|}\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2 \tag{S.2}$$

Combining (S.1) *and* (S.2) *yields:*

$$\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2 \leq \frac{3\lambda_n\sqrt{|A|}}{2\Phi}$$

This proves (8) in Theorem 1. Based on (8) and (S.2)*, we have*

$$\frac{1}{n}\|\widetilde{\mathbf{X}}\boldsymbol{\theta} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}\|_2^2 \leq \frac{9\lambda_n^2|A|}{4\Phi}$$

This proves (7) in Theorem 1.

2 Supplementary Numerical Results

(a): Selection criteria for the tuning parameter λ

For simulation study presented in Section 3, we compare the performance of SCC-MST using both AIC and BIC (denoted as SCC-AIC and SCC-BIC, respectively). Since the true values of $\boldsymbol{\beta}$ are known, we also include the minimum mean square error (MSE), which corresponds to the “optimal” λ as a benchmark (denoted as SCC-Min). As shown in Figure 1, Both AIC and BIC select values of λ that produce fairly close MSE values to that of the SCC-Min.

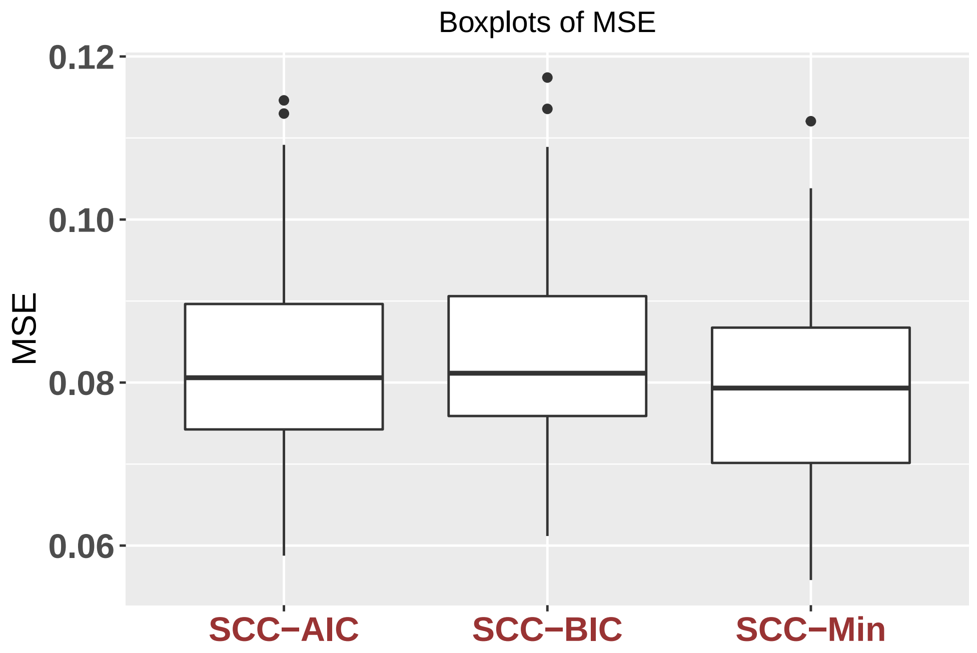

Figure 1: Study (a): Boxplots of MSE_β for the SCC model using AIC (denoted as SCC-AIC) and BIC (denoted as SCC-BIC). The boxplot of the minimum MSE_β of SCC (denoted as SCC-Min) is also included as a benchmark.

(b): Comparisons of the MSE_β derived from GWR using different bandwidths

For the simulation study in Section 3, the competing method GWR involves the specification for the tuning parameter bandwidth. We use the bandwidth selected by cross-validation for the GWR model. Below, we test whether increasing the bandwidth can improve the performance of GWR method. We examined the performance of GWR using the bandwidth obtained from the cross validation method multiplied by a factor of c for a range of c values. The boxplots of MSE_β with various values of c from 100 simulations are shown in Figure 2. The results indicate that the mean MSE_β over 100 simulations first decreases as c increases, and then increases again as c increases. The smallest mean MSE_β, 0.283, is achieved around $c = 1.2$. The reduction in mean MSE_β is about 21% from that of the model using bandwidth selected by cross-validation (i.e., $c = 1$), suggesting that the cross validation method works reasonably well. For the values of c that we tested, the mean MSE_β derived from the GWR method are all much larger than the value 0.087 derived from the SCC method.

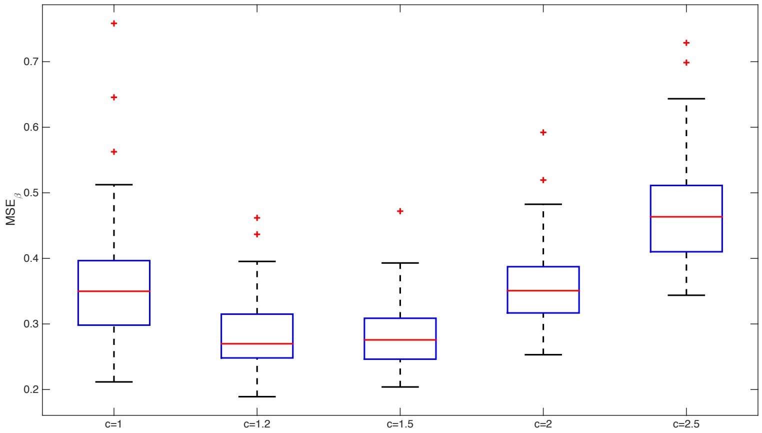

Figure 2: Study (b): Boxplots of MSE_β derived from the GWR method using different values of c ($c = 1$ corresponding to the bandwidth determined from the cross validation). Here the true $\boldsymbol{\beta}$ is clusterly distributed and the explanatory variables are associated with weak spatial correlation.

(c): Performance of SCC on regular grids

To investigate the performance of the SCC method on regular grids, we generate data at regular lattices from $[0, 1]\times[0, 1]$ with different resolutions, ranging from $n = 20^2, 30^2, 40^2, \ldots, 80^2$. We consider a model with only one varying intercept $\beta_0(\mathbf{s})$ for simplicity shown as in Figure 3, and generate y from the model $y(\mathbf{s}) = \beta_0(\mathbf{s}) + \epsilon(\mathbf{s})$ at each location, where $\epsilon(\mathbf{s})$ are white noises with standard deviation to be 0.1. We implement both SCC and GWR to estimate $\beta_0(\mathbf{s})$ and calculate the corresponding MSE_β. As can be seen from Figure 4, the SCC method produces much smaller MSE_β than that of the GWR method, similar as the case with irregular points. In addition, it is noticeable that MSE_β for the SCC method is decreasing as n increases.

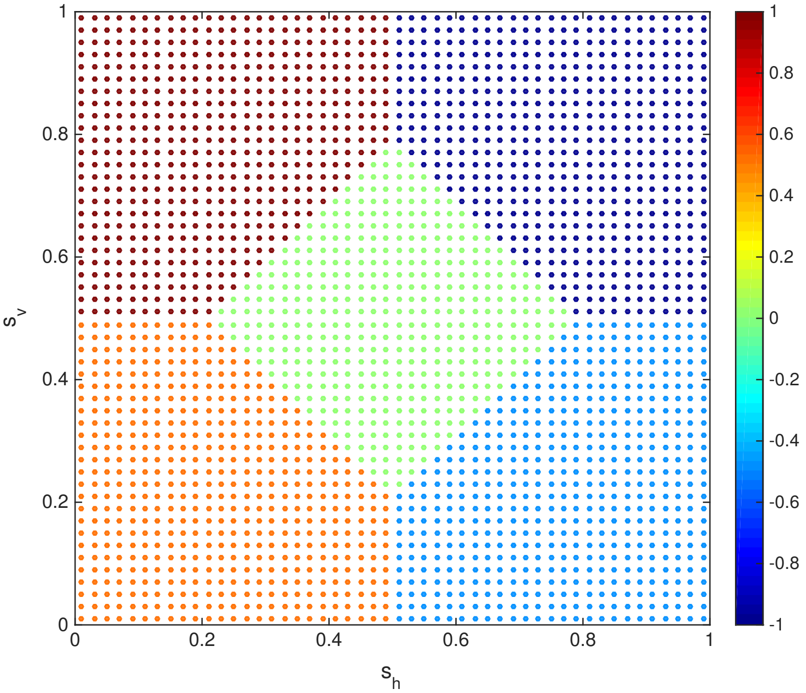

Figure 3: Study (c): The spatial distribution of β_0 on regular grids.

(d): Performance of SCC in a hybrid scenario

This study considers the case where $\beta_1(s_i)$ and $\beta_2(s_i)$ are spatially clustered but the intercept $\beta_3(s_i)$ is smoothly varying. Here the spatial structures of $\beta_1(s_i)$ and $\beta_2(s_i)$ are the same as those in Study 1, but $\beta_3(s_i)$ is generated from a Gaussian spatial process with a zero mean and an isotropic exponential covariance function. The variance parameter is chosen to be 0.5, with $\phi = 0.3$. It should be noted that $\beta_3(s_i)$ can be interpreted as the spatial random effect, as in SVC.

If the covariance matrix of $\beta_3(s_i) + \epsilon(s_i)$ is known, then, by multiplying the inverse of its Cholesky matrix with each term in (9), the model is transformed into a spatial regression problem with identically independently distributed errors and hence can be solved using our proposed SCC method. We provide the results from this method (the "transformed SCC", or SCC-T, henceforth) as a benchmark for comparison. It should be noted that the covariance matrix of $\beta_3(s_i)$ and $\epsilon(s_i)$ is unknown in practice.

The estimated values obtained from GWR, SCC, and SCC-T in one simulation are illustrated in Figure 5. As the SCC-T method treats $\beta_3(s_i)$ as a spatial random effect and thus does not estimate its value at individual points, we only show the estimates of $\beta_1(s_i)$ and $\beta_2(s_i)$. The coefficients obtained from SCC and SCC-T are similar to each other. Both agree reasonably well with the true model. In contrast, the estimates obtained from GWR are quite noisy and fail to reproduce the clustered patterns that exist in the true model.

Table 1: Summary of Study (d): the mean MSE_β for the GWR, SCC and SCC-T methods over 100 simulations, under various spatial correlations for predictors.

Spatial correlation	MSE_β		
	GWR	SCC	SCC-T
Weak	0.27	0.15	0.11
Moderate	0.74	0.22	0.20
Strong	2.47	0.38	0.34

Comparisons of the values of MSE_β further demonstrate the advantage of SCC and SCC-T, especially in the presence of strong spatial correlation among the predictors (Figure 6 and Table 1). Moreover, consistent with the case study shown in Figure 5, the estimation performances of SCC and SCC-T are comparable, as the MSE_β values derived from the two methods are fairly close. Note that SCC-T exhibits an ideal case in which the covariance of the spatial random effect is fully accounted for. The comparison thus suggests that the performance of SCC is robust not only in cases with spatially clustered random effects, but also those with spatially smoothly varying random effects.

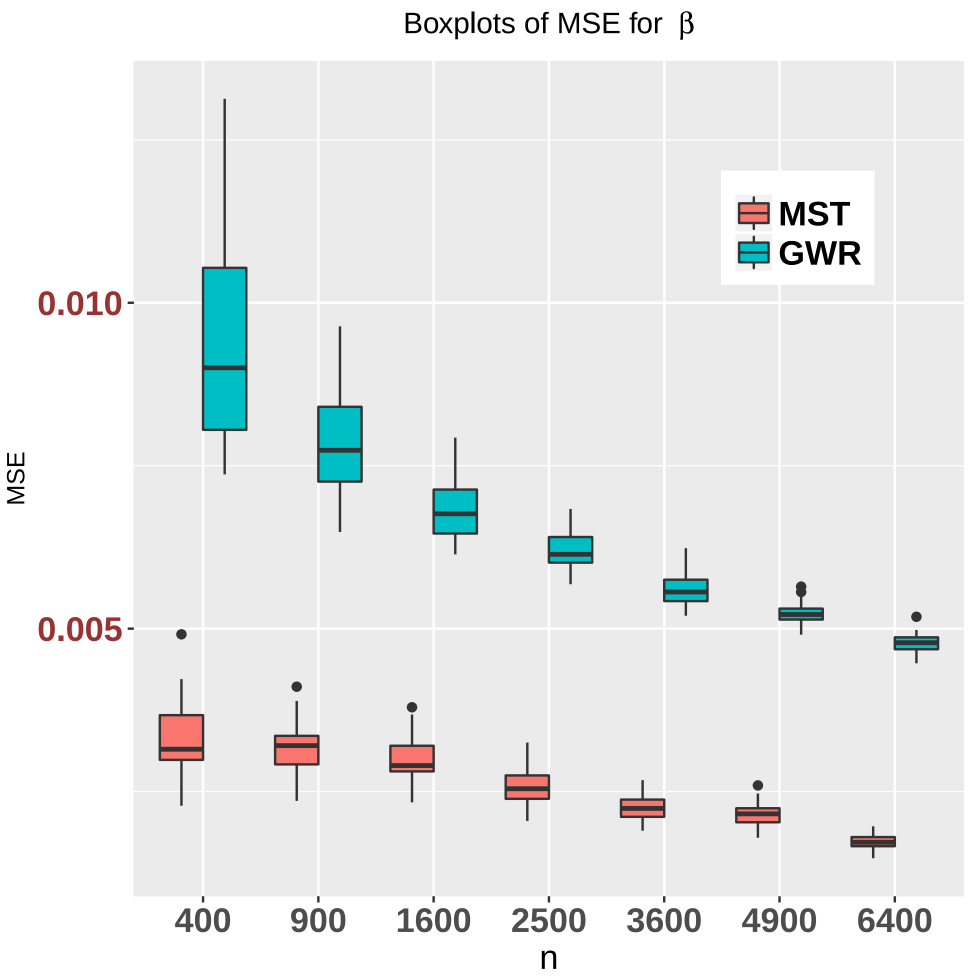

Figure 4: Study (c): Boxplots of MSE_β for SCC and GWR with a varying number of n locations on regular grids.

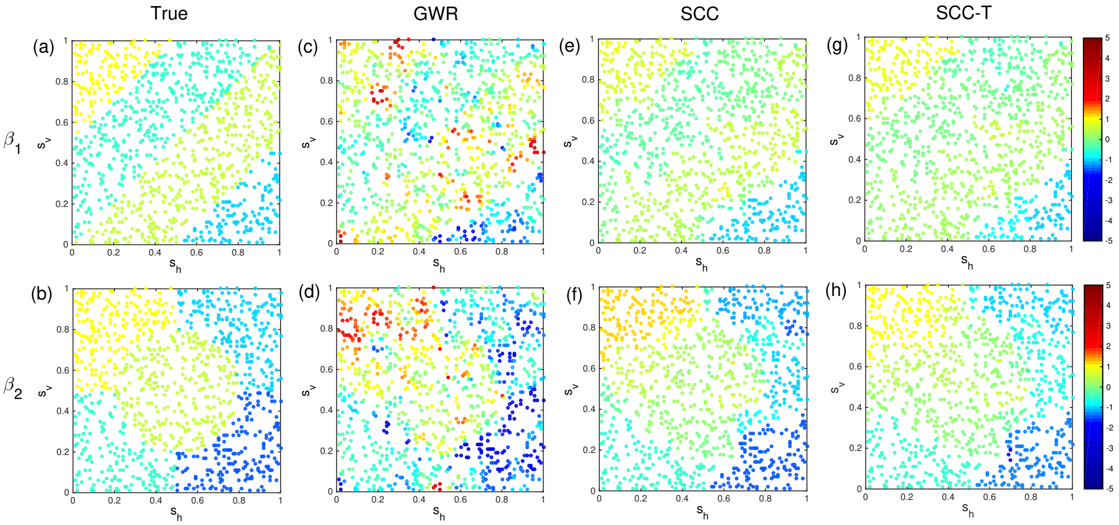

Figure 5: Study (d): spatial structures of (a,b) true coefficient β_1 and β_2. The estimated coefficient surfaces from the (c,d) GWR method, (e,f) SCC method, and (g,h) SCC-T method in one simulation with the spatial range parameter $\phi = 0.3$ for predictors.

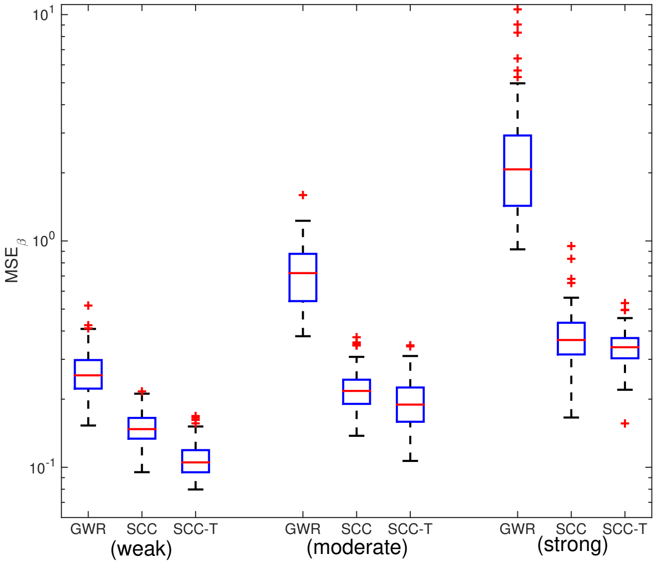

Figure 6: Study (d): the boxplots of MSE_β for the GWR, SCC and SCC-T methods under 3 setting of spatial correlations (weak, moderate and strong) for predictors, based on 100 simulated datasets. Here only β_1 and β_2 are included in computing MSE_β.