Supplement to “Spatial Variable Selection and An Application to Virginia Lyme Disease Emergence”

Yimeng Xie[1], Li Xu[1], Jie Li[1], Xinwei Deng[1], Yili Hong[1],
Korine Kolivras[2], and David N. Gaines[3]

[1]Department of Statistics, Virginia Tech, Blacksburg, VA 24061
[2]Department of Geography, Virginia Tech, Blacksburg, VA, 24061
[3]Virginia Department of Health, Richmond, VA, 23219

1 Additional Plots for Data Analysis

Figure 1 shows the histogram of pairwise correlations among all the covariates in the Virginia Lyme disease data. Figure 2 shows the plot of the observed versus expected number of counts for census tracts for the data.

2 Bootstrap Algorithm for Confidence Intervals

Here we describe the parametric bootstrap algorithm used in Section 4.4 of the paper.

Algorithm: Parametric Bootstrap for Confidence Intervals

1. Simulate $\boldsymbol{b}^* = (b_1^*, \ldots, b_i^*, \ldots, b_n^*)'$ from $\mathrm{N}(\mathbf{0}, \Sigma_{\widehat{\boldsymbol{\theta}}})$.

2. Simulate y_i^* from Poisson$\{\exp[\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + b_i^* + \log(m_i)]\}$, for $i = 1, \ldots, n$.

3. Apply the variable selection procedure in **Algorithm 1** or **2** to the samples generated in step 2 to obtain $\widehat{\boldsymbol{\beta}}^*$ and $\widehat{\boldsymbol{\theta}}^*$.

4. Repeat steps 1 to 3 B times to obtain B sets of bootstrap parameter estimates for inference.

5. Let θ be a general notation for any unknown quantity that is of interest and $\widehat{\theta}$ be the estimate. Sort the B bootstrap estimates $\widehat{\theta}^{*1}, \ldots, \widehat{\theta}^{*B}$ in increasing order and obtain $\widehat{\theta}^{*(b)}, b = 1, \ldots, B$.

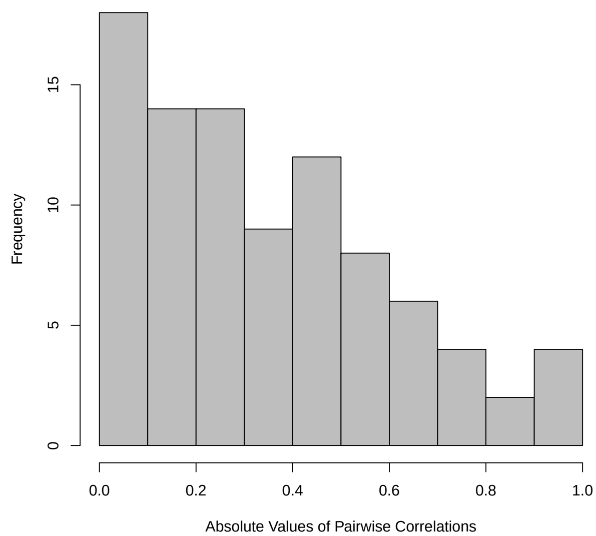

Figure 1: Histogram of pairwise correlations among all the covariates in the Lyme disease data.

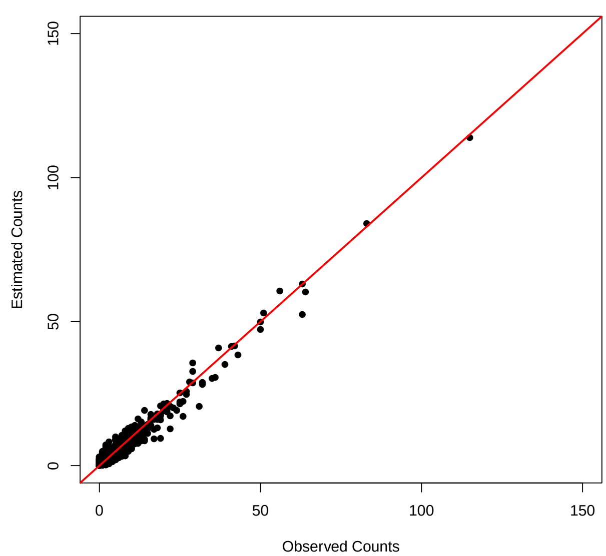

Figure 2: Plot of the observed versus expected number of counts for census tracts for the Virginia Lyme disease data. The 45-degree line is for reference.

Table 1: Computing time for one trial corresponding to the scenarios in Table 6 of the paper (Intel Core i7-860 CPU with 2.80 GHz and 8 GB RAM). The time unit is in seconds.

Method	Cases	Consider spatial correlation	Ignore spatial correlation
APL.AEN	Case 1	313.51	236.00
	Case 2	1715.21	1072.73
	Case 3	434.59	326.83
	Case 4	413.73	299.38
	Case 5	524.79	239.61
PQL.AEN	Case 1	53.00	9.47
	Case 2	26.02	10.28
	Case 3	73.11	9.61
	Case 4	40.09	8.59
	Case 5	100.52	12.13

6. We use bias-corrected bootstrap confidence interval (CI) procedure to obtain a CI for θ. The lower and upper bounds of the approximate $100(1-\alpha)\%$ CI for θ is

$$\left[\widehat{\theta}^{*([l])}, \quad \widehat{\theta}^{*([u])}\right]$$

where $l = B\Phi_{\text{nor}}\left(2z_q + z_{\alpha/2}\right)$ and $u = B\Phi_{\text{nor}}\left(2z_q + z_{1-\alpha/2}\right)$. Here $z_p = \Phi_{\text{nor}}^{-1}(p)$ is the p quantile of the standard normal distribution, q is the proportion of the B values of $\widehat{\theta}^*$ that are less than $\widehat{\theta}$, and $[\cdot]$ is the round function.

3 Computing Time for Simulations

Table 1 shows the computing time for one trial corresponding to the scenarios in Table 6 of the paper (Intel Core i7-860 CPU with 2.80 GHz and 8 GB RAM). The time unit is in seconds.

4 Additional Comparisons with Existing Methods

Tables 2-5 show the comparisons with existing methods for model selection results based on simulated samples. The setting is the same as in Tables 2-5 of the paper, respectively. The metrics used are *aver.size* (AS), *corr.coef* (CC), and *mis.coef* (MC). The existing methods under consideration are P-value-based and backward selection methods, and the GLMM Lasso method.

Table 2: Comparisons with existing methods for model selection results based on simulated samples. The setting is the same as in Table 2 of the paper. The metrics used are *aver.size* (AS), *corr.coef* (CC), and *mis.coef* (MC).

Cases	P-value-based			Backward			glmmLasso		
	AS	CC	MC	AS	CC	MC	AS	CC	MC
True value	5	10	0	5	10	0	5	10	0
Case 1	5.80	9.20	0.00	5.70	9.30	0.00	5.52	8.64	0.84
Case 2	5.56	9.44	0.00	5.59	9.41	0.00	2.90	9.84	2.26
Case 3	5.75	9.25	0.00	5.76	9.24	0.00	6.22	8.11	0.67
Case 4	5.83	9.17	0.00	5.91	9.09	0.00	5.99	8.25	0.76
Case 5	5.52	9.47	0.01	5.69	9.30	0.00	3.00	9.77	2.23

Table 3: Comparisons with existing methods for model selection results based on simulated samples. The setting is the same as in Table 3 of the paper. The metrics used are *aver.size* (AS), *corr.coef* (CC), and *mis.coef* (MC).

Cases	P-value-based			Backward			glmmLasso		
	AS	CC	MC	AS	CC	MC	AS	CC	MC
True value	5	10	0	5	10	0	5	10	0
Case 1	5.84	9.16	0.00	5.88	9.12	0.00	6.18	7.98	0.85
Case 2	5.60	9.29	0.11	5.56	9.33	0.11	3.42	9.50	2.09
Case 3	5.83	9.14	0.03	5.79	9.18	0.03	7.35	7.01	0.64
Case 4	5.75	9.22	0.03	5.63	9.32	0.04	7.00	7.31	0.69
Case 5	5.64	9.31	0.05	5.70	9.25	0.05	3.36	9.49	2.16

Table 4: Comparisons with existing methods for model selection results based on simulated samples. The setting is the same as in Table 4 of the paper. The metrics used are *aver.size* (AS), *corr.coef* (CC), and *mis.coef* (MC).

Cases	P-value-based			Backward			glmmLasso		
	AS	CC	MC	AS	CC	MC	AS	CC	MC
True value	5	10	0	5	10	0	5	10	0
Case 1	5.74	9.26	0.00	5.80	9.20	0.00	5.57	8.70	0.74
Case 2	5.53	9.47	0.00	5.54	9.46	0.00	2.76	9.90	2.34
Case 3	5.69	9.30	0.00	5.62	9.38	0.00	5.99	8.25	0.76
Case 4	5.70	9.30	0.00	5.68	9.32	0.00	6.03	8.22	0.75
Case 5	5.50	9.50	0.00	5.66	9.34	0.00	2.83	9.78	2.39

Table 5: Comparisons with existing methods for model selection results based on simulated samples. The setting is the same as in Table 5 of the paper. The metrics used are *aver.size* (AS), *corr.coef* (CC), and *mis.coef* (MC).

Cases	P-value-based			Backward			glmmLasso		
	AS	CC	MC	AS	CC	MC	AS	CC	MC
True value	10	10	0	10	10	0	10	10	0
Case 1	10.36	9.19	0.45	10.41	9.18	0.41	7.46	8.74	3.80
Case 2	10.17	9.07	0.76	10.21	9.10	0.69	10.07	7.72	2.21
Case 3	10.08	9.11	0.81	10.15	9.12	0.73	8.51	8.40	3.09
Case 4	9.95	9.18	0.87	10.07	9.13	0.79	8.20	8.53	3.26
Case 5	9.74	9.22	1.04	9.79	9.18	1.04	10.69	7.55	1.76