Simulation results for penalization methods for local variable selection in a logistic regression setting

Wesley Brooks

---

1. Simulation

1.1. Simulation Setup

Data were simulated on the spatial domain $[0,1]^2$, which was divided into a 30×30 grid. Each of $p = 5$ covariates $X_1, \ldots, X_5$ was simulated by a Gaussian random field (GRF) with mean zero and exponential spatial covariance $\text{Cov}\left(X_{ji}, X_{ji'}\right) = \sigma_x^2 \exp\left(-\tau_x^{-1}\delta_{ii'}\right)$ where $\sigma_x^2 = 1$ is the variance, $\tau_x = 0$ is the range parameter, and $\delta_{ii'}$ is the Euclidean distance $\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|_2$. Correlation was induced between the covariates by multiplying the $\boldsymbol{X}$ matrix by $\boldsymbol{R}$, where $\boldsymbol{R}$ is the Cholesky decomposition of the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{R}'\boldsymbol{R}$. The covariance matrix $\boldsymbol{\Sigma}$ is a 5×5 matrix that has ones on the diagonal and ρ for all off-diagonal entries, where ρ is the between-covariate correlation.

The simulated response was $y_i \sim \text{Binomial}(10, p_i)$ for $i = 1, \ldots, 900$ where $p_i = \exp(\eta_i)/(1+\exp(\eta_i))$ and $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}_i$. The simulated data included the response y and five covariates $x_1, \ldots, x_5$. The true data-generating model uses only x_1, so $x_2, \ldots, x_5$ are included to assess performance in variable-selection.

There were six simulation settings, each of which was simulated 100 times. For each of the twelve settings, $\beta_1(\boldsymbol{s})$, the true coefficient surface for x_1, was nonzero in at least part of the spatial domain $[0,1]^2$. There were four other simulated covariates, but their true coefficient surfaces were zero across the area under simulation. The twelve simulation settings are described in Table 1. Three

parameters were varied to produce the twelve settings: there were three functional forms for the coefficient surface $\beta_1(\boldsymbol{s})$, data was simulated both with ($\rho = 0.5$) and without ($\rho = 0$) correlation between the covariates.

The three coefficient surfaces used to produce the response variable in the simulations are pictured in Figure 1. The first is a "step" function, which is equal to zero in 40% of the spatial domain, equal to one in a different 40% of the spatial domain, and increases linearly in the middle 20% of the domain. The second is a gradient function, which increases linearly from zero at one end of the domain to one at the other. The final coefficient function is a parabola taking its maximum value of 0.535 at the center of the domain and falling to zero at each corner of the domain. The parabola is computed by finding the squared distance of each sampling location from the domain's center, multiplying by -1 and then adding an offset so that the corner points are equal to zero.

The performance of the penalized GWR methods (AL via `glmnet`, and the AEN via `enet`) was compared to that of oracular GWR (O-GWR), which is ordinary GWR with "oracular" variable selection, meaning that exactly the correct set of covariates was used to fit the GWR model at each location in the simulation. Also included in the comparison was the GWR algorithm of **?** without variable selection (`gwr`). Finally, there is a category of simulation results using the three penalized GWR methods for local variable selection and then ordinary GWR for coefficient estimation.

Results from the simulation were summarized at five locations on the simulated grid (see Figure 2). The five key locations were chosen because they represent interesting regions of the β_1 coefficient surfaces. The results of variable selection and coefficient estimation are presented in the tables below.

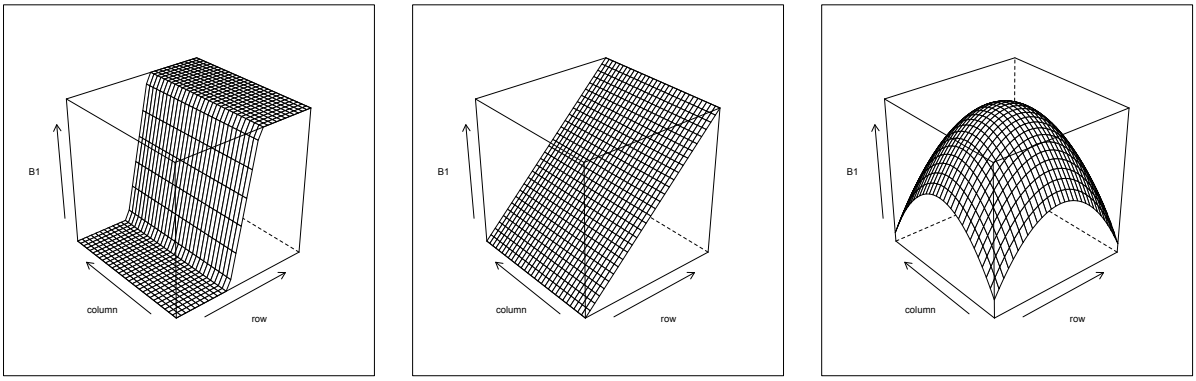

Figure 1: The actual β_1 coefficient surface used in the simulation.

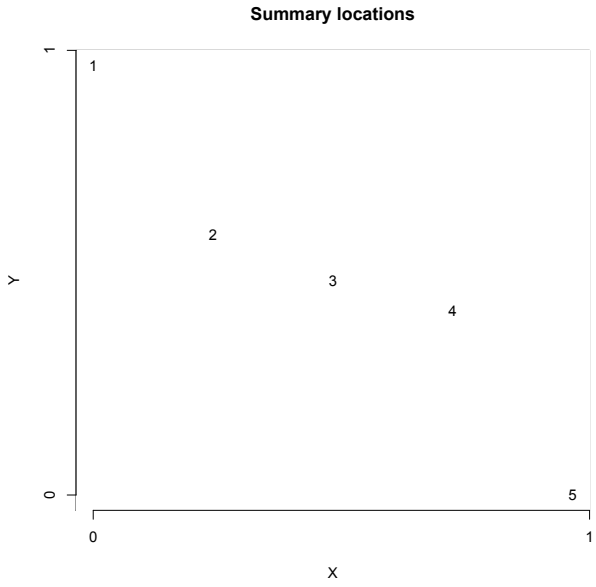

Figure 2: Locations where the variable selection and coefficient estimation of GWL were summarized.

Setting	function	ρ
1	step	0
4	step	0.5
5	gradient	0
8	gradient	0.5
10	parabola	0
11	parabola	0.5

Table 1: Simulation parameters for each setting.

	step				gradient				parabola			
	`enet`		`glmnet`		`enet`		`glmnet`		`enet`		`glmnet`	
location	β_1	β_2 - β_5	β_1	β_2 - β_5	β_1	β_2 - β_5	β_1	β_2 - β_5	β_1	β_2 - β_5	β_1	β_2 - β_5
1	1.00	0.04	1.00	0.05	0.99	0.10	0.96	0.08	1.00	0.03	1.00	0.02
	0.86	0.08	0.82	0.07	0.84	0.07	0.88	0.05	0.97	0.06	0.97	0.07
2	1.00	0.07	1.00	0.06	1.00	0.06	1.00	0.05	1.00	0.08	1.00	0.07
	1.00	0.06	1.00	0.06	1.00	0.07	0.99	0.04	0.98	0.08	0.99	0.07
3	0.99	0.06	0.99	0.06	0.97	0.08	0.92	0.04	1.00	0.08	1.00	0.07
	0.84	0.08	0.82	0.07	0.81	0.11	0.80	0.08	0.95	0.08	0.96	0.08
4	0.64	0.06	0.59	0.06	0.51	0.12	0.40	0.07	1.00	0.06	1.00	0.06
	0.48	0.07	0.49	0.07	0.52	0.07	0.51	0.07	0.95	0.07	0.93	0.06
5	0.03	0.03	0.03	0.03	0.02	0.03	0.03	0.05	0.93	0.05	0.94	0.04
	0.06	0.04	0.04	0.05	0.04	0.03	0.06	0.06	0.70	0.07	0.70	0.07

Table 2: Selection frequency for the simulation experiment

function	location	enet	glmnet	u.enet	u.glmnet	oracular	gwr
step	1	0.025	*0.023*	0.127	0.124	0.082	**0.005**
		0.186	0.216	0.376	0.375	*0.134*	**0.009**
	2	0.024	0.024	*0.021*	**0.021**	0.021	0.042
		0.063	0.068	*0.054*	0.056	**0.042**	0.070
	3	0.011	0.010	0.007	0.007	**0.004**	*0.005*
		0.043	0.047	0.049	0.054	*0.009*	**0.008**
	4	*0.014*	**0.014**	0.019	0.018	0.021	0.042
		0.036	*0.039*	0.042	0.046	0.047	0.074
	5	*0.001*	0.002	0.004	0.004	**0.000**	0.007
		0.006	*0.002*	0.024	0.009	**0.000**	0.011
gradient	1	*0.045*	0.073	0.134	0.205	0.101	**0.011**
		0.218	0.179	0.425	0.369	*0.154*	**0.022**
	2	0.027	0.021	0.021	**0.017**	*0.018*	0.044
		0.071	0.071	*0.056*	0.061	**0.043**	0.075
	3	0.014	0.022	0.011	0.021	**0.005**	*0.005*
		0.047	0.045	0.045	0.044	*0.008*	**0.008**
	4	*0.012*	**0.011**	0.016	0.014	0.020	0.044
		0.028	*0.038*	0.047	0.048	0.043	0.082
	5	*0.002*	0.003	0.009	0.009	**0.000**	0.010
		0.004	0.022	0.038	0.043	**0.000**	0.015
parabola	1	0.069	0.070	**0.007**	*0.007*	0.010	0.016
		0.094	0.096	0.078	0.085	*0.045*	**0.042**
	2	0.003	0.003	*0.001*	0.001	**0.001**	0.001
		0.013	0.008	0.013	0.009	*0.002*	**0.002**
	3	0.001	0.001	0.001	*0.001*	**0.001**	0.001
		0.017	0.015	0.019	0.017	*0.002*	**0.002**
	4	0.003	0.003	0.001	*0.001*	**0.001**	0.001
		0.014	0.016	0.012	0.015	**0.002**	*0.003*
	5	0.068	0.069	0.004	*0.004*	**0.000**	0.016
		0.051	0.052	0.019	*0.019*	**0.000**	0.044

Table 3: Mean squared error of $\hat{\beta}_1$ (**minimum**, *next best*).

function	location	enet	glmnet	u.enet	u.glmnet	oracular	gwr
step	1	-0.029	*-0.020*	0.038	0.033	0.034	**-0.004**
		-0.195	-0.211	-0.082	-0.075	*0.053*	**-0.017**
	2	-0.119	-0.119	*-0.110*	**-0.110**	-0.124	-0.196
		-0.178	-0.186	**-0.145**	*-0.150*	-0.175	-0.253
	3	*-0.014*	**-0.010**	0.017	0.015	0.021	0.040
		-0.027	-0.031	*0.009*	**0.004**	0.050	0.059
	4	*0.059*	**0.049**	0.074	0.065	0.129	0.196
		0.075	*0.076*	0.088	0.090	0.193	0.263
	5	*-0.006*	-0.006	-0.009	-0.010	**0.000**	-0.006
		-0.009	*-0.000*	-0.025	-0.008	**0.000**	-0.011
gradient	1	-0.077	-0.073	0.028	**-0.014**	0.050	*-0.017*
		-0.214	-0.167	-0.067	-0.068	*0.035*	**0.006**
	2	-0.130	*-0.099*	-0.103	**-0.083**	-0.110	-0.199
		-0.221	-0.216	**-0.167**	-0.184	*-0.182*	-0.263
	3	-0.056	-0.056	**-0.009**	-0.030	*0.017*	0.034
		-0.094	-0.077	-0.056	-0.056	**0.017**	*0.055*
	4	0.027	**0.010**	0.043	*0.020*	0.129	0.199
		0.073	*0.089*	0.105	0.105	0.189	0.275
	5	*-0.005*	-0.009	-0.009	-0.012	**0.000**	-0.009
		-0.011	-0.011	-0.036	-0.021	**0.000**	*-0.007*
parabola	1	-0.248	-0.253	**0.010**	0.011	*0.011*	-0.111
		-0.242	-0.248	*-0.014*	-0.022	**-0.007**	-0.182
	2	-0.047	-0.048	*0.002*	**0.001**	0.004	0.002
		-0.044	-0.035	**0.003**	0.011	*0.008*	-0.011
	3	0.005	0.005	*0.002*	**0.001**	0.003	0.002
		-0.017	-0.012	-0.013	-0.007	**0.003**	*0.006*
	4	0.043	0.045	*0.006*	0.007	**0.004**	0.008
		0.006	**0.002**	-0.014	-0.023	*0.004*	0.020
	5	0.249	0.253	*0.002*	0.003	**0.000**	0.113
		0.182	0.186	*-0.001*	0.004	**0.000**	0.187

Table 4: Bias of $\hat{\beta}_1$ (**minimum**, *next best*).

function	location	enet	glmnet	u.enet	u.glmnet	oracular	gwr
step	1	0.024	*0.023*	0.127	0.124	0.081	**0.005**
		0.149	0.173	0.373	0.373	*0.133*	**0.009**
	2	0.010	0.010	0.009	0.009	*0.006*	**0.003**
		0.032	0.034	0.033	0.034	*0.012*	**0.006**
	3	0.011	0.010	0.007	0.007	*0.004*	**0.003**
		0.043	0.047	0.050	0.055	*0.007*	**0.004**
	4	0.011	0.012	0.014	0.014	*0.004*	**0.003**
		0.030	0.033	0.035	0.038	*0.009*	**0.005**
	5	*0.001*	0.002	0.004	0.004	**0.000**	0.007
		0.006	*0.002*	0.024	0.009	**0.000**	0.011
gradient	1	*0.040*	0.068	0.134	0.207	0.099	**0.011**
		0.174	*0.153*	0.424	0.368	0.154	**0.022**
	2	0.011	0.012	0.010	0.010	*0.006*	**0.005**
		0.022	0.025	0.028	0.028	*0.010*	**0.006**
	3	0.011	0.019	0.011	0.020	*0.004*	**0.004**
		0.039	0.039	0.043	0.042	*0.008*	**0.005**
	4	0.011	0.011	0.014	0.013	**0.003**	*0.004*
		0.023	0.031	0.037	0.037	*0.007*	**0.006**
	5	*0.002*	0.003	0.009	0.009	**0.000**	0.010
		0.004	0.022	0.037	0.043	**0.000**	0.015
parabola	1	0.007	*0.006*	0.007	0.007	0.010	**0.004**
		0.035	0.035	0.079	0.085	0.046	**0.009**
	2	0.001	*0.001*	0.001	0.001	**0.001**	0.001
		0.011	0.007	0.013	0.009	*0.002*	**0.002**
	3	0.001	0.001	0.001	*0.001*	**0.001**	0.001
		0.017	0.015	0.019	0.017	*0.002*	**0.002**
	4	0.001	0.001	0.001	*0.001*	**0.001**	0.001
		0.014	0.017	0.012	0.014	**0.002**	*0.002*
	5	0.006	0.005	0.004	0.004	**0.000**	*0.003*
		0.018	0.018	0.020	0.020	**0.000**	*0.009*

Table 5: Variance of $\hat{\beta}_1$ (**minimum**, *next best*).

function	location	enet	glmnet	u.enet	u.glmnet	oracular	gwr
step	1	**0.100**	0.101	*0.100*	0.101	0.111	0.118
		0.594	**0.564**	0.594	*0.564*	0.694	0.850
	2	0.196	**0.194**	0.196	*0.194*	0.225	0.244
		1.019	**1.001**	1.019	**1.001**	1.171	1.123
	3	*0.232*	0.233	**0.232**	0.233	0.255	0.262
		0.850	**0.833**	0.850	**0.833**	1.025	1.020
	4	**0.241**	0.250	*0.241*	0.250	0.269	0.288
		0.950	**0.950**	0.950	**0.950**	1.045	1.053
	5	0.231	**0.224**	0.231	*0.224*	0.293	0.234
		0.675	0.697	**0.675**	0.697	0.782	0.716
gradient	1	*0.151*	0.169	**0.151**	0.169	0.213	0.247
		0.559	**0.552**	0.559	*0.552*	0.757	0.895
	2	0.275	**0.273**	0.275	*0.273*	0.311	0.332
		0.897	0.953	**0.897**	0.953	1.000	1.048
	3	0.257	**0.246**	0.257	*0.246*	0.275	0.265
		0.620	0.652	*0.620*	0.652	0.673	0.664
	4	0.293	**0.259**	0.293	**0.259**	0.304	0.333
		0.748	*0.743*	0.748	**0.743**	0.815	0.802
	5	0.259	**0.203**	0.259	**0.203**	0.278	0.238
		0.961	*0.915*	0.961	**0.915**	1.127	0.972
parabola	1	0.224	0.232	0.224	0.232	*0.223*	**0.222**
		0.669	0.671	*0.669*	0.671	0.723	0.757
	2	0.216	0.218	*0.216*	0.218	0.221	**0.210**
		0.814	0.836	*0.814*	0.836	0.863	0.832
	3	0.241	*0.241*	0.241	0.241	0.249	**0.229**
		1.094	1.096	**1.094**	1.096	1.135	1.117
	4	0.276	0.277	*0.276*	0.277	0.281	**0.262**
		0.882	*0.875*	0.882	0.875	0.885	**0.870**
	5	**0.197**	0.202	**0.197**	0.202	0.222	0.202
		1.257	*1.256*	1.257	**1.256**	1.289	1.275

Table 6: Mean squared error of $\hat{Y}$ (**minimum**, *next best*).