高维数据

唐洁, 邹云龙

2022 年 11 月 8 日

目录

1 空间模型

1.1 模型介绍

含空间自相关误差的空间自回归模型 (SARARmodel):

$$Y_n = \rho_1 W_n Y_n + X_n\beta + u_{(n)}, u_{(n)} = \rho_2 M_n u_{(n)} + \epsilon_{(n)}, \tag{1}$$

其中，n 是空间单元数量，$\rho_j, j = 1, 2$ 是空间自回归系数且 $|\rho_j|$<1，j=1,2，$X_n = (x_1, x_2, ..., x_n)'$ 是 $n \times p$ 维解释变量的样本资料矩阵，β 是 $p \times 1$ 维 X_n 的回归系数向量，$Y_n = (y_1, y_2, ..., y_n)'$ 是 $n \times 1$ 维响应变量，W_n 是解释变量 Y_n 的空间邻接权重矩阵，M_n 是扰动项 $u_{(n)}$ 的空间邻接权重矩阵，$\epsilon_{(n)}$ 是 $n \times 1$ 维空间误差向量，且满足

$$E\epsilon_{(n)} = 0, Var(\epsilon_{(n)}) = \sigma^2 I_n.$$

本汇报将呈现 $p \to \infty$ 的情形. 符号说明，$\mathbf{1}_k$ 表示数字 1 组成的 k 维列向量.

由 Qin , 得到关于 $\theta = (\beta', \rho_1, \rho_2, \sigma^2)' \in R^{p+3}$ 的经验 (对数) 似然比统计量:

$$\ell_n(\theta) = -2\log L_n(\theta) = 2\sum_{i=1}^{n} \log\{1 + \lambda'(\theta)\omega_i(\theta)\},$$

$$L_n(\theta) = \sup\left\{\prod_{i=1}^{n}(np_i) : p_1 \geq 0, \cdots, p_n \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i\omega_i(\theta) = 0\right\},$$

其中

$$\omega_i(\theta) = \begin{pmatrix} b_i\epsilon_i \\ \tilde{g}_{ii}(\epsilon_i^2 - \sigma^2) + 2\epsilon_i\sum_{j=1}^{i-1}\tilde{g}_{ij}\epsilon_j + s_i\epsilon_i \\ \tilde{h}_{ii}(\epsilon_i^2 - \sigma^2) + 2\epsilon_i\sum_{j=1}^{i-1}\tilde{h}_{ij}\epsilon_j \\ \epsilon_i^2 - \sigma^2 \end{pmatrix}_{(p+3)\times 1}.$$

情形 1 当 p 固定时，模型 (1) 下，当 $\theta = \theta_0$, $n \to \infty$，有

$$el = \ell_n(\theta_0) \xrightarrow{d} \chi^2_{p+3}.$$

情形 2 当 $p \to \infty$ 时，模型 (1) 下，当 $\theta = \theta_0$, $p = cn^{index}$，$n \to \infty$，有

$$hel = \ell_n^h(\theta_0) \xrightarrow{d} \chi^2_1,$$

其中 $\ell_n^h(\theta) = -2\log L_n^h(\theta) = 2\sum_{i=1}^{n}\log\{1 + \lambda'(\theta)\omega_i^h(\theta)\}$, 且 $\omega_i^h(\theta) = \omega_i'(\theta)\mathbf{1}_{p+3}, i = 1, 2, \cdots, n.$

1.2 分析汇报

在空间模型中，$p = cn^{index}$, 其中 p 为 X_n 的维数.

当 $index = 0$，维数为固定常数，由表 1-2 可看出，情形 1 得到验证. 当 $index > 0$，$p \to \infty$，由表 1-2 可看出，情形 2 得到验证，尤其是当 $index = 0.5$，由表 1-2 可知, EL 的覆盖率都几乎为 0，而 HEL 仍然接近名义水平. 换而言之，HEL 犯第一类错误的频率比 EL 要低，并且随着维数的增加，HEL 还很接近名义水平.

另外，我们设计了功效函数实验，比较二者犯第二类错误的表现. 表 3可知, 随着参数与真值参数的距离越远，HEL 犯第二类错误的频率越接近于 0，即功效函数越接近于 1. 而 EL 的表现劣于 HEL，因为 EL 在参数与真值参数的距离较近的时候，随着维数变大都趋于 1.

综上，HEL 比 EL 在第一类错误和第二类错误的表现中都好.

1.3 实验设计

我们通过模拟比较 EL 和 HEL 优劣，进行模拟 1000 次，在模拟中，使用如下模型:

$$Y_n = \rho_1 W_n Y_n + X_n \beta + u_{(n)}, u_{(n)} = \rho_2 M_n u_{(n)} + \epsilon_{(n)},$$

其中 $\rho_1 = 0.85$, $\rho_2 = 0.15$, $\beta_0 = \mathbf{1}_{p+3}$, $p = [cn^{index}]$, [ ] 表示取整函数，$\{X_i\}$ 服从 $N(\mathbf{0}, \Sigma_P)$，其中 $\Sigma_P = I_{\{i=j, 1 \le i,j \le p\}}$, $\epsilon_i' s$ 分别来自 $N(0, \sigma_0^2)$, 其中 $\sigma_0^2 = 1$. 空间权重矩阵采用皇后邻接，考虑空间单元的 4 种理想情况：规则正方形网格 $n = m \times m$，m=10、15、20、30 分别表示 W_n 为 $grid_{100}$、$grid_{225}$、$grid_{400}$ 和 $grid_{900}$.

记 $\theta' = (\beta', \rho_1, \rho_2, \sigma^2)$，$\beta = \beta_0 + \Delta$, $\sigma^2 = \sigma_0^2 + \Delta$, Δ 取值依次为 0、0.1、1 以及 2.

原假设 $H_0 : \theta = \theta_0$，即 $\Delta = 0$. 备择假设 $H_1 : \theta \neq \theta_0$，即 $\Delta \neq 0$. 取定 α, $0 < \alpha < 1$, 设 $z_\alpha(p+3)$ 满足 $P(\chi^2_{p+3} > z_\alpha(p+3)) = \alpha$，$z_\alpha(1)$ 满足 $P(\chi^2_1 > z_\alpha(1)) = \alpha$.

实验一：原假设下置信区域覆盖率

$\Delta = 0$, $c = (3, 4, 5)$，$index = (0, 0.14, 0.16, 0.24, 0.3, 0.4, 0.5)$，分别给出 $\ell_n(\theta_0) \le z_\alpha(p+3)$ 和 $\ell_n^h(\theta_0) \le z_\alpha(1)$ 在模拟中出现的比例，其中 θ_0 是 θ 的真实值.

实验二：备择假设下拒绝区域覆盖率

$\Delta = (0, 0.1, 1, 2)$, $c = 3$, $index = (0, 0.1, 0.2, 0.3, 0.4, 0.5)$，分别给出 $\ell_n(\theta) \ge z_\alpha(p+3)$ 和 $\ell_n^h(\theta) \ge z_\alpha(1)$ 在模拟中出现的比例，其中 $\theta \neq \theta_0$ 是 θ 的非真实值. 值得一提的是，当 $\Delta = 0.1$ 时，$\|\theta - \theta_0\| \le (n/p)^{-1/3}$.

2 模拟结果

2.1 实验一

表 1: $1-\alpha=0.90$

p, EL, HEL \ n / *index*	100	100	100	225	225	225	400	400	400	900	900	900
0	3	0.825	0.883	3	0.857	0.900	3	0.879	0.883	3	0.892	0.898
0	4	0.773	0.897	4	0.857	0.885	4	0.878	0.889	4	0.876	0.878
0	5	0.742	0.888	5	0.846	0.900	5	0.866	0.892	5	0.888	0.904
0.14	6	0.694	0.896	6	0.831	0.906	7	0.871	0.901	8	0.875	0.903
0.14	8	0.627	0.891	9	0.814	0.896	9	0.850	0.903	10	0.866	0.897
0.14	10	0.566	0.882	11	0.763	0.887	12	0.840	0.903	13	0.865	0.890
0.16	6	0.729	0.894	7	0.813	0.892	8	0.838	0.905	9	0.876	0.909
0.16	8	0.652	0.876	10	0.768	0.904	10	0.835	0.907	12	0.867	0.900
0.16	10	0.567	0.873	12	0.723	0.893	13	0.799	0.883	15	0.863	0.892
0.24	9	0.602	0.895	11	0.735	0.895	13	0.828	0.914	15	0.877	0.902
0.24	12	0.464	0.892	15	0.681	0.888	17	0.744	0.908	20	0.869	0.906
0.24	15	0.331	0.895	18	0.613	0.901	21	0.751	0.886	26	0.832	0.910
0.3	12	0.468	0.892	15	0.678	0.907	18	0.772	0.894	23	0.818	0.887
0.3	16	0.287	0.901	20	0.558	0.901	24	0.703	0.890	31	0.753	0.895
0.3	20	0.147	0.896	25	0.429	0.909	30	0.596	0.893	38	0.749	0.888
0.4	19	0.211	0.890	26	0.419	0.884	33	0.549	0.891	46	0.705	0.884
0.4	25	0.055	0.892	35	0.191	0.906	44	0.335	0.889	61	0.577	0.892
0.4	32	0.005	0.876	44	0.067	0.891	55	0.211	0.887	76	0.379	0.910
0.5	30	0.013	0.887	45	0.063	0.884	60	0.128	0.895	90	0.230	0.882
0.5	40	0.000	0.895	60	0.007	0.911	80	0.007	0.893	120	0.045	0.894
0.5	50	0.000	0.890	75	0.000	0.904	100	0.000	0.903	150	0.004	0.897

表 2: $1-\alpha=0.95$

p, EL, HEL \ n / *index*	100	100	100	225	225	225	400	400	400	900	900	900
0	3	0.890	0.938	3	0.930	0.946	3	0.931	0.937	3	0.954	0.942
0	4	0.858	0.952	4	0.924	0.945	4	0.932	0.945	4	0.941	0.936
0	5	0.821	0.936	5	0.915	0.947	5	0.922	0.944	5	0.938	0.955
0.14	6	0.769	0.942	6	0.903	0.967	7	0.934	0.949	8	0.940	0.945
0.14	8	0.721	0.937	9	0.879	0.941	9	0.911	0.954	10	0.923	0.953
0.14	10	0.662	0.938	11	0.835	0.936	12	0.903	0.958	13	0.918	0.948
0.16	6	0.800	0.942	7	0.887	0.941	8	0.915	0.953	9	0.931	0.953
0.16	8	0.747	0.941	10	0.847	0.948	10	0.897	0.954	12	0.919	0.949
0.16	10	0.653	0.937	12	0.810	0.948	13	0.892	0.938	15	0.921	0.940
0.24	9	0.697	0.953	11	0.812	0.956	13	0.896	0.953	15	0.934	0.953
0.24	12	0.549	0.946	15	0.787	0.948	17	0.843	0.953	20	0.924	0.955
0.24	15	0.433	0.936	18	0.691	0.955	21	0.833	0.942	26	0.904	0.955
0.3	12	0.555	0.944	15	0.772	0.952	18	0.838	0.944	23	0.886	0.953
0.3	16	0.360	0.943	20	0.672	0.950	24	0.778	0.938	31	0.843	0.934
0.3	20	0.192	0.947	25	0.526	0.955	30	0.707	0.951	38	0.850	0.953
0.4	19	0.280	0.943	26	0.526	0.937	33	0.651	0.952	46	0.791	0.943
0.4	25	0.086	0.935	35	0.270	0.958	44	0.440	0.947	61	0.703	0.946
0.4	32	0.008	0.937	44	0.100	0.938	55	0.275	0.949	76	0.518	0.966
0.5	30	0.019	0.946	45	0.090	0.947	60	0.184	0.952	90	0.324	0.929
0.5	40	0.000	0.948	60	0.010	0.954	80	0.012	0.936	120	0.087	0.943
0.5	50	0.000	0.956	75	0.000	0.951	100	0.000	0.951	150	0.008	0.944

2.2 实验二

越接近0.05越好

||theta−theta0||<(n/p)^{−1/3}
不要拒绝那么彻底比较好

越接近1越好

表 3: $1-\alpha=0.95$

N	p	$\triangle=0$		$\triangle=0.1$		$\triangle=1$		$\triangle=2$	
		EL	HEL	EL	HEL	EL	HEL	EL	HEL
100	3	0.146	0.058	0.205	0.091	0.994	0.083	1.000	0.976
	5	0.180	0.062	0.256	0.091	1.000	0.427	1.000	0.999
	8	0.282	0.066	0.325	0.082	1.000	0.732	1.000	1.000
	12	0.453	0.053	0.475	0.082	1.000	0.851	1.000	1.000
	19	0.755	0.058	0.726	0.065	1.000	0.891	1.000	1.000
	30	0.979	0.046	0.968	0.091	1.000	0.945	1.000	1.000
225	3	0.079	0.050	0.140	0.094	1.000	0.358	1.000	1.000
	5	0.097	0.044	0.129	0.089	1.000	0.907	1.000	1.000
	9	0.114	0.063	0.181	0.099	1.000	0.994	1.000	1.000
	15	0.220	0.060	0.263	0.100	1.000	1.000	1.000	1.000
	26	0.513	0.057	0.525	0.100	1.000	1.000	1.000	1.000
	45	0.925	0.045	0.911	0.108	1.000	1.000	1.000	1.000
400	3	0.072	0.061	0.146	0.141	1.000	0.735	1.000	1.000
	5	0.062	0.061	0.160	0.166	1.000	0.999	1.000	1.000
	10	0.092	0.047	0.143	0.117	1.000	1.000	1.000	1.000
	18	0.137	0.046	0.271	0.141	1.000	1.000	1.000	1.000
	33	0.342	0.048	0.586	0.142	1.000	1.000	1.000	1.000
	60	0.807	0.050	0.984	0.177	1.000	1.000	1.000	1.000
900	3	0.049	0.058	0.212	0.226	1.000	0.989	1.000	1.000
	6	0.050	0.055	0.168	0.221	1.000	1.000	1.000	1.000
	12	0.061	0.055	0.216	0.224	1.000	1.000	1.000	1.000
	23	0.101	0.046	0.552	0.235	1.000	1.000	1.000	1.000
	46	0.221	0.059	0.994	0.272	1.000	1.000	1.000	1.000
	90	0.654	0.049	1.000	0.334	1.000	1.000	1.000	1.000