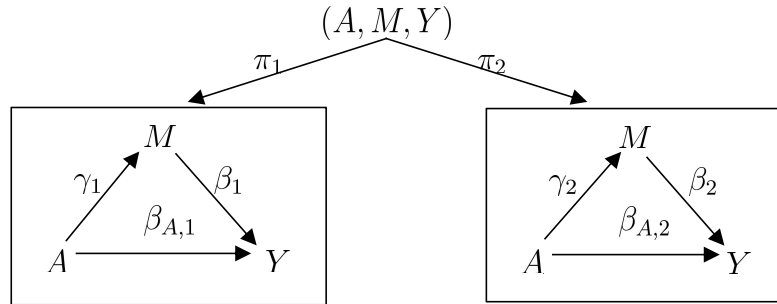
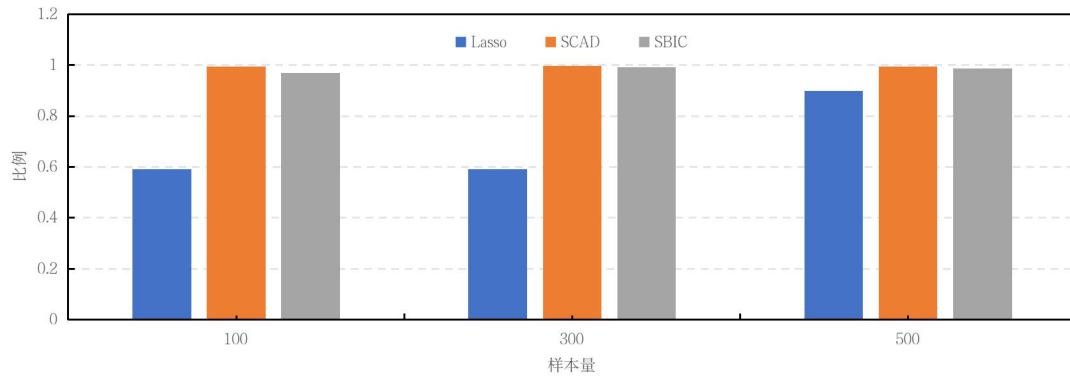
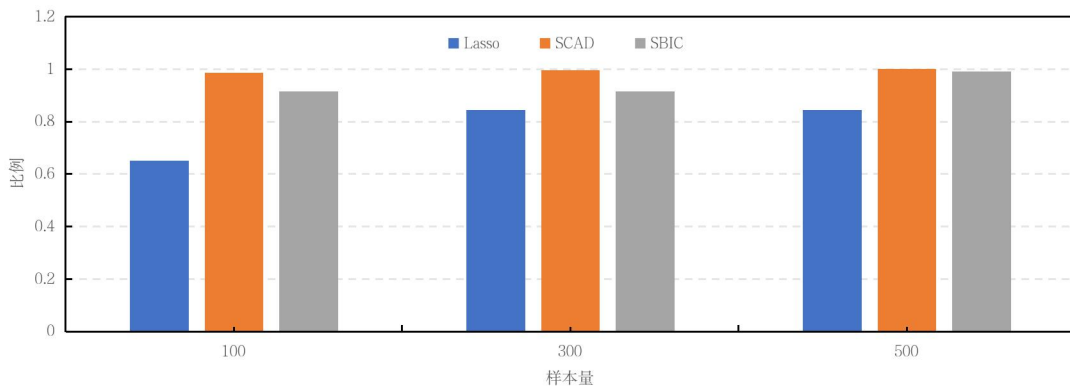


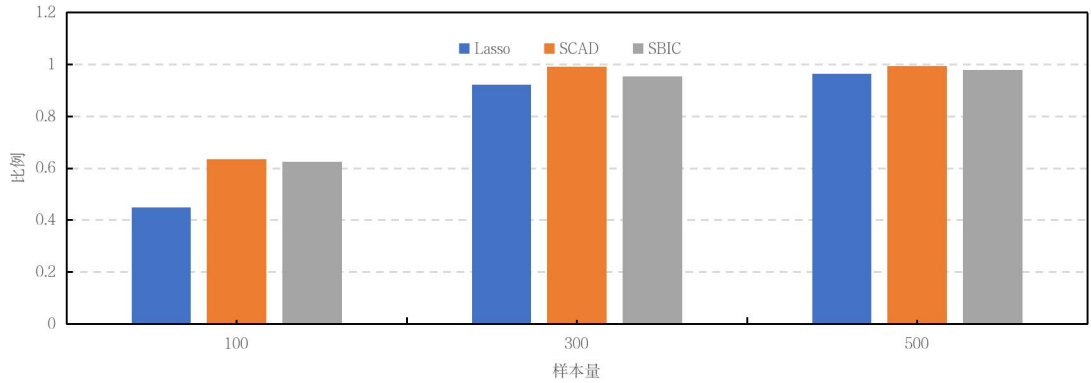
附录

附图

附图 1 两个子群的异质中介效应模型的因果有向无环图 (DAG)^①附图 2 Case I 当 $K = 4$ 三种方法正确选择子群数量的比例^②附图 3 Case II 当 $K = 4$ 三种方法正确选择子群数量的比例

^① π_1 和 π_2 分别表示两个子群的比例。 γ_1 是 A 对 M 的影响， β_1 是 M 对 Y 的影响， $\beta_{A,1}$ 是子群 1 中不由 M 介导的 A 对 Y 的直接作用。 γ_2 ， β_2 ， $\beta_{A,2}$ 在子群 2 中被类似地定义。

^② 不同的颜色代表不同的估计子群数量的方法，蓝色代表 Lasso 方法，橙色代表 SCAD 方法，灰色代表 SBIC 方法。



附图 4 Case III 当 $K = 4$ 三种方法正确选择子群数量的比例

数值模拟结果

附表 1 展示了 Case I 中 Lasso 和 SCAD 方法正确估计子群数量的比例。可以看到 SCAD 方法比 Lasso 方法选择正确子群数量的比例高。对于 Lasso 方法来说，样本量为 100 和 300 时正确选择比例几乎相同，当样本量增大到 500 时，正确选择子群数量的比例增大。当样本量固定时，初始最大子群数的大小不会影响正确选择的比例。对于 SCAD 方法来说，初始最大子群数固定时，样本量越大，正确选择的比例越大。样本量固定时，初始最大子群数为 4 时，正确选择子群数量的比例最大。

附表 1		Case I 中 Lasso 和 SCAD 方法正确选择子群数量的比例				
	Lasso			SCAD		
	$K = 4$	$K = 6$	$K = 8$	$K = 4$	$K = 6$	$K = 8$
$N = 100$	0.59	0.59	0.59	0.994	0.974	0.984
$N = 300$	0.59	0.59	0.59	0.998	0.99	0.988
$N = 500$	0.898	0.898	0.898	0.998	0.996	0.988

附表 2 展示了 Case II 中 Lasso 和 SCAD 方法正确估计子群数量的比例。可以看到 SCAD 方法比 Lasso 方法正确选择子群数量的比例高。对于 Lasso 方法来说，样本量增大时正确比例也变大，当样本量增大到 300 及以上时，正确选择比例不再发生变化。当样本量固定时，初始最大子群数的大小不会影响正确选择的比例。对于 SCAD 方法来说，初始最大子群数固定时，样本量越大，正确选择的比例越大。样本量固定时，初始最大子群数为 4 时，正确选择子群数量的比例最大。

附表 2		Case II 中 Lasso 和 SCAD 方法正确选择子群数量的比例				
	Lasso			SCAD		
	$K = 4$	$K = 6$	$K = 8$	$K = 4$	$K = 6$	$K = 8$
$N = 100$	0.65	0.65	0.65	0.986	0.978	0.968
$N = 300$	0.84	0.84	0.84	0.996	0.984	0.986
$N = 500$	0.84	0.84	0.84	1	0.992	0.988

附表 3 展示了在 Case I 中当 $K = 4$ 时估计参数的平均偏差和标准误。可以看到在不同样本量设置情形下，除个别情况外，三种方法中 SCAD 方法估计参数的平均偏差和标准误最小，SCAD 方法参数估计的效果最好。整体来看，样本量越大，这三种方法估计参数的平均偏差和标准误越小。当 $K = 4$ 时，Case II 和 Case III 估计参数的平均偏差和标准误结论相似。

附表 3 Case I 当 $K = 4$ 时估计参数的平均偏差和标准误						
n	k	方法	$\text{bias}(\hat{\pi}_k)(\text{se})^{\textcircled{3}}$	$\text{bias}(\hat{\gamma}_k)(\text{se})$	$\text{bias}(\hat{\beta}_k)(\text{se})$	$\text{bias}(\hat{\gamma}_k\hat{\beta}_k)(\text{se})$
100	1	Lasso	-0.003(0.006)	0.025(0.020)	-0.011(0.021)	-0.036(0.044)
		SCAD	0.000(0.002)	0.009(0.007)	0.010(0.008)	-0.014(0.016)
		SBIC	0.003(0.004)	-0.018(0.016)	0.020(0.016)	0.042(0.033)
	2	Lasso	-0.003(0.006)	-0.001(0.020)	0.005(0.021)	0.031(0.048)
		SCAD	0.000(0.002)	0.019(0.007)	0.005(0.007)	0.022(0.016)
		SBIC	0.003(0.004)	0.020(0.017)	0.012(0.017)	0.041(0.041)
300	1	Lasso	-0.003(0.006)	0.025(0.020)	0.005(0.021)	0.031(0.048)
		SCAD	0.001(0.001)	0.000(0.004)	0.001(0.004)	0.002(0.009)
		SBIC	0.001(0.001)	0.000(0.004)	0.001(0.004)	-0.001(0.008)
	2	Lasso	0.003(0.006)	-0.001(0.020)	-0.001(0.007)	0.004(0.005)
		SCAD	-0.001(0.001)	-0.007(0.004)	-0.003(0.004)	-0.013(0.008)
		SBIC	-0.001(0.001)	-0.003(0.004)	0.006(0.004)	0.001(0.008)
500	1	Lasso	0.000(0.001)	-0.002(0.003)	0.003(0.003)	0.003(0.007)
		SCAD	0.000(0.001)	0.001(0.003)	0.000(0.003)	0.000(0.007)
		SBIC	-0.001(0.001)	0.003(0.003)	0.000(0.003)	0.000(0.007)
	2	Lasso	0.000(0.001)	0.002(0.003)	-0.006(0.003)	-0.010(0.007)
		SCAD	0.000(0.001)	0.001(0.003)	0.000(0.003)	0.004(0.006)
		SBIC	0.001(0.001)	0.002(0.003)	-0.002(0.003)	0.003(0.007)

附表 4 展示了在 Case II 中当 $K = 4$ 时估计参数的平均偏差和蒙特卡罗标准误。总的来说，SCAD 的效果也是最好的。无论样本量大小如何，SCAD 估计的两个子群所占比例的偏差都接近于零。当样本量达到 500 时，SCAD 方法和 SBIC 方法估计的中介效应平均偏差都接近。

附表 5 展示了在 Case III 中当 $K = 4$ 时估计参数的平均偏差和蒙特卡罗标准误。与其他例子其他样本量相比，当样本量为 100 时，三个方法估计参数的平均偏差和蒙特卡罗标准误都相对较大。但当样本量中等偏大时，效果会变好很多。特别是当样本量为 500 时，三种方法估计的第二个子群的中介效应平均偏差和蒙特卡罗标准误都接近于 0，效果极好。

整体来说，SCAD 和 SBIC 方法比 Lasso 方法效果要好。尽管有些 SBIC 的参数估计的平均偏差和蒙特卡罗标准误要比 SCAD 估计的小，但总体 SCAD 方法估计的平均偏差和蒙特卡罗标准误小的情况更多，加之 SCAD 选择正确子群数量的比例更高，因此 SCAD 比 SBIC 方法效果要好。

^③ bias 表示偏差，se 表示标准误。

附表 4 Case II 当 $K = 4$ 时估计参数的平均偏差和标准误

n	k	方法	$\text{bias}(\hat{\pi}_k)(\text{se})$	$\text{bias}(\hat{\gamma}_k)(\text{se})$	$\text{bias}(\hat{\beta}_k)(\text{se})$	$\text{bias}(\hat{\gamma}_k\hat{\beta}_k)(\text{se})$
100	1	Lasso	-0.014(0.005)	-0.020(0.023)	0.001(0.018)	0.017(0.018)
		SCAD	0.000(0.002)	-0.014(0.008)	-0.023(0.006)	-0.004(0.005)
		SBIC	0.002(0.002)	-0.030(0.009)	-0.022(0.005)	-0.006(0.006)
	2	Lasso	0.014(0.005)	0.037(0.014)	-0.029(0.025)	-0.025(0.017)
		SCAD	0.000(0.002)	0.032(0.006)	0.000(0.010)	0.002(0.007)
		SBIC	-0.002(0.002)	0.027(0.005)	0.000(0.010)	-0.007(0.007)
300	1	Lasso	-0.009(0.001)	0.009(0.005)	0.002(0.003)	0.008(0.003)
		SCAD	0.000(0.001)	-0.014(0.005)	-0.002(0.003)	-0.004(0.003)
		SBIC	0.002(0.002)	-0.030(0.009)	-0.022(0.005)	-0.006(0.006)
	2	Lasso	0.009(0.001)	0.012(0.003)	-0.001(0.007)	0.004(0.005)
		SCAD	0.000(0.001)	0.000(0.003)	-0.002(0.006)	-0.008(0.004)
		SBIC	-0.009(0.001)	0.009(0.005)	0.002(0.003)	0.008(0.003)
500	1	Lasso	0.000(0.001)	-0.002(0.003)	0.003(0.003)	0.003(0.007)
		SCAD	0.000(0.001)	-0.003(0.004)	0.001(0.002)	0.001(0.002)
		SBIC	0.000(0.001)	-0.006(0.004)	0.002(0.002)	-0.001(0.002)
	2	Lasso	0.009(0.001)	0.012(0.003)	-0.001(0.007)	0.004(0.005)
		SCAD	0.000(0.001)	0.004(0.002)	-0.001(0.005)	0.000(0.003)
		SBIC	0.000(0.001)	0.003(0.002)	-0.001(0.005)	0.000(0.003)

附表 5 Case III 当 $K = 4$ 时估计参数的平均偏差和标准误

n	k	方法	$\text{bias}(\hat{\pi}_k)(\text{se})$	$\text{bias}(\hat{\gamma}_k)(\text{se})$	$\text{bias}(\hat{\beta}_k)(\text{se})$	$\text{bias}(\hat{\gamma}_k\hat{\beta}_k)(\text{se})$
100	1	Lasso	-0.015(0.008)	0.010(0.009)	-0.300(0.295)	-0.014(0.047)
		SCAD	-0.002(0.002)	-0.022(0.004)	-0.073(0.025)	-0.025(0.016)
		SBIC	0.003(0.002)	-0.015(0.004)	-0.067(0.023)	-0.020(0.017)
	2	Lasso	-0.005(0.008)	-0.056(0.025)	0.036(0.049)	0.014(0.010)
		SCAD	0.004(0.003)	-0.063(0.010)	0.017(0.021)	0.012(0.004)
		SBIC	0.011(0.003)	-0.037(0.010)	0.028(0.020)	0.007(0.005)
	3	Lasso	0.020(0.013)	0.081(0.024)	0.080(0.048)	0.039(0.041)
		SCAD	-0.002(0.004)	0.035(0.008)	0.074(0.019)	0.018(0.015)
		SBIC	0.001(0.001)	0.020(0.007)	0.088(0.026)	0.025(0.016)
	1	Lasso	-0.002(0.001)	-0.003(0.002)	0.001(0.008)	0.004(0.006)
		SCAD	0.001(0.001)	0.000(0.002)	0.010(0.008)	0.007(0.006)
		SBIC	0.002(0.001)	-0.001(0.002)	-0.002(0.009)	0.000(0.007)
		Lasso	0.003(0.002)	0.000(0.004)	0.010(0.009)	0.001(0.000)

300	2	SCAD	0.000(0.002)	-0.009(0.004)	-0.002(0.009)	0.001(0.000)
		SBIC	-0.002(0.002)	-0.002(0.004)	0.000(0.009)	0.001(0.000)
	3	Lasso	-0.001(0.002)	-0.002(0.003)	0.010(0.007)	0.004(0.004)
		SCAD	-0.002(0.002)	0.002(0.003)	-0.005(0.007)	-0.009(0.004)
		SBIC	0.000(0.002)	0.000(0.003)	-0.014(0.006)	-0.008(0.004)
	500	1	Lasso	-0.002(0.001)	0.002(0.001)	0.006(0.007)
SCAD			0.001(0.001)	-0.001(0.001)	-0.002(0.006)	-0.001(0.005)
SBIC			0.002(0.001)	-0.001(0.001)	0.012(0.007)	0.011(0.005)
2		Lasso	-0.004(0.001)	-0.004(0.003)	0.008(0.007)	0.000(0.000)
		SCAD	0.001(0.001)	-0.005(0.003)	-0.006(0.006)	0.000(0.000)
		SBIC	0.001(0.001)	-0.005(0.003)	0.006(0.007)	0.000(0.000)
3		Lasso	0.006(0.001)	0.002(0.002)	-0.005(0.005)	-0.005(0.003)
		SCAD	-0.001(0.001)	-0.001(0.002)	0.001(0.005)	-0.003(0.003)
		SBIC	-0.003(0.001)	-0.004(0.002)	0.005(0.005)	0.000(0.003)

附表 6 展示了在 Case I 中当 $K = 6$ 和 $K = 8$ 时估计参数的平均偏差和蒙特卡罗标准误。附表 7 展示了在 Case II 中当 $K = 6$ 和 $K = 8$ 时估计参数的平均偏差和蒙特卡罗标准误。与例一相同，SCAD 方法参数估计的效果总体上要好于 Lasso 方法。当 $K = 6, n = 500$ 、 $K = 8, n = 300$ 以及 $K = 8, n = 500$ 的情况下，SCAD 方法估计两个子群比例的平均偏差几乎接近 0，效果很好。

附表 6 Case I 当 $K = 6$ 和 $K = 8$ 时估计参数的平均偏差和标准误

K	n	k	方法	$\text{bias}(\hat{\pi}_k)(\text{se})$	$\text{bias}(\hat{\gamma}_k)(\text{se})$	$\text{bias}(\hat{\beta}_k)(\text{se})$	$\text{bias}(\hat{\gamma}_k \hat{\beta}_k)(\text{se})$
6	100	1	Lasso	-0.003(0.006)	0.025(0.020)	-0.011(0.021)	-0.036(0.044)
			SCAD	-0.001(0.002)	0.003(0.007)	-0.011(0.008)	-0.022(0.016)
		2	Lasso	0.003(0.006)	-0.001(0.020)	0.005(0.021)	0.031(0.048)
			SCAD	0.001(0.002)	0.002(0.007)	0.009(0.008)	0.015(0.016)
	300	1	Lasso	-0.003(0.006)	0.025(0.020)	-0.011(0.021)	-0.036(0.044)
			SCAD	-0.001(0.001)	0.006(0.004)	-0.002(0.004)	-0.005(0.009)
		2	Lasso	0.003(0.006)	-0.001(0.020)	0.005(0.021)	0.031(0.048)
			SCAD	0.001(0.001)	-0.002(0.004)	0.001(0.004)	0.005(0.009)
	500	1	Lasso	0.000(0.001)	-0.002(0.003)	0.003(0.003)	0.003(0.007)
			SCAD	-0.001(0.001)	0.002(0.003)	0.004(0.003)	0.000(0.006)
		2	Lasso	0.000(0.001)	0.002(0.003)	-0.006(0.003)	-0.010(0.007)
			SCAD	-0.001(0.001)	0.004(0.003)	-0.002(0.003)	0.000(0.007)
	100	1	Lasso	-0.003(0.006)	0.025(0.020)	-0.011(0.021)	-0.036(0.044)
			SCAD	0.000(0.002)	0.009(0.007)	0.006(0.007)	0.004(0.016)
		2	Lasso	0.003(0.006)	-0.001(0.020)	0.005(0.021)	0.031(0.048)
			SCAD	0.000(0.002)	0.014(0.007)	0.006(0.008)	0.040(0.016)

8	300	1	Lasso	-0.003(0.006)	0.025(0.020)	-0.011(0.021)	-0.036(0.044)
			SCAD	0.000(0.001)	-0.003(0.004)	-0.005(0.004)	-0.001(0.009)
		2	Lasso	0.003(0.006)	-0.001(0.020)	0.005(0.021)	0.031(0.048)
			SCAD	0.000(0.001)	0.006(0.004)	-0.002(0.004)	0.008(0.009)
	500	1	Lasso	0.000(0.001)	-0.002(0.003)	0.003(0.003)	0.003(0.007)
			SCAD	0.001(0.001)	0.009(0.003)	-0.001(0.003)	-0.008(0.006)
		2	Lasso	0.000(0.001)	0.002(0.003)	-0.006(0.003)	-0.010(0.007)
			SCAD	-0.001(0.001)	0.002(0.003)	-0.003(0.003)	0.004(0.006)

附表 7 Case II 当 $K = 6$ 和 $K = 8$ 时估计参数的平均偏差和标准误

K	n	k	方法	$\text{bias}(\hat{\pi}_k)(\text{se})$	$\text{bias}(\hat{\gamma}_k)(\text{se})$	$\text{bias}(\hat{\beta}_k)(\text{se})$	$\text{bias}(\hat{\gamma}_k\hat{\beta}_k)(\text{se})$
6	100	1	Lasso	-0.014(0.005)	-0.020(0.023)	0.001(0.018)	0.017(0.018)
			SCAD	0.002(0.002)	-0.031(0.009)	-0.016(0.006)	-0.006(0.006)
		2	Lasso	0.014(0.005)	0.037(0.014)	-0.029(0.025)	-0.025(0.017)
			SCAD	0.001(0.002)	0.030(0.005)	0.011(0.010)	0.006(0.008)
	300	1	Lasso	-0.009(0.001)	0.009(0.005)	0.002(0.003)	0.008(0.003)
			SCAD	-0.001(0.001)	0.006(0.005)	-0.006(0.003)	0.004(0.008)
		2	Lasso	0.009(0.001)	0.012(0.003)	-0.001(0.007)	0.004(0.005)
			SCAD	0.001(0.001)	0.002(0.003)	0.003(0.006)	-0.001(0.004)
	500	1	Lasso	-0.009(0.001)	0.009(0.005)	0.002(0.003)	0.008(0.003)
			SCAD	0.000(0.001)	-0.003(0.004)	0.000(0.002)	-0.001(0.002)
		2	Lasso	0.000(0.001)	0.012(0.003)	-0.001(0.007)	0.004(0.005)
			SCAD	0.009(0.001)	0.001(0.002)	0.007(0.005)	0.003(0.003)
8	100	1	Lasso	-0.014(0.005)	-0.020(0.023)	0.001(0.018)	0.017(0.018)
			SCAD	-0.002(0.002)	-0.031(0.009)	-0.025(0.006)	-0.009(0.005)
		2	Lasso	0.014(0.005)	0.037(0.014)	-0.029(0.025)	-0.025(0.017)
			SCAD	0.002(0.002)	0.025(0.005)	0.020(0.010)	0.001(0.007)
	300	1	Lasso	-0.009(0.001)	0.009(0.005)	0.002(0.003)	0.008(0.003)
			SCAD	0.000(0.001)	-0.004(0.005)	-0.003(0.003)	-0.001(0.003)
		2	Lasso	0.009(0.001)	0.012(0.003)	-0.001(0.007)	0.004(0.005)
			SCAD	0.000(0.001)	0.003(0.003)	0.012(0.006)	0.005(0.004)
	500	1	Lasso	-0.009(0.001)	0.009(0.005)	0.002(0.003)	0.008(0.003)
			SCAD	0.000(0.001)	0.004(0.004)	0.003(0.002)	0.003(0.002)
		2	Lasso	0.009(0.001)	0.012(0.003)	-0.001(0.007)	0.004(0.005)
			SCAD	0.000(0.001)	0.006(0.002)	-0.002(0.005)	0.003(0.003)

附表 8 展示了在 Case III 中当 $K = 6$ 时估计参数的平均偏差和标准误。SCAD 和 Lasso 方法参数估计的效果都比较好,尤其是在样本量中等偏大时。相比之下,SCAD 方法比 Lasso 方法效果更好,估计参数的偏差和标准误更小。而且随着样本量的增大,SCAD 和 Lasso 方

法参数估计的效果变得更好。

附表 9 展示了 Case III 中当 $K = 8$ 时估计参数的平均偏差和标准误。与 $K = 6$ 时类似，SCAD 方法比 Lasso 方法效果更好，且随着样本量的增大，两种方法参数估计的效果变得更好。

附表 8 Case III 当 $K = 6$ 时估计参数的平均偏差和标准误

n	k	方法	$\text{bias}(\hat{\pi}_k)(\text{se})$	$\text{bias}(\hat{\gamma}_k)(\text{se})$	$\text{bias}(\hat{\beta}_k)(\text{se})$	$\text{bias}(\hat{\gamma}_k\hat{\beta}_k)(\text{se})$
100	1	Lasso	-0.015(0.008)	0.010(0.009)	-0.300(0.295)	-0.014(0.047)
		SCAD	-0.001(0.002)	-0.008(0.005)	-0.092(0.023)	-0.046(0.016)
	2	Lasso	-0.005(0.008)	-0.056(0.025)	0.036(0.049)	0.014(0.010)
		SCAD	0.003(0.003)	-0.027(0.010)	0.030(0.021)	0.007(0.003)
	3	Lasso	0.020(0.013)	0.081(0.024)	0.080(0.048)	0.039(0.041)
		SCAD	-0.002(0.004)	0.028(0.008)	0.071(0.018)	0.016(0.012)
300	1	Lasso	-0.002(0.001)	-0.003(0.002)	0.001(0.008)	0.004(0.006)
		SCAD	0.000(0.001)	-0.002(0.002)	-0.007(0.008)	-0.005(0.006)
	2	Lasso	0.003(0.002)	0.000(0.004)	0.010(0.009)	0.001(0.000)
		SCAD	0.000(0.002)	-0.001(0.004)	0.001(0.009)	0.001(0.000)
	3	Lasso	-0.001(0.002)	-0.002(0.003)	0.010(0.007)	0.004(0.004)
		SCAD	0.000(0.002)	0.005(0.003)	-0.011(0.007)	-0.005(0.005)
500	1	Lasso	-0.002(0.001)	0.002(0.001)	0.006(0.007)	0.003(0.005)
		SCAD	0.000(0.001)	0.001(0.001)	0.003(0.007)	0.000(0.005)
	2	Lasso	-0.004(0.001)	-0.004(0.003)	0.008(0.007)	0.000(0.000)
		SCAD	-0.002(0.001)	-0.002(0.003)	0.001(0.007)	0.000(0.000)
	3	Lasso	0.006(0.001)	0.002(0.002)	-0.005(0.005)	-0.005(0.003)
		SCAD	0.002(0.001)	0.004(0.002)	-0.002(0.005)	-0.002(0.003)

附表 9 Case III 当 $K = 8$ 时估计参数的平均偏差和标准误

n	k	方法	$\text{bias}(\hat{\pi}_k)(\text{se})$	$\text{bias}(\hat{\gamma}_k)(\text{se})$	$\text{bias}(\hat{\beta}_k)(\text{se})$	$\text{bias}(\hat{\gamma}_k\hat{\beta}_k)(\text{se})$
100	1	Lasso	-0.015(0.008)	0.010(0.009)	-0.300(0.295)	-0.014(0.047)
		SCAD	-0.002(0.002)	-0.018(0.004)	-0.054(0.025)	-0.027(0.017)
	2	Lasso	-0.005(0.008)	-0.056(0.025)	0.036(0.049)	0.014(0.010)
		SCAD	0.004(0.003)	-0.058(0.010)	0.024(0.020)	0.010(0.003)
	3	Lasso	0.020(0.013)	0.081(0.024)	0.080(0.048)	0.039(0.041)
		SCAD	-0.002(0.004)	0.043(0.008)	0.043(0.020)	-0.015(0.011)
300	1	Lasso	-0.002(0.001)	-0.003(0.002)	0.001(0.008)	0.004(0.006)
		SCAD	0.000(0.001)	-0.001(0.002)	-0.007(0.009)	-0.005(0.007)
	2	Lasso	0.003(0.002)	0.000(0.004)	0.010(0.009)	0.001(0.000)
		SCAD	-0.003(0.002)	-0.003(0.004)	-0.005(0.008)	0.001(0.000)

500	3	Lasso	-0.001(0.002)	-0.002(0.003)	0.010(0.007)	0.004(0.004)
		SCAD	0.003(0.002)	0.005(0.003)	-0.001(0.007)	0.000(0.004)
	1	Lasso	-0.015(0.001)	0.002(0.002)	0.014(0.007)	0.008(0.005)
		SCAD	0.000(0.001)	-0.002(0.001)	-0.001(0.006)	-0.002(0.005)
	2	Lasso	-0.008(0.002)	0.003(0.004)	0.019(0.008)	0.000(0.000)
		SCAD	0.000(0.001)	0.002(0.003)	-0.001(0.007)	0.000(0.000)
	3	Lasso	0.023(0.002)	0.005(0.003)	-0.008(0.006)	-0.002(0.004)
		SCAD	0.000(0.001)	0.003(0.002)	0.000(0.005)	0.002(0.003)

附录 1 定理及证明

定理1 在 (9) 式中,若 $\sqrt{n}\lambda \rightarrow \infty$, $\lambda \rightarrow 0$ 且 $\varepsilon = o(1/\sqrt{n})$, 则存在 $l_{p,2}$ 的一个局部最大值点 $(\hat{\theta}, \hat{\eta})$, 使得 $\hat{\theta} = O_P(1/\sqrt{n})$, $\hat{q} \xrightarrow{P} q$, 其中 $\hat{q}$ 为估计的子群个数, q 为真实的子群个数.

证明: 为了证明该定理, 首先证明存在一个最大值点 $(\hat{\theta}, \hat{\eta})$, 使得 $\hat{\theta} = O_P(1/\sqrt{n})$. 事实上, 只需要证明对于一个较大的常数 C , $l_{p,2}(\theta, \eta) < l_{p,2}(0, \eta)$, 其中 $\theta = C/\sqrt{n}$, η 在一个局部紧集里. 如果 $\theta = C/\sqrt{n}$, 则

$$\begin{aligned}
& l_{p,2}(\theta, \eta) - l_{p,2}(0, \eta) \\
&= \sum_{i=1}^n \log f(Y_i, M_i \mid A_i, X_i; \theta, \eta) - n\lambda D_f \sum_{k=1}^K [\log(\varepsilon + P_\lambda(\pi_k)) - \log(\varepsilon)] \\
&\quad - \left(\sum_{i=1}^n \log g_0(Y_i, M_i \mid A_i, X_i) - n\lambda D_f \sum_{l=1}^q [\log(\varepsilon + P_\lambda(\pi_l^0)) - \log(\varepsilon)] \right) \\
&= \sum_{i=1}^n [\log f(Y_i, M_i \mid A_i, X_i; \theta, \eta) - \log g_0(Y_i, M_i \mid A_i, X_i)] \\
&\quad - n\lambda D_f \sum_{k=1}^{K-q} [\log(\varepsilon + P_\lambda(\pi_k)) - \log(\varepsilon)] \\
&\quad - n\lambda D_f \sum_{k=K-q+1}^K [\log(\varepsilon + P_\lambda(\pi_k)) - \log(\varepsilon)] \\
&\quad + n\lambda D_f \sum_{l=1}^q [\log(\varepsilon + P_\lambda(\pi_l^0)) - \log(\varepsilon)] \\
&= \sum_{i=1}^n [\log f(Y_i, M_i \mid A_i, X_i; \theta, \eta) - \log g_0(Y_i, M_i \mid A_i, X_i)] \\
&\quad - n\lambda D_f \sum_{k=K-q+1}^K [\log(\varepsilon + P_\lambda(\pi_k)) - \log(\varepsilon + P_\lambda(\pi_{k-K+q}^0))] \\
&\quad - n\lambda D_f \sum_{k=1}^{K-q} \left[\log \left(\frac{\varepsilon + P_\lambda(\pi_k)}{\varepsilon} \right) \right] \\
&\leq \sum_{i=1}^n [\log f(Y_i, M_i \mid A_i, X_i; \theta, \eta) - \log g_0(Y_i, M_i \mid A_i, X_i)] \\
&\quad - n\lambda D_f \sum_{k=K-q+1}^K [\log(\varepsilon + P_\lambda(\pi_k)) - \log(\varepsilon + P_\lambda(\pi_{k-K+q}^0))] \\
&\triangleq I_1 + I_2.
\end{aligned}$$

对于 I_2 , 由于 $\theta = C/\sqrt{n}$ 和 ρ_l 上的约束条件 ($l = 1, \dots, q$), 当 $k > K - q$ 时, $|\pi_k - \pi_{k-K+q}^0| \leq C/\sqrt{n}$. 由惩罚函数的性质, 有

$$\begin{aligned} |I_2| &= \left| -n\lambda D_f \sum_{k=K-q+1}^K [\log(\varepsilon + P_\lambda(\pi_k)) - \log(\varepsilon + P_\lambda(\pi_{k-K+q}^0))] \right| \\ &= \left| -n\lambda D_f \sum_{k=K-q+1}^K \left[\log\left(\varepsilon + \frac{(a+1)\lambda}{2}\right) - \log\left(\varepsilon + \frac{(a+1)\lambda}{2}\right) \right] \right| \\ &= 0. \end{aligned}$$

对于 I_1 , 如果 $\theta = C/\sqrt{n}$, 有

$$\begin{aligned} I_1 &= \sum_{i=1}^n \frac{f(Y_i, M_i | A_i, X_i; \theta, \eta) - g_0(Y_i, M_i | A_i, X_i)}{g_0(Y_i, M_i | A_i, X_i)} \\ &\quad - \frac{1}{2} \sum_{i=1}^n \left(\frac{f(Y_i, M_i | A_i, X_i; \theta, \eta) - g_0(Y_i, M_i | A_i, X_i)}{g_0(Y_i, M_i | A_i, X_i)} \right)^2 (1 + o_P(1)). \end{aligned}$$

对 $f(Y, M | A, X; \theta, \eta)$ 进行泰勒展开, 有

$$f(Y, M | A, \mathbf{X}; \theta, \boldsymbol{\eta}) = g_0(Y, M | A, \mathbf{X}) + \theta \cdot f'(Y, M | A, X; 0, \boldsymbol{\eta})(1 + o_P(1)).$$

则

$$\begin{aligned} I_1 &= \left\{ \sum_{i=1}^n \theta \frac{f'(Y_i, M_i | A_i, X_i; 0, \boldsymbol{\eta})}{g_0(Y_i, M_i | A_i, X_i)} - \frac{1}{2} \sum_{i=1}^n \theta^2 \left(\frac{f'(Y_i, M_i | A_i, X_i; 0, \boldsymbol{\eta})}{g_0(Y_i, M_i | A_i, X_i)} \right)^2 \right\} \\ &\quad \times (1 + o_P(1)). \end{aligned}$$

根据中心极限定理, 有

$$\sqrt{n} \cdot \left(\frac{1}{n} \sum_{i=1}^n \frac{f'(Y_i, M_i | A_i, \mathbf{X}_i; 0, \boldsymbol{\eta})}{g_0(Y_i, M_i | A_i, \mathbf{X}_i)} - E \left(\frac{f'(Y_i, M_i | A_i, \mathbf{X}_i; 0, \boldsymbol{\eta})}{g_0(Y_i, M_i | A_i, \mathbf{X}_i)} \right) \right) = O_P(1),$$

由 Bartlett 恒等式, $E f' / g_0 = 0$, 则

$$\sum_{i=1}^n \frac{f'(Y_i, M_i | A_i, \mathbf{X}_i; 0, \boldsymbol{\eta})}{g_0(Y_i, M_i | A_i, \mathbf{X}_i)} = O_P(\sqrt{n}).$$

由大数定律, 有

$$\sum_{i=1}^n \left(\frac{f'(Y_i, M_i | A_i, \mathbf{X}_i; 0, \boldsymbol{\eta})}{g_0(Y_i, M_i | A_i, \mathbf{X}_i)} \right)^2 = O_P(n).$$

因此

$$I_1 = \frac{C}{\sqrt{n}} \cdot O_P(\sqrt{n}) - \frac{C^2}{n} \cdot O_P(n).$$

当 C 足够大时, 在惩罚似然函数中, I_1 的第二项占主导地位, 则有 $l_{p,2}(\theta, \boldsymbol{\eta}) < l_{p,2}(0, \boldsymbol{\eta})$

依概率趋于 1 成立. 因此存在一个最大值点 $(\hat{\theta}, \hat{\boldsymbol{\eta}})$, 使得 $\hat{\theta} = O_P(1/\sqrt{n})$.

接下来, 证明存在一个满足 $\hat{\theta} = O_P(1/\sqrt{n})$ 对应的最大值点 $(\hat{\theta}, \hat{\eta})$, 有 $\hat{q} = q$ 或 $\hat{\pi}_k = 0, k = 1, \dots, K - q$. 事实上, 当 $\hat{\theta} = O_P(1/\sqrt{n})$ 时, 由 ξ_k 的约束条件, 有 $\hat{\pi}_k = O_P(1/\sqrt{n}), k = 1, \dots, K - q$. 对于约束条件 $\sum_{k=1}^K \hat{\pi}_k = 1$, 考虑拉格朗日乘子 ω , 只需要证明当 $\varepsilon_n = C/\sqrt{n}, k \leq K - q$ 时,

$$\left. \frac{\partial l^*(\theta)}{\partial \pi_k} \right|_{\pi_k = \hat{\pi}_k} < 0, \hat{\pi}_k < \varepsilon_n \quad (\text{A.1})$$

对于 $(\hat{\theta}, \hat{\eta})$ 依概率趋于1, 其中 $l^*(\theta) = l_{p,2}(\theta) - \omega(\sum_{k=1}^K \hat{\pi}_k - 1)$.

为了证明 (A.1) 成立, 首先考虑当 $k > K - q$ 时关于 π_k 的偏导数, 令

$$\left. \frac{\partial l^*(\theta)}{\partial \pi_k} \right|_{\pi_k = \hat{\pi}_k} = \sum_{i=1}^n \frac{\phi(\alpha_k)}{\sum_{k=1}^K \hat{\pi}_k \phi(\alpha_k)} - n\lambda D_f \frac{P'_\lambda(\hat{\pi}_k)}{\varepsilon + P_\lambda(\hat{\pi}_k)} - \omega = 0. \quad (\text{A.2})$$

根据大数定律, (A.2) 中的第一项是 $O_P(n)$. 给定 $k > K - q, \hat{\theta} = O_P(1/\sqrt{n})$, 因为 $\hat{\pi}_k > a\lambda$, 此时 (A.2) 中的第二项 $P_\lambda(\hat{\pi}_k)$ 为常数, 则 $P'_\lambda(\hat{\pi}_k)$ 等于 0, 因此第二项为 0. 因此 $\omega = O_P(n)$.

接下来, 考虑当 $k \leq K - q$ 时关于 π_k 的偏导数,

$$\left. \frac{\partial l^*(\theta)}{\partial \pi_k} \right|_{\pi_k = \hat{\pi}_k} = \sum_{i=1}^n \frac{\phi(\alpha_k)}{\sum_{k=1}^K \hat{\pi}_k \phi(\alpha_k)} - n\lambda D_f \frac{1}{\varepsilon + \hat{\pi}_k} - \omega. \quad (\text{A.3})$$

(A.3) 中的第一项和第三项 $\omega = O_P(n)$. 对于第二项, 因为 $\hat{\pi}_k = O_P(1/\sqrt{n}), \varepsilon = o(1/\sqrt{n}), \sqrt{n}\lambda \rightarrow \infty$, 则

$$\frac{n\lambda D_f / (\varepsilon + \hat{\pi}_k)}{n} = \lambda D_f \frac{1}{\varepsilon + \hat{\pi}_k} = O_P(\sqrt{n}\lambda) \rightarrow \infty.$$

则 (A.3) 中的第二项占主导地位, 因此我们证明了 (A.1), 或者等价地证明了, 当 $n \rightarrow \infty$ 时, $\hat{\pi}_k = 0, k = 1, \dots, K - q$ 依概率趋于1.

定理2 在(8)式中, 若 $\sqrt{n}\lambda \rightarrow C, \varepsilon = o(1/\sqrt{n} \log n)$, 其中 C 是一个常数, 则存在 $l_{p,1}$ 的一个局部最大值点 $(\hat{\theta}, \hat{\eta})$, 使得 $\hat{\theta} = O_P(1/\sqrt{n}), \hat{q} \xrightarrow{P} q$, 其中 $\hat{q}$ 为估计的子群个数, q 为真实的子群个数.

证明: 与定理1的证明类似, 首先证明当 $\lambda = C/\sqrt{n}$ 时, 存在一个最大值点 $(\hat{\theta}, \hat{\eta})$ 使得 $\hat{\theta} = O_P(1/\sqrt{n})$. 只需要证明, 对于一个大常数 $C_1, l_{p,1}(\theta, \eta) < l_{p,1}(0, \eta)$, 其中 $\theta = C_1/\sqrt{n}, \eta$ 在一个局部紧集里. 设 $\theta = C_1/\sqrt{n}$, 则

$$\begin{aligned}
& l_{p,1}(\theta, \eta) - l_{p,1}(0, \eta) \\
&= \sum_{i=1}^n \log f(Y_i, M_i \mid A_i, X_i; \theta, \eta) - n\lambda D_f \sum_{k=1}^K [\log(\varepsilon + \pi_k) - \log(\varepsilon)] \\
&\quad - \left(\sum_{i=1}^n \log g_0(Y_i, M_i \mid A_i, X_i) - n\lambda D_f \sum_{l=1}^q [\log(\varepsilon + \pi_l^0) - \log(\varepsilon)] \right) \\
&\leq \sum_{i=1}^n [\log f(Y_i, M_i \mid A_i, X_i; \theta, \eta) - \log g_0(Y_i, M_i \mid A_i, X_i)] \\
&\quad - n\lambda D_f \sum_{k=K-q+1}^K [\log(\varepsilon + \pi_k) - \log(\varepsilon + \pi_{k-K+q}^0)] \\
&\triangleq II_1 + II_2.
\end{aligned}$$

对于 II_2 , 因为 $\theta = C_1/\sqrt{n}$ 和 $\rho_l, l = 1, \dots, q$ 上的约束条件, 当 $k > K - q$ 时, 有 $|\pi_k - \pi_{k-K+q}^0| \leq C_1/\sqrt{n}$. 由惩罚函数的性质可以得到

$$\begin{aligned}
|II_2| &= \left| -n\lambda D_f \sum_{k=K-q+1}^K [\log(\varepsilon + \pi_k) - \log(\varepsilon + \pi_{k-K+q}^0)] \right| \\
&= \left| n\lambda D_f \sum_{k=K-q+1}^K \left[\frac{\pi_k - \pi_{k-K+q}^0}{\varepsilon + \pi_{k-K+q}^0} (1 + o(1)) \right] \right| \\
&= O(\sqrt{n}) \cdot \frac{qC_1}{\sqrt{n}} (1 + o(1)) \\
&= O(C_1).
\end{aligned}$$

对于 II_1 , 类似于定理1的证明, 有

$$II_1 = \frac{C_1}{\sqrt{n}} \cdot O_P(\sqrt{n}) - \frac{C_1^2}{n} \cdot O_P(n).$$

当 C_1 足够大时, II_1 的第二项在惩罚似然函数中占主导地位. 有 $l_{p,1}(\theta, \eta) - l_{p,1}(0, \eta) < 0$ 依概率趋于1. 因此存在一个最大值点 $(\hat{\theta}, \hat{\eta})$ 使得 $\hat{\theta} = O_P(1/\sqrt{n})$ 依概率趋于1成立.

接下来证明存在一个满足 $\hat{\theta} = O_P(1/\sqrt{n})$ 的最大值点 $(\hat{\theta}, \hat{\eta})$, 使得 $\hat{q} = q$

或 $\hat{\pi}_k = 0, k = 1, \dots, K - q$ 依概率趋于1成立.

首先证明对于任意一个 $l_p(\theta, \eta)$ 的极大值点 (θ^*, η^*) 满足 $|\theta^*| \leq C_1/\sqrt{n}$, 如果存在 $k \leq K - q$ 使得 $C_1/\sqrt{n} \geq \pi_k^* \geq 1/(\sqrt{n} \log n)$, 那么一定存在 $l_{p,1}(\theta, \eta)$ 的另一个极大值点 θ , 满足 $|\theta| \leq C_1/\sqrt{n}$. 这意味着在 $|\theta| \leq C_1/\sqrt{n}$ 的紧域中, $l_{p,1}(\theta, \eta)$ 的最大值点应该满足对于任意 $k < K - q + 1, \pi_k < 1/(\sqrt{n} \log n)$. 因此, 它相当于证明, 对于任何这类 $l_p(\theta^*, \eta^*)$ 的极大值点, $|\theta^*| \leq C_1/\sqrt{n}$, 总是有 $l_{p,1}(\theta^*, \eta^*) < l_{p,1}(0, \eta^*)$ 依概率趋于1. 与上面的分析类似, 有

$$\begin{aligned}
& l_{p,1}(\theta^*, \eta^*) - l_{p,1}(0, \eta^*) \\
&= \sum_{i=1}^n [\log f(Y_i, M_i \mid A_i, X_i; \theta^*, \eta^*) - \log g_0(Y_i, M_i \mid A_i, X_i)] \\
&\quad - n\lambda D_f \sum_{k=K-q+1}^K [\log(\varepsilon + \pi_k^*) - \log(\varepsilon + \pi_{k-K+q}^0)] \\
&\quad - n\lambda D_f \sum_{k=1}^{K-q} \log \frac{\varepsilon + \pi_k^*}{\varepsilon} \\
&\triangleq III_1 + III_2 + III_3.
\end{aligned}$$

如前所述, 得到 $III_1 + III_2 = -O_p(C_1^2)$. 对于 III_3 , 因为 $\varepsilon = o(1/(\sqrt{n} \log n))$, 有

$$|III_3| = O(n \cdot C/\sqrt{n}) \log \frac{\varepsilon + \pi_k^*}{\varepsilon} = O(\sqrt{n}).$$

注意到 III_3 总是负的, 在 III_1 和 III_2 中占主导地位, 因此有 $l_{p,1}(\theta^*, \eta^*) < l_{p,1}(0, \eta^*)$.

下面证明存在一个满足 $\hat{\theta} = O_P(1/\sqrt{n})$ 对应的最大值点 $(\hat{\theta}, \hat{\eta})$, 有 $\hat{q} = q$ 或 $\hat{\pi}_k = 0$, $k = 1, \dots, K - q$ 依概率趋于1成立. 对于约束 $\sum_{k=1}^K \hat{\pi}_k = 1$, 考虑拉格朗日乘子 ω , 只需要证明

$$\left. \frac{\partial l^*(\theta)}{\partial \pi_k} \right|_{\pi_k = \hat{\pi}_k} < 0, \quad \hat{\pi}_k < \frac{1}{\sqrt{n} \log n} \quad (\text{A.4})$$

对于 $(\hat{\theta}, \hat{\eta})$ 依概率趋于1成立, 其中 $l^*(\theta) = l_{p,1}(\theta) - \omega(\sum_{k=1}^K \hat{\pi}_k - 1)$.

为了证明 (A.4) 成立, 首先考虑当 $k > K - q$ 时关于 π_k 的偏导数, 令

$$\left. \frac{\partial l^*(\theta)}{\partial \pi_k} \right|_{\pi_k = \hat{\pi}_k} = \sum_{i=1}^n \frac{\phi(\alpha_k)}{\sum_{k=1}^K \hat{\pi}_k \phi(\alpha_k)} - n\lambda D_f \frac{1}{\varepsilon + \hat{\pi}_k} - \omega = 0. \quad (\text{A.5})$$

很明显, 根据大数定律, (A.5) 的第一项是 $O_P(n)$. 如果 $k > K - q$, $\hat{\theta} = O_P(1/\sqrt{n})$, 很容易知道 $\hat{\pi}_k = \pi_{k-K+q}^0 + O_P(1/\sqrt{n}) > \frac{1}{2} \min(\pi_1^0, \dots, \pi_q^0)$, 那么第二项为 $O_P(n\lambda) = o_P(n)$.

因此 $\omega = O_P(n)$.

接下来, 考虑 $k \leq K - q$, $\hat{\pi}_k < 1/(\sqrt{n} \log n)$ 时关于 π_k 的偏导数

$$\left. \frac{\partial l^*(\theta)}{\partial \pi_k} \right|_{\pi_k = \hat{\pi}_k} = \sum_{i=1}^n \frac{\phi(\alpha_k)}{\sum_{k=1}^K \hat{\pi}_k \phi(\alpha_i)} - n\lambda D_f \frac{1}{\varepsilon + \hat{\pi}_k} - \omega. \quad (\text{A.6})$$

(A.6) 中的第一项和第三项是 $O_P(n)$. 对于第二项, 因为 $\hat{\pi}_k = O_P(1/(\sqrt{n} \log n))$,

$\lambda = C/\sqrt{n}$, $\varepsilon = o(1/(\sqrt{n} \log n))$, 有

$$\frac{n\lambda D_f / (\varepsilon + \hat{\pi}_k)}{n} = \lambda D_f \frac{1}{\varepsilon + \hat{\pi}_k} = O_p(\lambda \cdot \sqrt{n} \log n) \rightarrow \infty.$$

则 (A.6) 中的第二项占主导地位, 因此证明了 (A.4), 或者等价证明了, 当 $n \rightarrow \infty$ 时,

$\hat{\pi}_k = 0, k = 1, \dots, K - q$ 依概率趋于 1 成立.

定理3 在 (9) 式中, 若 $\sqrt{n}\lambda \rightarrow \infty, \lambda \rightarrow 0$ 且 $\varepsilon = o(1/\sqrt{n})$, $\hat{q}_{\lambda_{BIC}}$ 表示 (9) 中使用调整参数 λ_{BIC} 选择子群的个数, λ_{BIC} 表示 BIC 方法选择的调整参数 λ, q 表示真实的子群个数, 有 $P(\hat{q}_{\lambda_{BIC}} = q) \rightarrow 1$.

证明: 首先, 给定 $\lambda^* = \sqrt{\log n/n}$, 通过定理 1 可知, 存在一个最大值使得 $\hat{q}$ 依概率趋于 q , $\hat{\pi}_{K-q+k} (k = 1, \dots, q)$ 是 $\pi_l^0 (l = 1, \dots, q)$ 的相合估计. 因此, 有

$$l_{p,2}(\hat{\theta}_{\lambda^*}) = l(\hat{\theta}_{\lambda^*}) - n\lambda^* D_f \cdot q \cdot \log \left(\frac{\varepsilon + \frac{(a+1)\lambda^*}{2}}{\varepsilon} \right),$$

其中 $\hat{\theta}_{\lambda^*}$ 为多元高斯混合模型参数的估计量. 另一方面, 当 q 是已知时, 最大似然估计 $\hat{\theta}_{MLE}$ 是相合的. 因此, 有

$$\begin{aligned} l_{p,2}(\hat{\theta}_{MLE}) &= l(\hat{\theta}_{MLE}) - n\lambda^* D_f \sum_{l=1}^q [\log(\varepsilon + P_{\lambda^*}(\hat{\pi}_{l,MLE})) - \log(\varepsilon)] \\ &= l(\hat{\theta}_{MLE}) - n\lambda^* D_f \cdot q \cdot \log \left(\frac{\varepsilon + \frac{(a+1)\lambda^*}{2}}{\varepsilon} \right) \\ &\geq l_{p,2}(\hat{\theta}_{\lambda^*}), \end{aligned}$$

式中, $\hat{\pi}_{l,MLE}$ 是 $\pi_l^0, l = 1, \dots, q$ 的极大似然估计, 然后由 $l_{p,2}(\hat{\theta}_{\lambda^*})$ 的定义,

当 $\lambda^* = \sqrt{\log n/n}$ 时, 我们有 Oracle 性质, 即惩罚似然估计 $\hat{\theta}_{\lambda^*}$ 依概率趋于 $\hat{\theta}_{MLE}$.

接下来我们确定了两种不同的情况, 即欠拟合模型和过拟合模型.

情况一: 欠拟合模型即 $\hat{q}_{\lambda} < q$, 其中 $\hat{q}_{\lambda}$ 为估计的子群个数, q 为真实的子群个数.

由 BIC 的定义, 有

$$BIC_{\lambda} = l(\hat{\theta}_{\lambda}) - \frac{1}{2}\hat{q}_{\lambda} D_f \log n \leq l(\hat{\theta}_{\hat{q}_{\lambda},MLE}) - \frac{1}{2}\hat{q}_{\lambda} D_f \log n,$$

其中 $\hat{\theta}_{\hat{q}_{\lambda},MLE}$ 是有限高斯混合模型当子群个数为 $\hat{q}_{\lambda}$ 的极大似然估计. 由大数定律及计算, 可以证明

$$\begin{aligned} &\frac{1}{n} \{l(\hat{\theta}_{\hat{q}_{\lambda},MLE}) - l(\hat{\theta}_{q,MLE})\} \\ &\xrightarrow{P} \frac{1}{n} \{l(\hat{\theta}_{\hat{q}_{\lambda},MLE}) - l(\hat{\theta}_{\lambda^*})\} = -\frac{1}{n} \{l(\hat{\theta}_{\lambda^*}) - l(\hat{\theta}_{\hat{q}_{\lambda},MLE})\} \hat{=} -\frac{1}{n} \sum_{i=1}^n \log \frac{g_{i0}}{g_{i\lambda}} \\ &\xrightarrow{P} -E \log \frac{g_{i0}}{g_{i\lambda}} = -K(g_{i0}, g_{i\lambda}), \end{aligned}$$

其中 $K(g_{i0}, g_{i\lambda})$ 为 g_{i0} 和 $g_{i\lambda}$ 之间的 KL 距离, 则有

$$\begin{aligned}
& BIC_{\lambda} - BIC_{\lambda^*} \\
& \leq l(\hat{\theta}_{\hat{q}_{\lambda}, MLE}) - l(\hat{\theta}_{\hat{q}_{\lambda^*}, MLE}) - \frac{1}{2}\hat{q}_{\lambda}D_f \log n + \frac{1}{2}\hat{q}_{\lambda^*}D_f \log n \\
& = l(\hat{\theta}_{\hat{q}_{\lambda}, MLE}) - l(\hat{\theta}_{q, MLE}) - \frac{1}{2}\hat{q}_{\lambda}D_f \log n + \frac{1}{2}qD_f \log n \\
& = -nK(g_{i0}, g_{i\lambda})(1 + o_P(1)) + \frac{1}{2}(q - \hat{q}_{\lambda})D_f \log n \\
& \leq -n \inf_{\lambda} K(g_{i0}, g_{i\lambda})(1 + o_P(1)) + \frac{1}{2}(q - \hat{q}_{\lambda})D_f \log n \\
& < 0.
\end{aligned}$$

这意味着

$$P\left(\sup_{\lambda: \{\hat{q}_{\lambda} < q\}} BIC_{\lambda} > BIC_{\lambda^*}\right) \rightarrow 0. \quad (\text{A.7})$$

情况二：过拟合模型即 $\hat{q}_{\lambda} > q$ ，其中 $\hat{q}_{\lambda}$ 为估计的子群个数， q 为真实的子群个数。

如果可以证明

$$l(\hat{\theta}_{\hat{q}_{\lambda}, MLE}) - l(\hat{\theta}_{q, MLE}) = O_P(1), \quad (\text{A.8})$$

则有

$$\begin{aligned}
& BIC_{\lambda} - BIC_{\lambda^*} \\
& \leq l(\hat{\theta}_{\hat{q}_{\lambda}, MLE}) - l(\hat{\theta}_{\hat{q}_{\lambda^*}, MLE}) - \frac{1}{2}\hat{q}_{\lambda}D_f \log n + \frac{1}{2}\hat{q}_{\lambda^*}D_f \log n \\
& = l(\hat{\theta}_{\hat{q}_{\lambda}, MLE}) - l(\hat{\theta}_{q, MLE}) - \frac{1}{2}\hat{q}_{\lambda}D_f \log n + \frac{1}{2}qD_f \log n \\
& = O_P(1) + \frac{1}{2}(q - \hat{q}_{\lambda})D_f \log n \\
& < 0.
\end{aligned}$$

这意味着

$$P\left(\sup_{\lambda: \{\hat{q}_{\lambda} > q\}} BIC_{\lambda} > BIC_{\lambda^*}\right) \rightarrow 0. \quad (\text{A.9})$$

为了证明 (A.8)，注意给定的有界 $\hat{q}_{\lambda}$ ，其中 $\hat{q}_{\lambda} > q$ ，由大数定律，有

$$\begin{aligned}
& \sup_{(\theta, \eta) \in \Theta} \left| \frac{1}{n} l(\theta) - E \log f(Y, M \mid A, \mathbf{X}; \theta, \eta) \right| \\
& = \sup_{(\theta, \eta) \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \log f(Y_i, M_i \mid A_i, \mathbf{X}_i; \theta, \eta) - E \log f(Y, M \mid A, \mathbf{X}; \theta, \eta) \right| \\
& \xrightarrow{P} 0,
\end{aligned}$$

其中 Θ 为紧参数空间，与定理 1 的证明类似，可知在 Θ 中有一个最大值点。

注意到对于任何 η ，当 $\theta > \varepsilon$ ， ε 是一个小常数时，由似然函数的性质我们有

$$E \log f(Y, M \mid A, \mathbf{X}; \theta, \eta) < E \log f(Y, M \mid A, \mathbf{X}; 0, \eta).$$

与 van der Vaart^[48]中定理 5.7 的证明类似，可以证明 θ 的极大似然估计依概率收敛到 0。

接下来，和定理 1 对 I_1 的证明类似，因为

$$E\left(\frac{f'(Y, M | A, \mathbf{X}; 0, \boldsymbol{\eta})}{g_0(Y, M | A, \mathbf{X})}\right) = 0,$$

有

$$\begin{aligned} & E(\log f(Y, M | A, \mathbf{X}; \theta, \boldsymbol{\eta}) - \log f(Y, M | A, \mathbf{X}; 0, \boldsymbol{\eta})) \\ &= -\frac{1}{2}\theta^2 E\left(\frac{f'(Y, M | A, \mathbf{X}; 0, \boldsymbol{\eta})}{g_0(Y, M | A, \mathbf{X})}\right)^2 (1 + o_P(1)). \end{aligned}$$

根据条件, $\boldsymbol{\eta}$ 在紧空间里, $\left(\frac{f'(Y, M | A, \mathbf{X}; 0, \boldsymbol{\eta})}{g_0(Y, M | A, \mathbf{X})}\right)^2$ 是 $\boldsymbol{\eta}$ 的连续函数. 因此, $E\left(\frac{f'(Y, M | A, \mathbf{X}; 0, \boldsymbol{\eta})}{g_0(Y, M | A, \mathbf{X})}\right)^2$

可以被一个常数所限制, 另外, 对于一个充分小的 $\theta \rightarrow 0$, 有

$$E(\log f(Y, M | A, \mathbf{X}; \theta, \boldsymbol{\eta}) - \log f(Y, M | A, \mathbf{X}; 0, \boldsymbol{\eta})) \leq -C\theta^2.$$

令 $G_n(\theta, \boldsymbol{\eta}) = 1/\sqrt{n} \sum_{i=1}^n \{\log f(Y_i, M_i | A_i, \mathbf{X}_i; \theta, \boldsymbol{\eta}) - E \log f(Y, M | A, \mathbf{X}; \theta, \boldsymbol{\eta})\}$. 对于 $|\theta| < \delta$, 其中 δ 是一个充分小的值, 考虑 $G_n(\theta, \boldsymbol{\eta}) - G_n(0, \boldsymbol{\eta})$, 根据定理 1 的证明, 有

$$\begin{aligned} & G_n(\theta, \boldsymbol{\eta}) - G_n(0, \boldsymbol{\eta}) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \{\log f(Y_i, M_i | A_i, \mathbf{X}_i; \theta, \boldsymbol{\eta}) - \log f(Y_i, M_i | A_i, \mathbf{X}_i; 0, \boldsymbol{\eta})\} \\ &\quad - \sqrt{n} E \{\log f(Y, M | A, \mathbf{X}; \theta, \boldsymbol{\eta}) - \log f(Y, M | A, \mathbf{X}; 0, \boldsymbol{\eta})\} \\ &= \frac{1}{\sqrt{n}} \left\{ \sum_{i=1}^n \theta \frac{f'(Y_i, M_i | A_i, \mathbf{X}_i; 0, \boldsymbol{\eta})}{g_0(Y_i, M_i | A_i, \mathbf{X}_i)} - \frac{1}{2} \sum_{i=1}^n \theta^2 \left(\frac{f'(Y_i, M_i | A_i, \mathbf{X}_i; 0, \boldsymbol{\eta})}{g_0(Y_i, M_i | A_i, \mathbf{X}_i)} \right)^2 \right\} \\ &\quad \times (1 + o_P(1)) + \frac{\sqrt{n}}{2} \theta^2 E \left(\frac{f'(Y, M | A, \mathbf{X}; 0, \boldsymbol{\eta})}{g_0(Y, M | A, \mathbf{X})} \right)^2 (1 + o_P(1)). \end{aligned}$$

对于 $\delta = C/\sqrt{n}$, C 为足够大的常数, 有

$$E \sup |G_n(\theta, \boldsymbol{\eta}) - G_n(0, \boldsymbol{\eta})| < C_1 \delta + \sqrt{n} C_2 \delta^2 + \sqrt{n} C_3 \delta^2 = O(\sqrt{n} \delta^2).$$

由 van der Vaart^[48]中定理 5.55, 有

$$\hat{\theta}_{\hat{q}_{\lambda, MLE}} = O_P(1/\sqrt{n}).$$

另外通过计算, 可以证明

$$\begin{aligned}
& l(\hat{\theta}_{\hat{q}_{\lambda}, MLE}) - l(\hat{\theta}_{q, MLE}) \\
&= \sqrt{n} \left\{ G_n(\hat{\theta}_{\hat{q}_{\lambda}, MLE}, \hat{\eta}_{\hat{q}_{\lambda}, MLE}) - G_n(\hat{\theta}_{q, MLE}, \hat{\eta}_{q, MLE}) \right\} \\
& \quad + nE \left\{ \log f(Y, M \mid A, X; \hat{\theta}_{\hat{q}_{\lambda}, MLE}, \hat{\eta}_{\hat{q}_{\lambda}, MLE}) \right. \\
& \quad \left. - \log f(Y, M \mid A, X; \hat{\theta}_{q, MLE}, \hat{\eta}_{q, MLE}) \right\} \\
&= \sqrt{n} \left\{ G_n(\hat{\theta}_{\hat{q}_{\lambda}, MLE}, \hat{\eta}_{\hat{q}_{\lambda}, MLE}) - G_n(0, \hat{\eta}_{\hat{q}_{\lambda}, MLE}) + G_n(0, \hat{\eta}_{\hat{q}_{\lambda}, MLE}) - G_n(0, \eta_0) \right. \\
& \quad \left. + G_n(0, \eta_0) - G_n(0, \hat{\eta}_{q, MLE}) + G_n(0, \hat{\eta}_{q, MLE}) - G_n(\hat{\theta}_{q, MLE}, \hat{\eta}_{q, MLE}) \right\} \\
& \quad + n \left\{ E(\log f(Y, M \mid A, X; \hat{\theta}_{\hat{q}_{\lambda}, MLE}, \hat{\eta}_{\hat{q}_{\lambda}, MLE})) \right. \\
& \quad - E(\log f(Y, M \mid A, X; 0, \hat{\eta}_{\hat{q}_{\lambda}, MLE})) \\
& \quad + E(\log f(Y, M \mid A, X; 0, \hat{\eta}_{\hat{q}_{\lambda}, MLE})) - E(\log f(Y, M \mid A, X; 0, \eta_0)) \\
& \quad + E(\log f(Y, M \mid A, X; 0, \eta_0)) - E(\log f(Y, M \mid A, X; 0, \hat{\eta}_{q, MLE})) \\
& \quad \left. + E(\log f(Y, M \mid A, X; 0, \hat{\eta}_{q, MLE})) - E(\log f(Y, M \mid A, X; \hat{\theta}_{q, MLE}, \hat{\eta}_{q, MLE})) \right\} \\
&= \sqrt{n} O_P(\sqrt{n} \delta^2) + n O_P(\delta^2) \\
&= O_P(1).
\end{aligned}$$

因此 (A.8) 得证, 从而 (A.9) 得证, 由 (A.7) 和 (A.9), 可知 $P(\hat{q}_{\lambda_{BIC}} = q) \rightarrow 1$, 这里 $\hat{q}_{\lambda_{BIC}}$ 表示 BIC 方法估计的子群个数.