高维Logistic回归模型的全局和同时假设检验

——统计计算课程项目报告

小组成员：付芷睿 2018302010035

韩硕 2018302060281

林宇帆 2018302010027

孙雪倩 2018302010111

1 研究背景

logistic回归模型在遗传学、金融和商业分析中已经得到了广泛的应用。在统计课程上我们已经对低维度的logistic回归模型有所接触，但是在实际应用中，模型的协变量的数量常常超过观测值的数量，成为了高维模型。在这种高维条件下，估计、假设检验和置信区间的构造等统计问题变得比经典的低维条件更加困难。

这篇论文考虑了在高维条件下的logistic回归模型：

$$\log(\frac{\pi_i}{1-\pi_i}) = X_i^T\beta, \qquad i = 1, \ldots, n, \tag{1}$$

其中，$\beta \in \mathbb{R}^p$ 是回归系数向量。观测值是相互独立的样本，$Z_i = (y_i, X_i)$，$i = 1, \ldots, n$，并且我们假定$y_i|X_i \sim \text{Bernoulli}(\pi_i)$且相互独立，$i = 1, \ldots, n$。

1.1 全局和同时性假设检验

在高维条件下的logistic回归中，通常要解决两个问题：

（1）协变量与结果之间是否有相关性；

（2）如果有，哪些协变量与结果相关。

要解决第一个问题，通常检验全局零假设$H_0:\beta=0$。为了解决第二个问题，通常同时检验零假设$H_{0,i}:\beta_i=0$，$i=1,\ldots,p$。

以前的学者们已经研究了很多与高维logistic回归模型相关的问题，比如对高维logistic回归的估计、Lipschitz损失函数下的高维广义线性模型（GLMs）、logistic回归的组Lasso、利用限制强凸性得到l_1正则极大似然估计在GLMs下的收敛速度等等，但是对于高维Logistic回归的假设检验与置信区间的研究却相对很少。

这篇论文从解决以上两个问题出发，分别考虑了对高维logistic回归模型的全局检验和多重检验。

首先，在全局检验中，统计量被构造为单个系数的标准化统计量平方的最大值，该统计量的构造基于两步标准化程序。第一步，利用广义低维投影（LDP）方法修正logistic Lasso估计量的偏差，第二步，通过估计量的标准误差将得到的近似无偏估计量归一化。

接着，在大规模的多重检验中，文章提出了一个数据驱动型检验程序，并证明其可以渐近地控制错误发现率(FDR)和错误发现变量(FDV)，最后通过与现有的其他方法进行比较，通过大量模拟来评估所提出的检验程序的数值性能。

这篇论文的贡献体现在：为高维logistic回归中的全局检验和大规模同时性检验提出了新程序。特别地，为了体现维数p可以远大于样本量n，文章在全局检验中设置了$\log p=O(n^{c_1})$，在多重检验中设置了$p=O(n_2^c)$，其中常数$c_1,c_2>0$。

2 模型描述

2.1 全局假设检验

文章首先考虑了logistic回归模型下的全局零假设检验：

$$H_0:\beta=0 \quad \text{versus} \quad H_1:\beta\neq 0, \tag{2}$$

这样的全局假设检验问题常用来检验协变量和结果之间是否相关。

全局假设检验过程中，首先需要建立一个修正偏差估计量，该估计量又建立于一个正则估计量，如l_1正则化的M-估计量。对于高维logistic回归，l_1正则化的M-估计量定义为：

$$\widehat{\beta}=\underset{\beta}{\operatorname{argmin}}\{\frac{1}{n}\Sigma_{i=1}^{n}[-y_i\beta^TX_i+log(1+e^{\beta^TX_i})]+\lambda||\beta||_1\} \tag{3}$$

2.1.1 利用广义低维投影（LDP）方法建立检验统计量

2.1中由logistic模型直接得到的估计量$\hat{\beta}$是有偏的，因此为了纠正偏差，下面利用广义低维投影（LDP）方法建立修正偏差统计量$\tilde{\beta}$，或称为广义LDP检验统计量。

设X为设计矩阵，其第i行表示为X_i。则由（1）定义的logistic回归模型可以表示为：

$$y_i=f(\beta^TX_i)+\epsilon_i, \tag{4}$$

其中$f(u)=e^u/(1+e^u)$，ϵ_i是误差项。为了纠正（3）式中初始估计量$\widehat{\beta}$的偏差，考虑$f(u_i)$在$\hat{u}_i$处的泰勒展开式：

$$f(u_i)=f(\hat{u}_i)+\dot{f}(\hat{u}_i)(u_i-\hat{u}_i)+Re_i \tag{5}$$

其中$u_i=\beta^TX_i$，$\hat{u}_i=\hat{\beta}^TX_i$，$Re_i$是余项。将（5）式代入回归模型（4）可得：

$$y_i-f(\hat{u_i})+\dot{f}(\hat{u_i})X_i^T\hat{\beta}=\dot{f}(\hat{u}_i)X_i^T\beta+(Re_i+\epsilon_i). \tag{6}$$

因此，通过改写logistic回归模型，可以将上式等号左边的$y_i-f(\hat{u_i})+\dot{f}(\hat{u_i})X_i^T\hat{\beta}$看作新的响应变量，其中$\dot{f}(\hat{u}_i)X_i$是新的协变量，$Re_i+\epsilon_i$是噪声。那么，$\beta$可以看作为这个近似线性模型的回归变量。

从而，修正偏差估计量，或者称为广义LDP估计量$\tilde{\beta}_j$可以被定义为：

$$\tilde{\beta}_j=\hat{\beta}_j+\frac{\Sigma_{i=1}^{n}v_{ij}(y_i-f(\hat{\beta}^TX_i))}{\Sigma_{i=1}^{n}v_{ij}\dot{f}(\hat{\beta}^TX_i)X_{ij}},\quad j=1,\dots,p, \tag{7}$$

其中，X_{ij}是X_i的第j个元素，$v_j=(v_{1j},v_{2j},\dots,v_{nj})^T$是得分向量。

要计算修正偏差估计量$\tilde{\beta}_j$，就需要计算得分向量v_j，因此考虑下面的协变量之间的节点向回归：

$$x_j=X_{-j}\gamma_j+\eta_j,\quad j=1,\dots,p, \tag{8}$$

其中$\gamma_j=\underset{\gamma\in\mathbb{R}^{p-1}}{\operatorname{argmin}}E[||x_j-X_{-j}\gamma||_2^2]$，$\eta_j$是误差项。

在实际中，为了得到误差项η_j的估计，文章利用节点向拉索（node-wise lasso）算法：

$$v_j(\lambda)=\hat{W}^{-1}\hat{\eta}_j(\lambda),\quad \hat{\eta}_j(\lambda)=x_j-X_{-j}\hat{\gamma}_j(\lambda),\tag{9}$$
$$\hat{\gamma}_j(\lambda)=\underset{b}{\operatorname{argmin}}\{\frac{||x_j-X_{-j}b||_2^2}{2n}+\lambda||b||_1\}。$$

显然，$v_j(\lambda)$依赖于调节参数λ。其中，$\hat{W}=diag(\dot{f}(\hat{u}_1),\ldots,\dot{f}(\hat{u}_n))$，$\dot{f}(x)=df(x)/dx$。再定义下列数量：

$$\zeta_j(\lambda)=\max_{k\neq j}\frac{|\langle v_j(\lambda),x_k\rangle_n|}{||v_j(\lambda)||_n},\quad \tau_j(\lambda)=\frac{||v_j(\lambda)||_n}{|\langle v_j(\lambda),x_j\rangle_n|}\tag{10}$$

因此，由（10）式中的$\zeta_j(\lambda)$和$\tau_j(\lambda)$就可以决定调节参数λ，进而得到$v_j(\lambda)$与误差项η_j的估计，最后得到修正偏差估计量$\tilde{\beta}_j$。

上述整个算法归纳在表1中。

表1.　计算V_J值的节点向拉索算法(9)
输入：　ζ_j 的上限ζ_j^*，默认值为$\zeta^*=\sqrt{2\log p}$，调节参数$k_0\in[0,1]$，$k_1\in(0,1]$；
Step1：　如果对所有$\lambda>0$都有$\zeta_j(\lambda)>\zeta_j^*$，令$\zeta_j^*=(1+k_1)\inf_{\lambda>0}\zeta_j(\lambda)$； $\lambda\leftarrow\max\{\lambda:\zeta_j(\lambda)\leq\zeta_j^*\}$，$\zeta_j^*\leftarrow\zeta_j(\lambda)$，$\tau_j^*\leftarrow\tau_j(\lambda)$；
Step2：　$\lambda_j\leftarrow\min\{\lambda:\tau_j(\lambda)\leq(1+k_0)\tau_j^*\}$；$v_j\leftarrow v_j(\lambda_i)$，$\tau_j\leftarrow\tau_j(\lambda_j)$，$\zeta_j\leftarrow\zeta_j(\lambda_j)$
输出：　λ，v_j，τ_j，ζ_j

根据以上步骤就可以得到$\tilde{\beta}_j$和τ_j，进而可以给出下面的标准检验统计量M_j：

$$M_j=\tilde{\beta}_j/\tau_j,\quad j=1,\ldots,p\tag{11}$$

给定一个阈值t，如果$|M_j|\geq t$，则拒绝$\beta_j=0$，$j=1,\ldots,p$。

因此，全局检验统计量可以被定义为：

$$M_n=\max_{1\leq j\leq p}M_j^2\tag{12}$$

本节建立全局检验统计量的思维流程如图1所示。

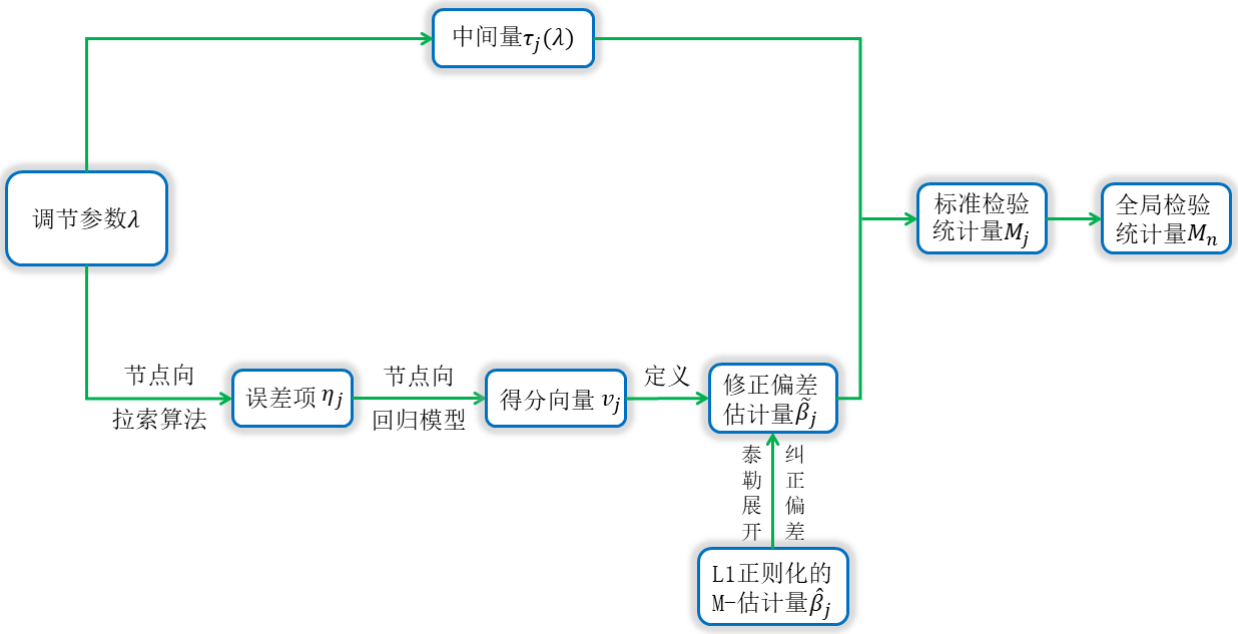

图1 建立全局检验统计量的思维流程图

2.1.2 渐近零分布

可以证明，在高斯或者有界条件下，M_n的渐近零分布是一个甘贝尔分布（Gumbel distribution）。

那么基于极限零分布，渐近α水平检验可以被定义为：

$$\Phi_\alpha(M_n) = I\{M_n \geq 2\log p - \log\log p + q_\alpha\}, \tag{13}$$

其中q_α是Gumbel分布的$1-\alpha$分位数，其累积分布函数为$exp(-\frac{1}{\sqrt{\pi}}exp(-x/2))$，即

$$q_\alpha = -\log(\pi) - 2\log\log(1-\alpha)^{-1}. \tag{14}$$

因此，当且仅当$\Phi_\alpha(M_n) = 1$时拒绝零假设H_0。

2.2 大规模多重检验

用β表示模型中的真系数向量，$H_0 = \{j : \beta_j = 0, j = 1, \ldots, p\}$，$H_1 = \{j : \beta_j \neq 0, j = 1, \ldots, p\}$。为了识别$H_1$中的索引，文章考虑了同时检验下面的这些零假设：

$$H_{0,j} : \beta_j = 0 \quad \text{versus} \quad H_{1,j} : \beta_j \neq 0, \quad 1 \leq j \leq p。 \tag{15}$$

除了尽可能多地识别非零的β_j外，为了获得有实际意义的结果，作者还希望控制FDR以及错误发现率（FDP），即FDV的数量。

2.2.1 建立多重检验过程

在上一节中，定义了标准统计量$M_j = \tilde{\beta}_j/\tau_j$，$j = 1, \ldots, p$。对于给定的阈值$t > 0$，若$|M_j| \geq t$，则拒绝假设$H_{0,j} : \beta_j = 0$，对每个$j = 1, \ldots p$。因此对于每个$t$，可以定义

$$\begin{aligned} \mathrm{FDP}_\theta(t) = \frac{\Sigma_{j \in H_0} I\{|M_j| \geq t\}}{\max\{\Sigma_{j=1}^p I\{|M_j| \geq t\}, 1\}}, \\ \mathrm{FDR}_\theta(t) = E_\theta[\mathrm{FDP}(t)], \end{aligned} \tag{16}$$

错误发现变量（FDV）的期望值为$\mathrm{FDV}_\theta(t) = E_\theta[\Sigma_{j \in H_0} I\{|M_j| \geq t\}]$。

2.2.1.1 控制FDR/FDP的程序

为了将FDR/FDP控制在预先设定的水平$0 < \alpha < 1$，可以设置阈值为：

$$\hat{t} = inf\{0 \leq t \leq b_p : \frac{\Sigma_{j \in H_0} I\{|M_j| \geq t\}}{\max\{\Sigma_{j=1} p I\{|M_j| \geq t\}, 1\}} \leq \alpha\} \tag{17}$$

其中，b_p的值稍后设置。

通常来说，$\tilde{t}_1$是未知的，需要被估计，因为它依赖于真零值H_0。设$G_0(t)$为在阈值水平t处被该过程拒绝的零错误值占所有真零值的比例，即

$$G_0(t) = \frac{1}{p_o} \Sigma_{j \in H_0} I\{|M_j| \geq t\} \tag{18}$$

其中，$p_0 = |H_0|$。在实际中，真正的备择假设是稀疏的。如果样本量大，则可以使用正态分布的尾部$G(t) = 2 - 2\Phi(t)$去近似$G_0(t)$。事实上，可以证明，对$b_p = \sqrt{2\log p - 2\log\log p}$，$sup_{0 \leq t \leq b_p} |\frac{G_0(t)}{G(t)} - 1| \to 0$，$(n, p) \to \infty$。这些过程可以总结为以下的可以控制FDR和FDP的logistic多重检验(LMT)程序。

过程1（LMT） 令$0 < \alpha < 1$，$b_p = \sqrt{2\log p - 2\log\log p}$，定义

$$\hat{t} = inf\{0 \leq t \leq b_p : \frac{pG(t)}{\max\{\Sigma_{j=1} p I\{|M_j| \geq t\}, 1\}} \leq \alpha\} \tag{19}$$

如果式（18）中的$\hat{t}$不存在，就令$\hat{t} = \sqrt{2\log p}$。当$|M_j| \geq \hat{t}$时拒绝$H_{0,j}$。

2.2.1.2 控制FDV的程序

对于大规模检测，有时直接控制FDV的数量，而不是控制不那么严格的FDR/FDP，尤其是在样本容量较小的情况下。事实上，FDV控制可以通过对上面介绍的FDP控制程序进行适当修改来实现。具体而言，论文提出了以下控制FDV(或FWER)的logistic多重测试(LMTV)程序。

过程2（LMT_V） 对于给定的FDV容忍值$r < p$（或者FWER的预期水平$0 < r < 1$），令$\hat{t}_{FDV} = G^{-1}(r/p)$。当$|M_j| \geq \hat{t}_{FDV}$就拒绝$H_{0,j}$。

3 模拟结果与分析

在本节中，我们按照论文的方法对于前面提出的检验方法进行数值测试。由于空间限制，对于全局和多重测试问题，我们只关注单个回归模型。

3.1 全局假设检验

在下面的模拟中，原文章分别考虑了多种维度、多种样本大小和不同模型的稀疏性的情况。对于备择假设，协变量的维数p从100，200、300到400取值，稀疏度k设置为2。样本大小n由比值$r = p/n$决定，该比值从0.2，0.4和1.2中取值。为了生成设计矩阵X，考虑用块相关的协变量进行高斯设计，使得$\Sigma = \Sigma_B$，其中Σ_B是一个$p \times p$的分块对角矩阵，包括10个大小相等的块，其对角线元素为1，非对角元素设置为0.7。在备择条件下，假设S是回归系数β的支持，且$|S| = k$，令$|\beta_j| = \rho 1\{j \in S\}$，对于$j = 1, \ldots, p$。$\rho = 0.75$且$\rho$和$-\rho$。设置$k_0 = 0$和$k_1 = 0.5$。

原文章为了评价提出的测试方法（以下称为“Proposed”方法）的表现，将测试结果与另外两种方法：U-S方法和LLR方法进行比较。因此同样地，我们也将三种方法进行了对比，表2显示了这些测试方法在1000次模拟中，置信水平$\alpha = 0.05$下的第一类错误。图2-4表示在不同条件设置下相应的方法表现。

结果表明，在r=0.2和r=0.4的情况下，proposed方法始终比另外两种方法的表现更好。并且在低维度，r=0.2的情况下，LLR方法几乎和proposed方法的表现一样好。另外，当n或p增加时，power会随之增加。不过在低维度，r=1.2的情况下，proposed方法略差于U-S方法，与原文章有轻微出入，如图5所示，这可能是由于模拟次数不够导致的。

表2 置信水平为0.05，不同n和p下三种方法的第一类错误

p/n	k=2		
	P=100	200	300
Proposed			
0.2	0.052	0.070	0.070
0.4	0.072	0.070	0.116
1.2	0.050	0.068	0.068
U-S			
0.2	0.042	0.026	0.024
0.4	0.040	0.044	0.032
1.2	0.048	0.036	0.026
LLR			
0.2	0.048	0.066	0.062
0.4	0.076	0.066	0.062

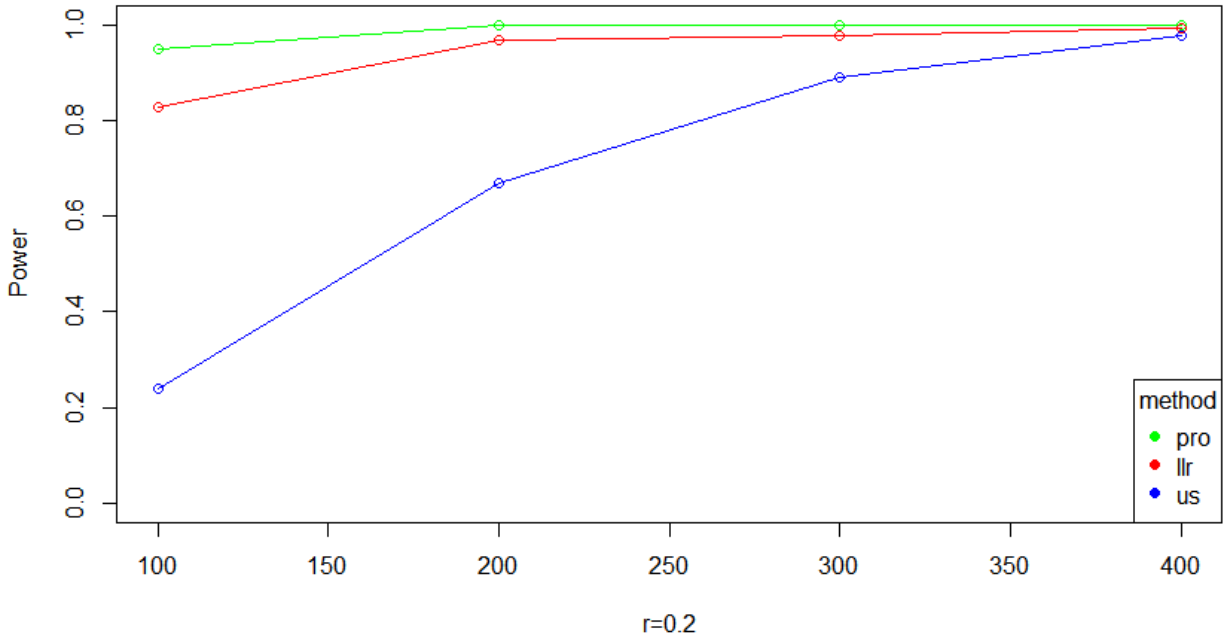

图2 r=0.2时三种方法的能力评估

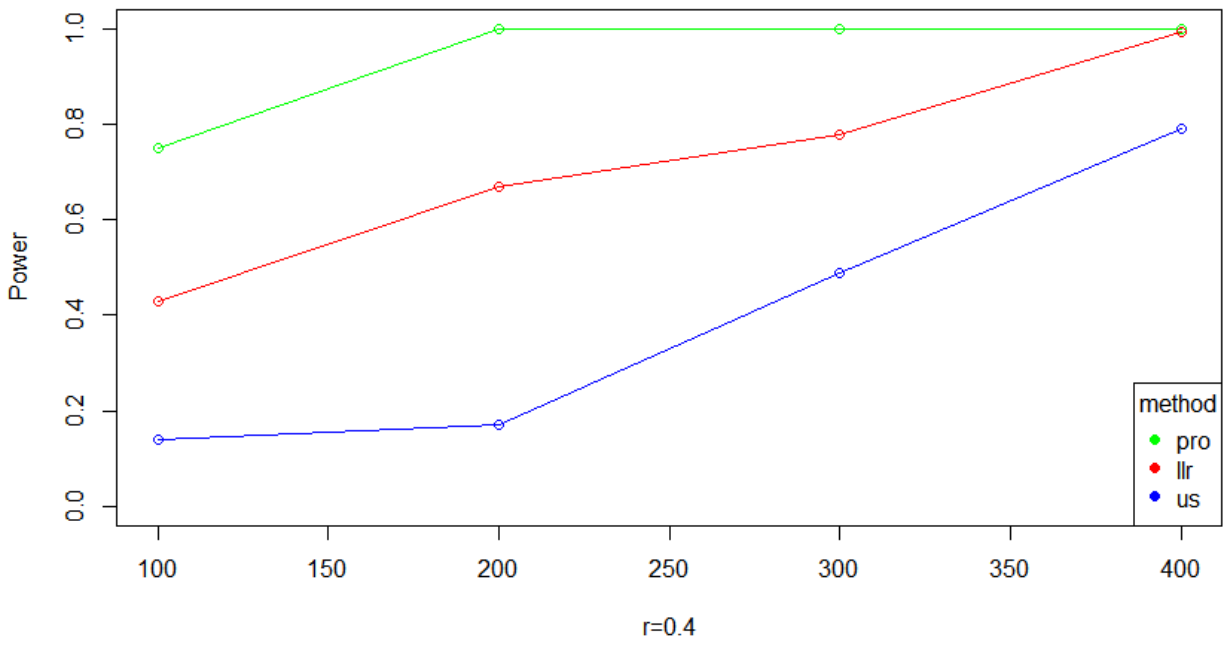

图3 r=0.4时三种方法的能力评估

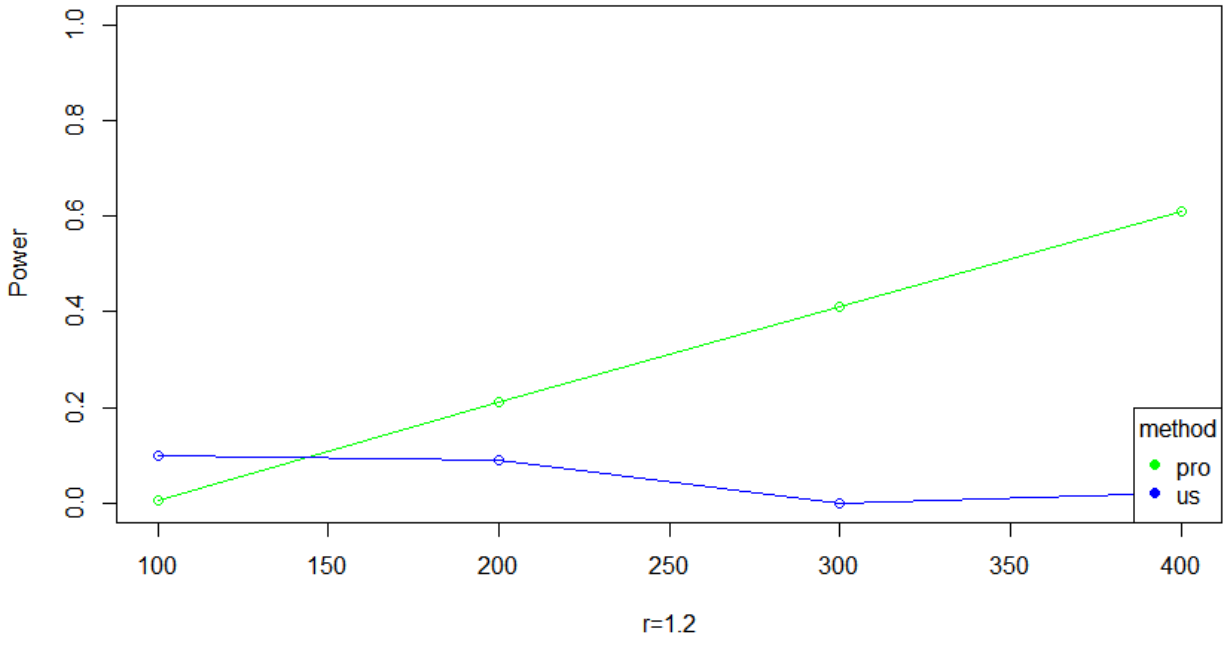

图4 r=1.2时两种方法的能力评估

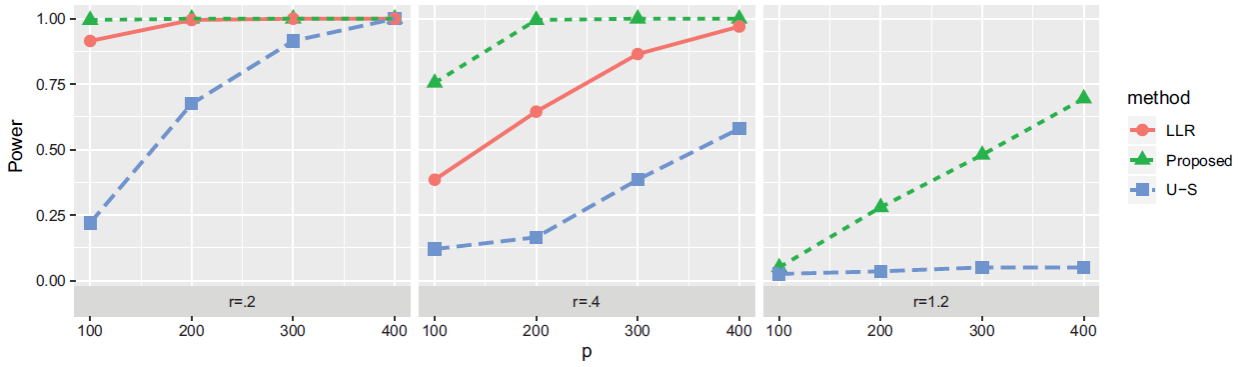

图5 原文章中对三种方法的能力评估

3.2 多重假设检验

3.2.1 FDR控制

在FDR控制中，原文章设置了$p = 800$，n从600，800，1000，1200和1400中取值，这样就总有$p > n/2$，即所有模型都是高维的。稀疏程度k从40，50和60中取值。给定支持S使得$|S| = k$，令$|\beta_j| = \rho 1j \in S$，对$j = 1, \ldots, p$，$\rho$和$-\rho$等比例。设计矩阵$X_i$产生于一个$(|X_i^T \beta| < 3)$截断多元高斯分布，其协方差矩阵$\Sigma = 0.01\Sigma_M$，其中$\Sigma_M$是一个$p \times p$的分块对角矩阵。$k_0$和$k_1$的选择与全局检验相同，另外，我们设置FDR水平为$\alpha = 0.2$。

文章将其提出的程序（以下简称为“LMT”）与以下方法进行比较：

（1）将基本LMT程序中的公式（19）的b_p替换成∞，称作“LMT0”方法；

（2）使用无偏估计量M_j的“BY”方法；

（3）一种BH方法应用于MLE统计进行单变量筛选的p值，称作“U-S”方法；

（4）“Knockoff”方法。

我们按照文章进行模拟，结果如图6-8所示，为基于1000次模拟的五种方法的“能力”（power）。这里的能力定义为正确发现的变量的数量与真正关联的变量的数量之比。

在模拟结果中，我们发现LMT和LMT0方法能正确控制FDR，并且是所有方法中最有能力的。特别地，LMT和LMT0的能力几乎相当，且能力随着稀疏程度的下降或者样本量n的增加而增加。而U-S方法，尽管正确控制了FDR，但是能力很弱，这是因为协变量之间的相关性。与原文的结果进行对比，我们发现在k=40时，LMT和LMT0方法的能力并不随着样本量的增加而增加，但是总趋势仍然是上升的，因此可以认为符合原文结论。

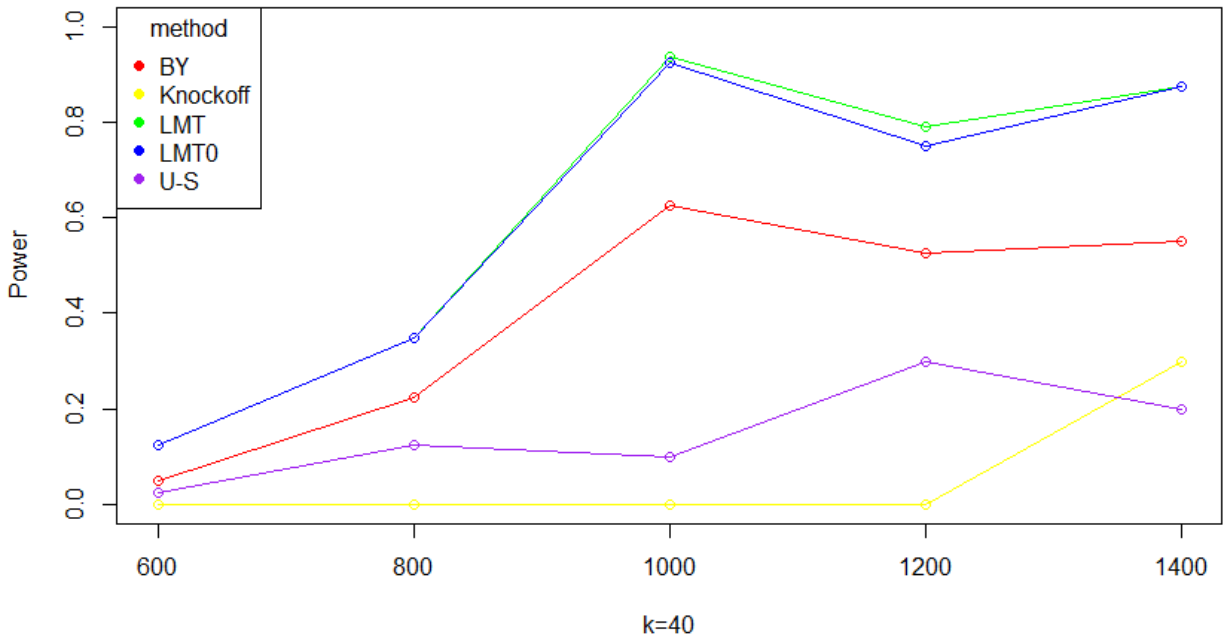

图6 k=40时不同方法控制FDR的能力评估

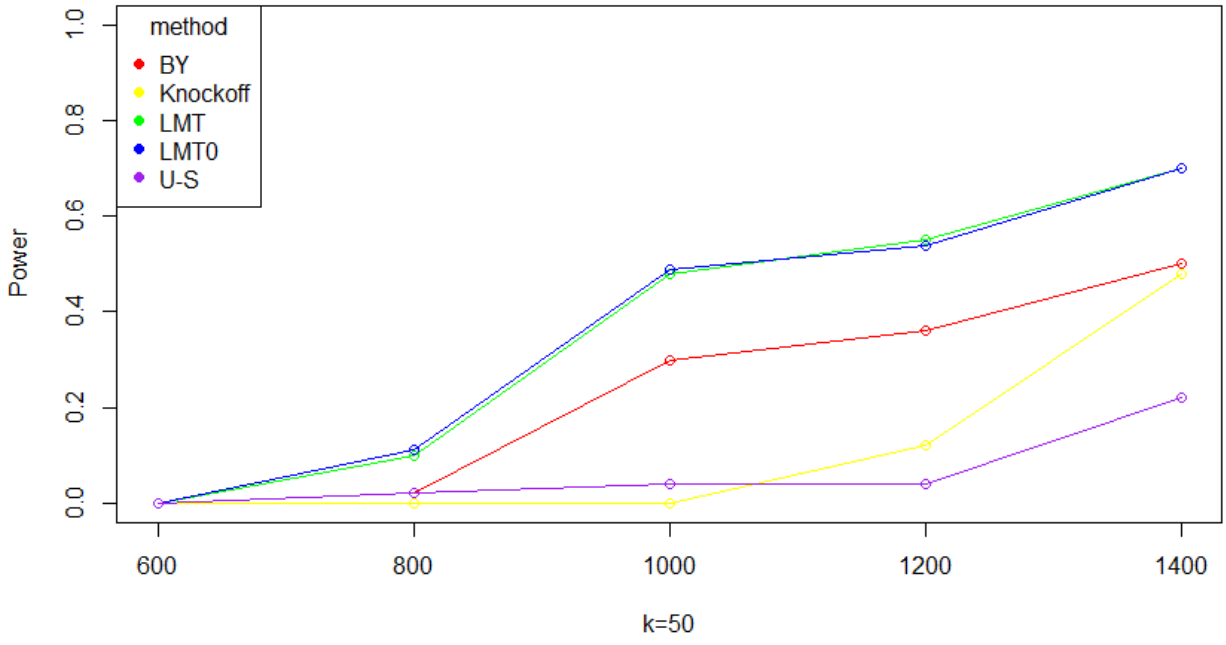

图7 k=50时不同方法控制FDR的能力评估

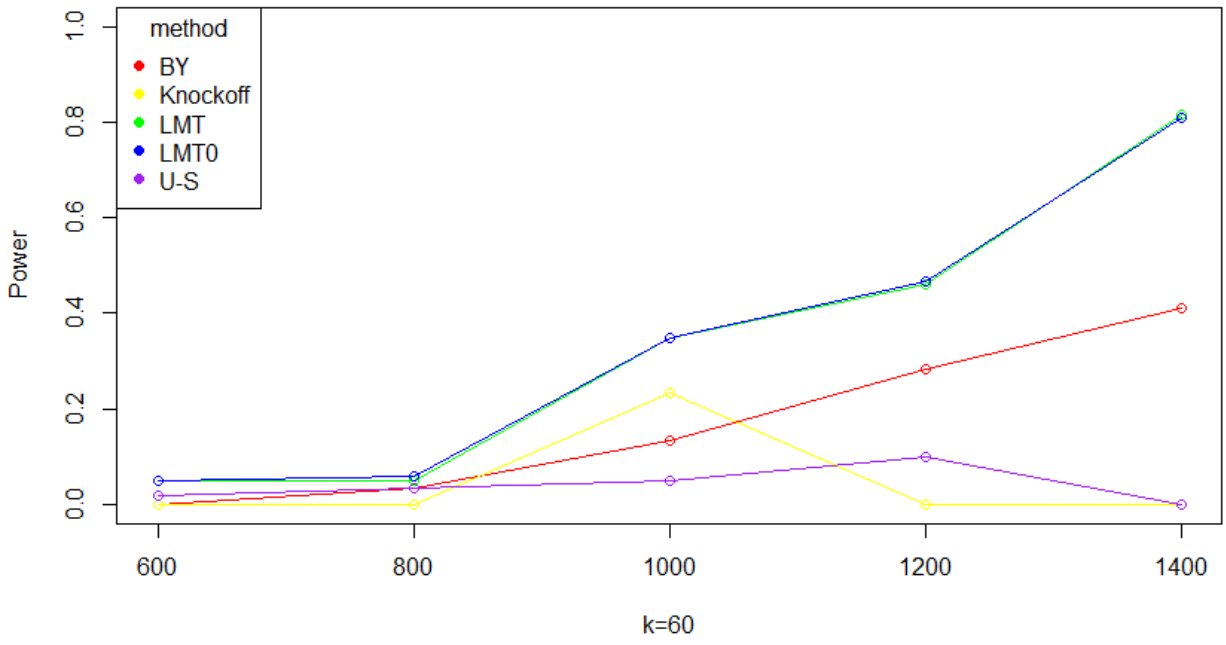

图8 k=60时不同方法控制FDR的能力评估

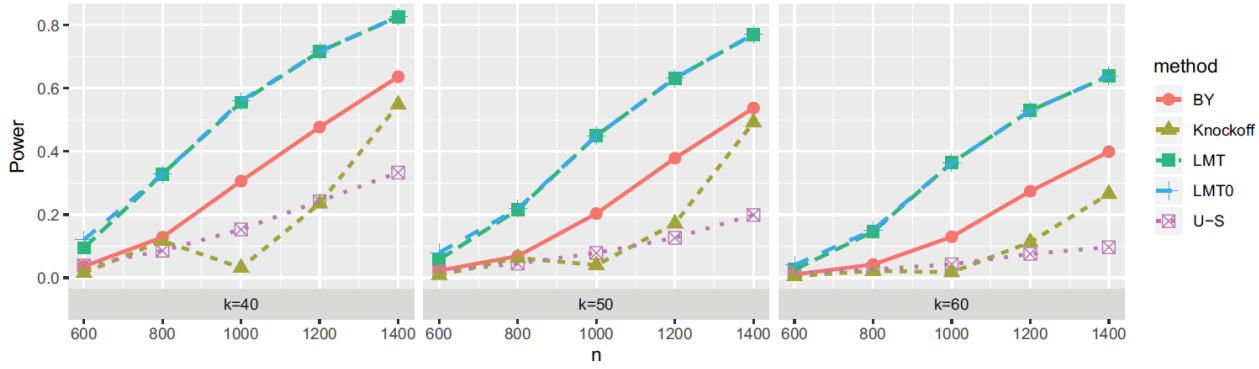

图9 原文章分别在k=40，50和60时不同方法控制FDR的能力评估

3.2.2 FDV控制

对于文章提出的测试方法控制FDV的能力（以下简称为LMT_V），设置FDV程度$r=10$，文章将测试方法应用在不同的环境设置下。特别地，设置$\rho=3$，$p\in\{800,1000,1200\}$，$k\in\{40,50,60\}$，n从400，600，800和1000中取值。设计协变量与3.2.1中类似的方法产生。我们按照文章方法进行模拟，得到的FDV检验结果和能力总结在表3中。

结果表明，文章提出的LMT_V在所有环境设置中都正确地控制了FDV，并且随着n增加，k减少或者p减少，能力会增加。

表3 控制FDV的能力及结果

ρ	p	k	Empirical FDV				Empirical power			
			n=400	600	800	1000	400	600	800	1000
3	800	40	3	6	6	10	0.125	0.175	0.375	0.525
		50	3	8	9	10	0.040	0.180	0.340	0.540
		60	8	4	10	4	0.033	0.117	0.267	0.400
	1000	40	3	8	9	7	0.050	0.200	0.200	0.550
		50	4	3	5	7	0.040	0.020	0.420	0.440
		60	2	9	7	11	0.050	0.083	0.150	0.350
	1200	40	2	4	7	6	0.075	0.150	0.250	0.400
		50	3	3	9	7	0.040	0.200	0.140	0.400
		60	5	3	5	3	0.017	0.017	0.183	0.283

4 总结

对于该论文，我们主要实现了其中三个模拟实例，分别是：（1）全局假设检验；（2）FDR控制程序；（3）FDV控制程序。由于内存空间及模拟次数的不足，与原文结果有少许出入，但是整体上仍然符合原文结论，成功地实现了这三个实例。

小组互评（满分100分）：

姓名	付芷睿	韩硕	林宇帆	孙雪倩
得分	100	100	100	100