Advances in Applied Mathematics 应用数学进展, 2021, 10(5), 1668-1680
Published Online May 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.105178

Hans汉斯

广义线性模型的平方根Lasso选择性推断

梁　博，石翔宇，张　齐*

青岛大学，山东　青岛
Email: lb17866635835@163.com, 2018020226@qdu.edu.cn, *qizhang@qdu.edu.cn

收稿日期：2021年4月25日；录用日期：2021年5月8日；发布日期：2021年5月27日

摘　要

已经有很多人对线性模型的相关问题做出了大量的选择性推断工作。但是，其适用范围并不全面。实际上，我们会遇到很多非正态非连续的数据，并且随机误差项是异方差等方面的问题，故本文将一些选择性推断的工作推广到广义线性模型上。我们建议使用数据雕琢方法，对数据样本先拆分再进行推断。另外，很多的模型变量选择的工作，都是基于已知噪声水平的。但是，在很多实际情况下，误差方差是未知的，而且在高维数据中，对它的估计存在一定难度。本文中，我们使用平方根lasso，进行变量选择，证明出对于广义线性模型，调优参数的选择不受数据中噪声波动的影响。因此，平方根lasso比lasso调优参数选择更方便，应用范围更广泛。模拟结果表明，使用数据雕琢得到的参数的置信区间更小。

关键词

广义线性模型，平方根Lasso，一致最优势无偏检验，置信区间

Selective Inference of Generalized Linear Models by the Square Root Lasso

Bo Liang, Xiangyu Shi, Qi Zhang*

Qingdao University, Qingdao Shandong
Email: lb17866635835@163.com, 2018020226@qdu.edu.cn, *qizhang@qdu.edu.cn

Received: Apr. 25th, 2021; accepted: May 8th, 2021; published: May 27th, 2021

Abstract

In view of the fact that many people have done a lot of work on the selective inference of linear models, however, its scope of application is not comprehensive. In fact, we will encounter many

*通讯作者。

文章引用: 梁博, 石翔宇, 张齐. 广义线性模型的平方根 Lasso 选择性推断[J]. 应用数学进展, 2021, 10(5): 1668-1680.
DOI: 10.12677/aam.2021.105178

problems such as non normal and non continuous data, and the random error term is heteroscedasticity. Therefore, this paper extends some selective inference work to generalized linear model. We suggest that we use the data carving to split the sample first and then infer. In addition, many of the work of model variable selection are based on the known noise level. However, in practical cases, the error variance is unknown, and in high-dimensional data, it is difficult to estimate. In this paper, we use the square root lasso to select variables, and prove that for the generalized linear model, the selection of tuning parameters is not affected by the noise fluctuations in the data. So the square root lasso is more convenient and widely used than the lasso, and the application is more extensive. Simulation shows that we get the shorter interval length by data carving.

Keywords

Generalized Linear Model, The Square Root Lasso, Uniformly Most Powerful Unbiased Test, Confidence Interval

1. 引言

完整的统计调查包括模型选择和推断两个过程。选择过程：我们使用数据来选择模型、进而确定出模型的假设检验、以及关于模型的任何其他问题。推断过程：我们根据所选择出的模型和数据，回答选择过程中提出的问题。简单地说，选择过程决定要问什么问题，推断过程决定回答那些问题。我们以前所做的工作大都是先用所有的数据进行模型选择，然后再进行模型推断。这样两个过程会重复利用数据，造成结果并不那么地精确。以至于，后来统计学家们思考有没有更简单精确的方法进行模型的选择性推断，数据(样本)分割方法[1]将数据分成两个独立随机部分，其中一部分数据用于模型选择，另一部分用于推断。Wasserman & Roeder [2]使用样本分割来解决高维模型中的一致变量选择问题。Meinshausen *et al.* [3]通过数据分割处理高维回归模型中的显著性问题，由此产生的 p 值可以控制错误发现率(FDR)和族错误率(FWER)。样本分割方法解决了控制选择性误差的问题，但这种方法减少了用于模型选择和推断过程中的数据量，成本较高。本文采用数据雕琢的方法，利用其中的一部分数据进行模型选择，另一部分数据和模型选择的剩余数据进行推断，它相对于数据分割方法提高了对于数据的利用率。

针对于线性模型，它仅在响应变量为正态随机变量时适用。Nelder [4]和 Wedderburn [5]将线性模型推广到广义线性模型。广义线性模型可以适用于以下情况：响应变量为正态和非正态变量，响应变量的分布可以是任意指数族分布。比如，二项分布和伽马分布。可以看出，广义线性模型能很好地解决回归模型中随机误差项的异方差问题。所以，把线性模型推广到广义线性模型上研究问题有重要意义。

近些年来，如生物信息、金融管理等领域产生的高维数据为模型选择带来了更大的挑战。这些领域的实验数据的维数有的甚至超过样本量大小。面对这样的高维数据，我们需要建立一个合适的数学模型，并且还要尽量使用少而有效的数据来进行建模分析，这就产生了一些新的模型选择方法。传统的模型选择方法包括最优子集选择法[6] [7]，AIC [8]和 BIC [9]等。当然，还有现在常用的模型选择方法，Robert Tibshirani 于 1996 年提出使用 lasso [10]，采用惩罚项压缩一些不重要的变量使其系数变为零。此外，与岭回归相比较，lasso 对于重要变量的系数压缩较轻，这样提高了参数估计的精度。自适应 lasso [11]和 Relaxed lasso [12]可以提高参数估计的准确性和一致性。Belloni *et al.* [13]提出平方根 lasso 调优参数 γ 的

选择与误差方差无关。这是相较于 lasso 方法，用平方根 lasso 选择模型的一个优点。利用平方根 lasso 选择变量对正态分布数据进行了大量的研究，但对非正态分布数据的研究还比较少。在本文中，我们尝试使用平方根 lasso 为广义线性模型选择变量，研究其拥有的特点。

在模型选择后，研究选定模型的推断工作，很多学者已经做了大量的工作。Lockhart *et al.* [14]基于 lasso 拟合值，推导出了原假设下的渐近检验统计量。Lee & Taylor [15]，Lee *et al.* [16]提出使用固定调优参数值的 lasso 进行精确检验。Belloni [17]，Belloni *et al.* [18]，Zhang & Zhang [19]，重点研究高维稀疏回归模型的推断。Tian *et al.* [20]提出在高维线性回归模型中，用平方根 lasso 进行方差未知的选择性推断。Lehmann & Romano [21]主要处理了关于指数族的检验问题。Lehmann & Scheffé [22]提出了一种对于指数族的一致最优势无偏(UMPU)检验的构造方法。Benjamini & Hochberg [23]提出通过控制 FDR 来解决多重显著性检验问题。Fithian *et al.* [24]提出了通过控制选择性第 I 型错误来进行模型选择后的推断。Shi *et al.* [25]将后选推断推广到广义线性模型，并提出了使用惩罚最小二乘法进行后选推断的新方法。综上所述，大部分工作是在数据服从正态分布时，对线性模型的选择性推断，而对高维异方差数据的研究较少。因此，我们基于平方根 lasso 方法对广义线性模型的选择性推断。

本篇文章的研究内容主要包括以下几个部分。在第二章中，我们提出广义线性模型的选择性推断包括两个过程：模型选择阶段和模型推断。我们应该注意控制在选定模型下的选择性第 I 类错误率。第三章，指出了我们使用平方根 lasso 和 lasso 选择变量，并且研究了选择事件的特征形式，将选择事件简写成有关于数据的表达式。在第四章中，我们给出了指数族分布的 UMPU 检验。第五章，给出了指数族分布的参数的充分统计量，以及参数所选模型下的分布形式。第六章，是对于线性回归模型和逻辑回归模型的模拟。

2. 广义线性模型

对于给定的数据 (x_i, y_i), $i=1,2,\cdots,n$，是由数据 $x_i=(x_{i1},x_{i2},\cdots,x_{ip})$ 和数据 y_i, $i=1,2,\cdots,n$ 组成，令 $X=\{x_{ij}\}$ 是设计矩阵，响应向量 $y\in R^n$。使用下面的广义线型模型来描述设计矩阵和响应向量的关系，

$$g(y)=X\beta+e \tag{2-(1)}$$

其中，β 是一个 p 维参数向量，e 是一个 n 维的随机向量，假设它满足以下条件：$E(e)=0$，$Var(e)=\sigma^2 I_n$。$g(\cdot)$是连接函数，这里把响应向量和设计矩阵二者进行了连接。

在广义线性模型中，被解释向量 y 服从的指数族分布可以被表示成如下的形式，

$$y\sim\exp\left\{\frac{yX\beta-B(X\beta)}{a(\varphi)}\right\}+c(y,\varphi) \tag{2-(2)}$$

其中，β 是感兴趣参数，φ 是讨厌参数。

在没有使用选择后推断方法的时候，统计人员在进行分析数据之前，往往根据经验指定模型一个模型 M 以及要检验的假设。在选定模型 M 和指定的原假设下，对于原假设的水平 α 检验，必须控制第 I 型错误率，

$$P_{M,H_0}(\text{reject } H_0)\le\alpha \tag{2-(3)}$$

从上面的公式可以看到，在计算第 I 类错误率时，是假定了数据来自模型 M，和原假设为真的条件。此时，若我们选定的模型 M 是不合理的，我们不能保证上面的拒绝率是小于水平 α 的。

因为在大多数的统计工作中，完全排除模型选择这一过程也是不现实的。所以统计专家会利用数据去检查他们的模型，如果诊断出问题，他们就会调整自己的模型。我们认为，如果模型的确定，是通过

先用我们手里的数据进行模型选择的，随之，确定原假设。我们就能有效地控制选择性第 I 型错误率，

$$P_{M,H_0}\left(\text{reject } H_0 \middle| M \text{ selected}\right) \le \alpha \tag{2-(4)}$$

如果用于模型选择和推断的数据是独立的，则上述条件概率(2-(4))可以写成无条件概率(2-(3))，从而得到如下公式，

$$P_{M,H_0}\left(\text{reject } H_0 \middle| M \text{ selected}\right) = P_{M,H_0}\left(\text{reject } H_0\right) \le \alpha \tag{2-(5)}$$

由此，我们通过数据分割方式将数据 y 分成两部分，即数据 $y=(y_1,y_2)$，让 y_1 独立于 y_2。那么，将样本数据 y_1 用于模型选择，样本推断过程仅仅依赖于数据 y_2。基于数据 y_2 的水平 α 的检验满足公式(2-(4))，因此，基于数据 y_2 的检验可以控制选择性第 I 类错误率(2-(3))。

数据拆分方法方便了我们的模型选择和推断的工作，也易于让人们理解，这就是当时这种方法被提出后大受欢迎的理由。但是，该方法也同时减少了模型选择和推断过程中的数据量，要付出很大代价。本文采用数据雕刻的方法，利用一部分数据 y_1 进行模型选择，另一部分数据 y_2 和选择模型中的数据一起来用于推断。此方法与数据分割方法相比，提高了数据的利用率，进而能够提高参数估计的精度。

3. 通过平方根 Lasso 选择变量

对于广义线性模型，我们考虑使用平方根 lasso 选择变量。它是对 lasso 方法的修正，通过解下面的公式，可以得到 β 的最优估计，

$$\hat{\beta} = \arg\min \left\| g(y) - X\beta \right\|_2 + \gamma \left\| \beta \right\|_1 \tag{3-(1)}$$

其中，参数 γ 是调优参数。第一项是广义线性模型的最小二乘目标函数，第二项是平方根 lasso 的惩罚项，使得许多的变量系数变成零，由此，平方根 lasso 产生了稀疏解，保留了 lasso 方法选变量的优点，并且通过定理 1 可观察到在调优参数的选择上要更加方便。

定理 1. 平方根 lasso 调优参数 γ 的选择独立于方差 σ。

以下集合定义为平方根 lasso 选择的模型，

$$\hat{M} = \left\{ j : \hat{\beta}_j \ne 0 \right\}$$

平方根 lasso 选择的变量符号定义如下，

$$\hat{s} = sign\left\{ \hat{\beta}_j : \hat{\beta}_j \ne 0 \right\}$$

注意到事件 $\left\{ \hat{M} = M \right\}$ 是通过平方根 lasso 进行变量选择产生的，接下来我们将继续注意这个选择事件。并将此选择事件进行具体的公式表达。

对于(3-(1))式的 KKT 条件的解，其结果等价于下式，

$$\begin{aligned} &\frac{X^{\mathrm{T}}\left(g(y) - X\hat{\beta}\right)}{\left\| g(y) - X\hat{\beta} \right\|_2} = \gamma \hat{s} \\ &\text{sign}\left(\hat{\beta}_{\hat{M}}\right) = \hat{s}_{\hat{M}} \\ &\left\| \hat{s}_{-\hat{M}} \right\|_\infty \le 1 \end{aligned} \tag{3-(2)}$$

子集 $\hat{M} = \left\{ i \in \{1,2,\cdots,p\} : \left|\hat{s}_j\right| = 1 \right\}$ 是使用平方根 lasso 方法得到的选择模型。$\hat{\beta}_j \ne 0, \left|\hat{s}_j\right| = 1$ 意味着选定模型的系数是非 0 的。明显地，从公式(3-(2))可以看出，有些系数是 0。但对于几乎所有的参数 γ，上述集合意味着预测变量具有非零系数。选择事件 $\left\{ \hat{M} = M \right\}$ 是通过对对每一个选择事件 $\left\{ \hat{M} = M, \hat{s} = s \right\}$ 的中

的符号函数 s 求并集得到的，并将选择事件 $\left\{\hat{M}=M,\hat{s}=s\right\}$ 表达成数据 y 的表达式。

引理 1. X 是设计矩阵，M 和 s 表示所选变量及其符号的候选集合。给出以下定义，

$$\begin{aligned} v^*(M,s) &= \left(X_M^{\mathrm{T}}X_M\right)^{-1}\left(X_M^{\mathrm{T}}g(y)-\gamma D(y)s\right) \\ u^*(M,s) &= X_{-M}^{\mathrm{T}}\left(X_M^{\mathrm{T}}\right)^{+}s+\frac{1}{\gamma D(y)}X_{-M}^{\mathrm{T}}\left(I-P_M\right)g(y) \end{aligned} \tag{3-(3)}$$

其中，$D(y)=\left\|g(y)-X\hat{\beta}\right\|_2$，$P_M=X_M\left(X_M^{\mathrm{T}}X_M\right)^{-1}X_M^{\mathrm{T}}$ 是 X_M 的投影矩阵。选择事件可以被表示成如下形式，

$$\left\{\hat{M}=M,\hat{s}=s\right\}=\left\{\operatorname{sign}\left(v^*(M,s)\right)=s,\left\|u^*\right\|_\infty<1\right\} \tag{3-(4)}$$

注意到此时，$\left\{\left\|u^*\right\|_\infty<1\right\}$ 是无效约束，表示当某些变量未被选入模型 $\hat{M}$。然而，当变量被选择进入模型时，使用条件 $\left\{\operatorname{sign}\left(v^*\right)=s\right\}$ 表示有效约束。

引理 2. v^* 和 u^* 是被定义成上面(3-(3))式。它们可以进一步被重写为如下的形式，

$$\left\{\operatorname{sign}\left(v^*(M,s)\right)=s,\left\|u^*\right\|_\infty<1\right\}=\left\{\begin{pmatrix}C_0^*(M,s)\left(I-P_M\right)\\C_1^*(M,s)\end{pmatrix}g(y)<\begin{pmatrix}b_0^*(M,s)\\b_1^*(M,s)\end{pmatrix}\right\} \tag{3-(5)}$$

其中，$C_0^*\left(I-P_M\right)$，b_0^* 对应约束条件 $\left\{\left\|u^*\right\|_\infty<1\right\}$，$C_1^*$，$b_1^*$ 对应于约束条件 $\left\{\operatorname{sign}\left(v^*(M,s)\right)=s\right\}$，它们被定义成下面的形式，

$$C_0^*(M,s)=\frac{1}{\gamma D(y)}\begin{pmatrix}X_{-M}\\-X_{-M}^{\mathrm{T}}\end{pmatrix}$$

$$b_0^*(M,s)=\begin{pmatrix}1-X_{-M}^{\mathrm{T}}\left(X_{-M}^{\mathrm{T}}\right)^{+}s\\1+X_{-M}^{\mathrm{T}}\left(X_{-M}^{\mathrm{T}}\right)^{+}s\end{pmatrix}$$

$$C_1^*(M,s)=-\operatorname{diag}(s)\left(X_M^{\mathrm{T}}X_M\right)^{-1}X_M^{\mathrm{T}}$$

$$b_1^*(M,s)=-\gamma D(y)\operatorname{diag}(s)\left(X_M^{\mathrm{T}}X_M\right)^{-1}s$$

定理 2. 基于引理 1 和引理 2，令 $C=\begin{pmatrix}C_0^*\left(I-P_M\right)\\C_1^*\end{pmatrix}$，$b=\begin{pmatrix}b_0^*(M,s)\\b_1^*(M,s)\end{pmatrix}$，我们选择事件的形式，

$$\left\{\hat{M}=M,\hat{s}=s\right\}=\left\{C^*(M,s)g(y)<b^*(M,s)\right\}$$

所以，选择事件 $\left\{\hat{M}=M\right\}$ 被表示成以下关于数据 y 的不等式的并集，

$$\left\{\hat{M}=M\right\}=\bigcup_{s\in\{-1,1\}}\left\{C^*(M,s)g(y)<b^*(M,s)\right\}$$

4. 广义线性模型的选择性推断

4.1. 选定模型下的参数推断

在上一章中，已经得到了选定的模型 $\hat{M}$，和它所对应的选择性事件 $\left\{\hat{M}=M\right\}$。在本章中，我们考虑以选定的模型 $\hat{M}$ 作为条件，响应向量 y 的分布的推断。此分布表示为，

$$L\left\{y\,|\left\{\hat{M}=M\right\}\right\} \tag{4-(1)}$$

数据 y 服从带有干扰参数的 p 维参数的指数族分布，

$$f_{\theta,\varphi}(y)=\exp\left\{\theta' T(y)+\varphi' U(y)-c(\theta,\varphi)\right\}h_0(y) \tag{4-(2)}$$

其中，θ 是感兴趣参数，φ 是讨厌参数。$T(y)$和 $U(y)$是它们相对应的充分统计量，它们分别是 k 维和 p-k 维的向量。

对于任意给定的事件 E，在 $y\in E$ 的条件下，y 的这个条件分布也是指数族分布，并且与数据 y 自身的分布拥有相同的参数和充分统计量，即，可以被表达成下面的形式，

$$f_{\theta,\varphi}(y\,|\,y\in E)=\exp\left\{\theta' T(y)+\varphi' U(y)-c_E(\theta,\varphi)\right\}h_0(y)1_E(y) \tag{4-(3)}$$

我们上面提到的模型选择事件 $\left\{\hat{M}=M\right\}$，它作为任意事件 E 的子集，故样本数据 y 在选定模型 $\hat{M}$ 下也是服从指数族，且它们的参数和充分统计量都是相同的。例如，在第三章中，我们提到的模型选择事件 $\left\{\hat{M}=M\right\}$，此选择事件被表示成数据 y 的相关表达式，是事件 $\left\{y\in E\right\}$ 的一中特例，故 y 的条件密度函数具有类似的形式，也可以写成指数族分布，

$$f_{\theta,\varphi}\left(y\,|\,\hat{M}=M\right)=\exp\left\{\theta^{*\prime}T^*(y)+\varphi^{*\prime}U^*(y)-c^*\left(\theta^*\right)\right\}h_0^*(y)1_{\hat{M}}(y) \tag{4-(4)}$$

其中 θ^*，φ^* 为选定模型的参数，$T^*(y)$ 和 $U^*(y)$ 为相应的充分统计量。接下来，我们通过消除干扰参数，来对感兴趣的参数进行推断。

如果一维参数 θ 是感兴趣参数，其它参数 φ 是干扰参数，T 和 U 是它们分别对应的充分统计量，那么，当讨厌参数的充分统计量 $\left\{U=u\right\}$ 作为已知条件时，以下条件分布仅依赖于感兴趣参数 θ，

$$(T\,|\,U=u)\sim\psi_\theta(t\,|\,u)=\exp\left\{\theta' t-c_\psi(\theta\,|\,u)\right\}\psi_0(t\,|\,u) \tag{4-(5)}$$

上述条件分布通过在 U 上的条件分布消除了讨厌参数 φ，当 k=1 时，我们得到了充分统计量 T 的单参数指数族。

根据公式(4-(4))，在选择事件 $\left\{\hat{M}=M\right\}$ 下，y 的分布仍然是指数族分布。根据公式(4-(5))感兴趣参数 θ 可由下面的条件分布进行推断，

$$L_\theta\left(T(y)\,|\,U(y),\hat{M}=M\right)$$

4.2. UMPU 选择性检验

给出了原假设 $H_0:\theta\in\Theta_0$ 和备择假设 $H_1:\theta\in\Theta_1$。如果满足以下不等式，则水平 α 检验 $\phi(y)$ 可以成为是选择性无偏的，

$$pow_\phi(\theta\,|\,E)=E_\theta\left[\phi(y)\,|\,E\right]\geq\alpha,\theta\in\Theta_1 \tag{4-(6)}$$

在所有满足上述公式的水平 α 的检验中，UMPU 水平 α 选择性检验的势函数是一致最优的。

引理 3. (UMPU 检验)y 服从指数族分布(4-(2))，k=1，在水平 α 检验下，考虑以下原假设和备择假设，

$$H_0:\theta=\theta_0\leftrightarrow H_1:\theta\neq\theta_0,$$

有 UMPU 检验 $\phi(y)=f\left(T(y),U(y)\right)$，其中，

$$f(t,u)=\begin{cases}1 & t<d_1(u)\text{或}t>d_2(u)\\ \pi_i & t=d_i(u)\\ 0 & d_1(u)<t<d_2(u)\end{cases}$$

其中， d_i 和 π_i 是下面方程组的解，

$$E_{\theta_0}\left[f(T,U)\,|\,U=u\right]=a$$
$$E_{\theta_0}\left[Tf(T,U)\,|\,U=u\right]=aE_{\theta_0}\left[f(T,U)\,|\,U=u\right]$$

由公式(4-(4))可知，当 y 服从指数族分布时，条件分布 $L\left(y\,|\left\{\hat{M}=M\right\}\right)$ 也是指数族分布，所以我们得到如下定理。

定理 3. (UMPU 选择性检验) y 遵循指数族分布(4-(2))，k=1，考虑以下在选择事件 $\left\{\hat{M}=M\right\}$ 下的原假设和备择假设，

$$H_0:\theta=\theta_0 \leftrightarrow H_1:\theta\neq\theta_0$$

有 UMPU 选择性检验， $\phi(y)=f\left(T(y),U(y)\right)$，满足，

$$f(t,u)=\begin{cases}1 & t<d_1(u)\text{或}t>d_2(u)\\ \pi_i & t=d_i(u)\\ 0 & d_1(u)<t<d_2(u)\end{cases}$$

这里的 d_i 和 π_i 满足下面的方程组，

$$E_{\theta_0}\left[f(T,U)\,|\,U=u,\hat{M}=M\right]=a$$
$$E_{\theta_0}\left[Tf(T,U)\,|\,U=u,\hat{M}=M\right]=aE_{\theta_0}\left[f(T,U)\,|\,U=u,\hat{M}=M\right]$$

在下一章中，让我们关注广义线性模型选择性推断的几个特定示例。

5. 例子

例 1. 针对于前面讨论的指数族框架下的一些结论，本章中，我们打算举几个具体例子。当数据 y 来自于多元正态分布时，

$$y\sim N_n\left(\mu,\sigma^2 I_n\right),\quad \mu=X\beta$$

在选定的模型 $\hat{M}$ 的指数族分布形式如下，

$$y\sim\exp\left\{\frac{1}{\sigma^2}\beta' X_M y-\frac{1}{2\sigma^2}\|y\|_2^2-C\left(\beta,\sigma^2\right)\right\}$$

若误差方差 σ^2 已知， $X_j'y$ 是用来推断参数 β 的充分统计量，

$$L_{\beta_j}\left(X_j'y\,|\,X_{M\setminus j}'y,\hat{M}=M\right)$$

若误差方差 σ^2 未知， $\left(X_j'y,\|y\|_2^2\right)$ 是参数 $\left(\frac{\beta_j}{\sigma^2},\sigma^2\right)$ 的充分统计量，依据下面的分布进行参数推断，

$$L_{\frac{\beta_j}{\sigma^2}}\left(X_j'y\,|\,X_{M\setminus j}'y,\|y\|_2^2,\hat{M}=M\right),$$

同样地，对于参数 σ^2 的推断，根据下面的分布，

$$L_{\sigma^2}\left(\|y\|_2^2\,|\,X_M'y,\hat{M}=M\right)$$

例 2. 假设数据 y 由伯努利分布生成，即， $y_i\sim b(1,p_i)$。

$$y_i\sim p_i^{y_i}\left(1-p_i\right)^{y_i},\ \ln\frac{p_i}{1-p_i}=\theta_i,\ i=1,2,\cdots,n$$

我们把上面这个分布写成指数族的表达形式，

$$y \sim \exp\left\{\sum_{i=1}^{n} y_i \ln\frac{p_i}{1-p_i} + c\left(p_i\right)\right\}$$

其中，这个分布有 n 个参数和相应的充分统计量。如果我们想要对感兴趣参数 θ_1 进行推断，则我们先选定需要的模型 $\left\{\hat{M}=M\right\}=\left\{i=1\right\}$，和讨厌参数 $\left\{\theta_i\right\}, i=2,3,\cdots,n$ 及所对应的充分统计量，故对它的推断依赖于以下的条件分布，

$$L_{\theta_1}\left(y_1 \mid y_2,\cdots,y_n,\left\{i=1\right\}\right),$$

例 3. 若数据 $y=\left(y_1,\cdots,y_n\right)$ 服从伽马分布 $Gamma\left(\alpha,\lambda\right)$，即，$y_i \sim \Gamma\left(\alpha,\lambda\right)$，

$$f\left(y_i\right)=\frac{\lambda^{\alpha}}{\Gamma\left(\alpha\right)} y_i^{\alpha-1}\exp^{-\lambda y_i}, \alpha,\lambda>0, i=1,\cdots,n$$

把上述分布进一步写成概率密度的表示形式，

$$y \sim \exp\left\{\left(\alpha-1\right)T\left(y\right)+\lambda N\left(y\right)-c\left(\alpha,\lambda\right)\right\}$$

其中，$T\left(y\right)=\sum_{i=1}^{n}\ln y_i$，$N\left(y\right)=\sum_{i=1}^{n} y_i$ 分别是参数 α，λ 的充分统计量。

在任一选择事件 E 下，我们想要对参数 α 进行推断，则依赖于下面的分布表达式，

$$L_{\alpha}\left(T\left(y\right) \mid N\left(y\right),E\right)$$

6. 模拟

6.1. 多元高斯分布

我们用平方根 lasso 对线性回归模型来选择变量。例如，设置 n=100，p=200，X 的行是来自变量之间具有成对相关系数为 $\rho=0.4$ 的等相关多元正态分布，y 服从以下的多元正态分布，

$$y=X\beta+e, Ee=0, Var\left(e\right)=I_n$$

将 β 设置为 200 维向量，它的前 7 个分量为 7，调整参数设置成 $E\left(\frac{\left\|X^{\mathrm{T}}e\right\|_{\infty}}{\left\|e\right\|_2}\right)$。设置置信水平 $\alpha=0.05$。

我们考虑将数据分成独立的两部分，以获得 Pscreen 和参数 β 的区间长度。其中，Pscreen 表示所有的 7 个变量被成功选择的概率。Carve(n)表示使用样本 y 的前 n 个数据进行模型选择，用剩余的数据和选择模型的数据对模型进行推断，以及 No carve 意味着使用所有的数据进行模型选择，然后再使用所有的数据进行推断。表 1 展示了我们的模拟运行结果。

Table 1. Pscreen & confidence interval lengths
表 1. Pscreen 和置信区间长度

	Pscreen	置信区间长度
Carve (50)	0.894737	0.415657
Carve (70)	0.932692	0.460035
Carve (80)	0.9375	0.717047
Carve (90)	0.953947	0.734330
No Carve	0.961539	0.568038

此外，Fithian *et al.* [25]利用 lasso 方法对线性模型进行了后选推断。当 $n=100$ 和 $p=200$ 时，X 服从多元正态分布，其变量之间两两相关 $\rho=0.3$，y 服从多元正态分布，列被规范化为长度为 1。Pscreen 和 carve (n)的含义与上面表 1 中的相同。Power 是样本观测值落到拒绝域的概率。表 2 的模拟运行结果表示如下。

Table 2. Pscreen & power
表 2. Pscreen 和 power

	Pscreen	Power	Level
Carve (50)	0.09	0.99	0.06
Carve (75)	0.68	0.97	0.05
No Carve	0.99	0.80	0.05

通过对上述结果的分析，用数据雕琢方法推断参数得到的表 1 的结果与表 2 的结果是一致的。Pscreen 表示通过雕琢选择所有 7 个变量的概率，概率会随着我们有更多数据用于变量选择而增加，并且它趋于 1。上面表 1 中的 Pscreen 符合这个理论规律，所得结果与表 2 的变化趋势一致。在表 2 中，power 随着用于选择模型的数据 n 的增加而逐渐减小，通过不进行数据雕琢的方式获得的 power，比数据雕琢获得的 power 小。结果表明，通过数据雕琢的得到的 power 较大。在表 1 中，通过数据雕琢获得的间隔的长度也随着用于模型选择的数据 n 的增加而增加。没有用数据雕琢方法得到的置信区间长度介于使用两次不同的数据雕琢后模型选择推断得到的两个区间长度值之间，说明数据雕琢方法可以使得参数区间估计更小。表 1 中的区间长度和表 2 中的 power 变化趋势表明，数据雕琢所得结果与表 2 的结果一致。

6.2. 伯努利分布

我们用平方根 lasso 模拟高斯分布，然后将此方法推广到非高斯分布。例如，我们使用数据雕琢方法来选择和推断 logistic 回归模型，在模型选择阶段，我们使用平方根 lasso 方法。当 $n=100$，$p=200$，预测变量 y 服从伯努利分布，且连接函数为 $g(y)=\ln\left(\frac{y}{1-y}\right)$ 时，我们有以下广义线性模，

$$g(y)=X\beta+e,\ Ee=0,\ Var(e)=I_n$$

将 β 设置为 200 维向量，其前 7 个分量为 7。在数据雕琢方法中，我们使用一部分数据进行选择，另一部分数据和选择模型数据一起用于推断。我们用平方根 lasso 选择模型，并在推断过程中设置置信水平 $\alpha=0.05$。我们考虑使用数据雕琢方法，将数据分别拆分成不同的部分，来获得 Pscreen，参数 β 的置信区间和覆盖率。其中，Pscreen 和置信区间的含义与第一个模拟中的含义相同，覆盖率表示参数 β 的覆盖率。表 3 显示了模拟结果。

Table 3. Pscreen & confidence interval lengths & coverage
表 3. Pscreen 和置信区间长度和覆盖率

	Pscreen	confidence interval lengths	coverage
Carve (50)	0.92045	0.42066	0.974
Carve (75)	0.93913	0.72218	0.986
Carve (90)	0.95679	0.85426	0.984
No carve	0.96067	0.67334	0.995

通过分析表 3 中得到的数据，我们得到以下结论：第一，对于广义线性模型，随着用于模型选择的数据的增加，我们得到的 Pscreen 会逐渐增大。结果表明，随着模型选择所用数据的增加，模型选择的效果会更好。第二，随着用于模型选择的数据的增加，参数的置信区间将会增大，但是，发现使用数据雕琢会比不使用数据雕琢方法获得的区间长度短。比如，直接用所有数据进行选择性推断获得的置信区间长度值，是在 carve (50)和 carve (75)之间的。这说明我们可以利用数据雕琢来减小参数的置信区间，从而使我们的选择性推断效果更好。第三，通过数据雕琢获得的参数覆盖率约为 0.97~0.98，而不通过数据雕刻获得的覆盖率略大。因此，结合这三个方面，可以利用数据雕琢进行模型选择性推断，这样使得结果更好。

参考文献

[1] Cox, D.R. (1975) A Note on Data-Splitting for the Evaluation of Significance Levels. *Narnia*, **62**, 441-444. https://doi.org/10.1093/biomet/62.2.441

[2] Wasserman, L. and Roeder, K. (2009) High-Dimensional Variable Selection. *The Annals of Statistics*, **37**, 2178-2201. https://doi.org/10.1214/08-AOS646

[3] Meinshausen, N., Meier, L. and Bühlmann, P. (2009) *p*-Values for High-Dimensional Regression. *Journal of the American Association Statistical*, **104**, 1671-1681. https://doi.org/10.1198/jasa.2009.tm08647

[4] Nelder, J.A. and Wedderburn, R. (1972) Generalized Linear Models. *Journal of the Royal Statistical Society*, **135**, 370-384. https://doi.org/10.2307/2344614

[5] Wedderburn, R. (1974) Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss—Newton Method. *Biometrika*, **61**, 439-447. https://doi.org/10.1093/biomet/61.3.439

[6] Hocking, R.R. (1976) A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression. *Biometrics*, **32**, 1-49. https://doi.org/10.2307/2529336

[7] Guyon, I. and Elisseeff, A. (2003) An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, **3**, 1157-1182.

[8] Akaike, H. (1998) Information Theory and an Extension of the Maximum Likelihood Principle. In: Parzen, E., Tanabe, K. and Kitagawa, G., Eds., *Selected Papers of Hirotugu Akaike*, Springer, New York, 199-213. https://doi.org/10.1007/978-1-4612-1694-0_15

[9] Schwarz, G.E. (1978) Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, 461-464. https://doi.org/10.1214/aos/1176344136

[10] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, *Series B*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[11] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429. https://doi.org/10.1198/016214506000000735

[12] Meinshausen, N. (2007) Relaxed Lasso. *Computational Statistics & Data Analysis*, **52**, 374-393. https://doi.org/10.1016/j.csda.2006.12.019

[13] Belloni, A., Chernozhukov, V. and Wang, L. (2011) Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming. *Biometrika*, **98**, 791-806.

[14] Lockhart, R., Taylor, J, Tibshirani, R.J. and Tibshirani, R. (2013) Rejoinder: “A Significance Test for the Lasso”. *The Annals of Statistics*, **42**, 518-531. https://doi.org/10.1214/14-AOS1175REJ

[15] Lee, J.D and Taylor, J.E. (2014) Exact Post Model Selection Inference for Marginal Screening. In: *Advances in Neural Information Processing Systems*, 136-144. https://proceedings.neurips.cc/paper/2014/file/a0a080f42e6f13b3a2df133f073095dd-Paper.pdf

[16] Lee, J.D., Sun, D.L., Sun, Y. and Taylor, J.E. (2016) Exact Post-Selection Inference, with Application to the Lasso. *Annals of Statistics*, **44**, 907-927. https://doi.org/10.1214/15-AOS1371

[17] Belloni, A., Chernozhukov, V. and Hansen, C. (2011) Inference for High-Dimensional Sparse Econometric Models. Centre for Microdata Methods and Practice, Institute for Fiscal Studies. https://arxiv.org/pdf/1201.0220v1.pdf

[18] Belloni, A., Chernozhukov, V., Fernández-Val, I. and Hansen, C. (2013) Program Evaluation and Causal Inference with High-Dimensional Data. *Econometrica*, **85**, 233-298. https://doi.org/10.3982/ECTA12723

[19] Zhang, C.H. and Zhang, S.S. (2014) Confidence Intervals for Low Dimensional Parameters in High Dimensional Li-

near Models. *Journal of the Royal Statistical Society*, *Series B*: *Statistical Methodology*, **76**, 217-242. https://doi.org/10.1111/rssb.12026

[20] Tian, X.Y., Loftus, J.R. and Taylor, J.E. (2018) Selective Inference with Unknown Variance via the Square-Root Lasso. *Biometrika*, **105**, 755-768. https://doi.org/10.1093/biomet/asy045

[21] Neath, A.A. (2006) Testing Statistical Hypotheses. *Journal of the American Statistical Association*, **101**, 847-848. https://doi.org/10.1198/jasa.2006.s100

[22] Scheffé, L.H. (1955) Completeness, Similar Regions, and Unbiased Estimation: Part II. *Sankhyā*: *The Indian Journal of Statistics*, **15**, 219-236.

[23] Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*. *Series B*: *Methodological*, **57**, 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[24] Fithian, W., Sun, D. and Taylor, J. (2014) Optimal Inference after Model Selection. arXiv: 1410.2597.

[25] Shi, X.Y., Liang, B. and Zhang, Q. (2020) Post-Selection Inference of Generalized Linear Models Based on the Lasso and the Elastic Net. *Communication in Statistics-Theory and Methods*, No. 725, 1-18. https://doi.org/10.1080/03610926.2020.1821892

附录

定理 1. 考虑广义线性模型，

$$g(y)=X\beta+\sigma^2\varepsilon, E\varepsilon=0, Var(\varepsilon)=I_n$$

$$Q(\beta)=\|g(y)-X\beta\|_2^2$$

$$\hat{\beta}=\arg\min\frac{1}{2}\|g(y)-X\beta\|_2+\gamma\|\beta\|_1=\arg\min\frac{1}{2}Q^{\frac{1}{2}}(\beta)+\gamma\|\beta\|_1$$

$$S:=\frac{\partial Q^{\frac{1}{2}}(\beta)}{\partial\beta}=\frac{\frac{\partial Q(\beta)}{\partial\beta}}{2Q^{\frac{1}{2}}(\beta)}=\frac{Exe}{E\|e\|_2}=\frac{Ex\varepsilon}{E\|\varepsilon\|_2}$$

S 是不依赖于误差方差的，故 S 不受问题中的所有噪声信息的影响。

设置参数 γ 来控制噪声，(3-(1)) 的次梯度是最佳条件，$-S_j+\gamma>0, S_j+\gamma>0$。即，$|S_j|+\gamma>0, j=1,2,\cdots,p$。

设置参数 γ，保证上述公式会成立，当 $\gamma>\max|S_j|=\|S\|_\infty$ 时，满足条件。通常，我们将参数设置为 $\gamma=k\|S\|_\infty$。

引理 1. 重写 KKT 条件，$X_{-\hat{M}}$ 表示不在模型中的变量。

$$\begin{aligned}
&X_{\hat{M}}^{\mathrm{T}}\left(X_{\hat{M}}\hat{\beta}_{\hat{M}}-g(y)\right)+\gamma D(y)\hat{s}_{\hat{M}}=0\\
&X_{-\hat{M}}^{\mathrm{T}}\left(X_{\hat{M}}\hat{\beta}_{\hat{M}}-g(y)\right)+\gamma D(y)\hat{s}_{-\hat{M}}=0\\
&\left\|\hat{s}_{-\hat{M}}\right\|_\infty<1\\
&sign\left(\hat{\beta}_{\hat{M}}\right)=\hat{s}_{\hat{M}}
\end{aligned}$$

由于 KKT 条件对于我们问题的解是充分必要的，所以当且仅当以下公式成立时，才能得到解，

$$\begin{aligned}
&X_{\hat{M}}^{\mathrm{T}}\left(X_{\hat{M}}v^*-g(y)\right)+\gamma D(y)s=0\\
&X_{-\hat{M}}^{\mathrm{T}}\left(X_{\hat{M}}v^*-g(y)\right)+\gamma D(y)u^*=0\\
&\left\|u^*\right\|_\infty<1\\
&sign\left(v^*\right)=s
\end{aligned}$$

从前两个公式，我们可以得到公式(3-(3))，最后两个公式，我们可以得到公式(3-(4))。

引理 2. 我们可以重写两个约束条件，

$$\begin{aligned}
\left\{sign\left(v^*(M,s)\right)=s\right\}&=\left\{diag(s)v^*>0\right\}\\
&=\left\{diag(s)\left(X_M^{\mathrm{T}}X_M\right)^{-1}\left(X_M^{\mathrm{T}}g(y)-\gamma D(y)s\right)>0\right\}\\
&=\left\{C_1^*(M,s)g(y)<b_1^*(M,s)\right\}
\end{aligned}$$

$$\begin{aligned}
\left\{\left\|u^*\right\|_\infty<1\right\}&=\left\{-1<X_{-M}^{\mathrm{T}}\left(X_M^{\mathrm{T}}\right)^+s+\frac{1}{\gamma D(y)}X_{-M}^{\mathrm{T}}\left(I-P_M\right)g(y)<1\right\}\\
&=\left\{C_0^*(M,s)g(y)<b_0^*(M,s)\right\}
\end{aligned}$$

定理 3.

由公式(4-(2))和(4-(3))可知，当数据 y 服从指数族分布时，条件分布 $L\left(y \mid y \in E\right)$ 也是指数族分布。选择事件 $\left\{\hat{M}=M\right\}$，作为事件 E 的特例，自然有 $L\left(y \mid \hat{M}=M\right)$ 服从指数族分布。我们得到了定理 3。