第 45 卷 第 1 期 应 用 数 学 学 报 Vol. 45 No. 1
2022 年 1 月 ACTA MATHEMATICAE APPLICATAE SINICA Jan., 2022

带有组结构的稀疏模型的参数估计和变量选择方法 *

张韵祺 张春明 †

(云南大学云南省统计建模与数据分析实验室，昆明 650091)

(威斯康星大学麦迪逊分校统计系，麦迪逊 53705, 美国)

(†E-mail: cmzhang@stat.wisc.edu)

唐年胜

(云南大学云南省统计建模与数据分析实验室，昆明 650091)

摘 要 本文针对带有组结构的广义线性稀疏模型, 引入布雷格曼散度作为一般性的损失函数, 进行参数估计和变量选择，使得该方法不局限于特定模型或特定的损失函数. 本文比较研究了 Ridge, SACD, Lasso, 自适应 Lasso, 组 Lasso, 分层 Lasso, 自适应分层 Lasso 和稀疏组 Lasso 共 8 种惩罚函数的特点和引入模型后参数估计和变量选择的方法，并给出了分层 Lasso 的坐标轴下降算法和稀疏组 Lasso 的加速全梯度更新算法. 模拟研究验证了组 Lasso, 分层 Lasso, 自适应分层 Lasso 和稀疏组 Lasso 能更好的利用数据的组结构信息, 自适应分层 Lasso 和稀疏组 Lasso 在变量选择准确性，参数估计精度方面优于其它方法，稀疏组 Lasso 在模型预测精度上达到最优. 作为实证研究，本文将带有稀疏组 Lasso 惩罚的逻辑斯蒂模型应用于骨关节炎患者的外周血单核细胞基因表达水平的分析，选出了 9 个基因集中共 136 个基因与骨关节炎有关，以期对后续生物医学研究有一定指导价值

关键词 Lasso; 布雷格曼散度；组结构；广义线性模型；稀疏模型

MR(2000) 主题分类 62J12; 62P10

中图分类 O212.1

本文 2019 年 12 月 24 日收到. 2020 年 7 月 15 日收到修改稿.

* 国家自然科学基金 (11690014), 美国自然科学基金 (DMS-1712418), 以及威斯康星校友研究基金资助

† 通讯作者.

1 引言

随着社交网络，电子商务，金融风控，云计算和基因工程等大数据应用的飞速发展，产生了大量几百维甚至上万维的高维或超高维数据. 同时，由于经济和时间成本等原因的限制，此类数据通常还伴随着小样本问题. 高维小样本数据的变量空间维数很高而样本空间维数很低，这会为建模带来许多问题. 样本量太小会导致过拟合，使得模型难以外推预测；而变量维数太高会使得估计过程产生病态方程组，模型没有唯一解且变量值的微小变化就会导致模型拟合值发生很大变化. 稀疏模型能够有效的解决此类问题，它具有变量选择功能，可以将大量的冗余变量去除，只保留与响应变量最相关的解释变量，简化了模型的同时却保留了数据集中最重要的信息，并且具有良好的可解释性. 惩罚方法常被用于稀疏建模，它通过最小化以下目标函数来估计参数$\boldsymbol{\beta}\in\mathbb{R}^p$

$$T(\boldsymbol{\beta})=L(\boldsymbol{\beta})+P_\lambda(\boldsymbol{\beta}),$$

其中， $L(\boldsymbol{\beta})$ 是衡量模型拟合程度的损失函数， $P_\lambda(\boldsymbol{\beta})$ 是评估参数 $\boldsymbol{\beta}$ 自然合理性的惩罚函数， λ 是正则化惩罚的调节参数.

近年来，众多学者提出了不同惩罚函数 $P_\lambda(\boldsymbol{\beta})$ 来构建稀疏模型. 经典的 Lasso (Least Absolute Shrinkage and Selection Operator, Lasso) 方法 [1] 使用 L_1 范数正则化惩罚得到$\boldsymbol{\beta}$的估计. 该方法在零点处求解次梯度得到稀疏解，使得模型向量的许多分量为零，实现模型稀疏化和变量选择. 其它方法如岭回归 (Ridge Regression)[2], 自适应 Lasso(Adaptive Lasso)[3] 和光滑截断绝对差 (Smoothly Clipped Absolute Deviation, SCAD)[4] 等都是通过改变惩罚函数得到不同的参数估计，从而达到降维和变量选择的目的.

以上方法都是单独考虑每个参数的估计，然而解释变量往往具有组结构，例如在基因表达分析中可把属于同一生物学路径或具有类似生物学功能的基因看做一个组；在股票分析中可把相同行业分类的上市公司看做一个组. 若将这种组结构信息作为先验信息对变量进行特征分组后再构造惩罚函数拟合模型，就可以将某些变量作为一个整体被同时选中进而参与模型的构建，或同时从模型中移除进而不参与模型的构造，即具有变量组选择的效果. Yuan 和 Lin[5] 提出了组 Lasso(Group Lasso) 方法，该方法将 L_2 范数作为正则化选项，实现了变量组水平上的稀疏性，从而具有变量组的选择能力. 为了同时实现组内和组间的变量选择， Zhou 和 Zhu[6] 提出了分层 Lasso(Hierarchical Lasso) 和自适应分层 Lasso(Adaptive Hierarchical Lasso) 的方法， Simon 和 Friedman 等人 [7] 提出了稀疏组 Lasso(Sparse Group Lasso) 的方法.

构建稀疏模型的另一个重要因素是损失函数 $L(\boldsymbol{\beta})$. 基于平方损失的传统方法适用于一般线性模型，而许多可用于广义线性模型的基于似然方法的损失函数要求知道响应变量的概率分布. 因而本文致力于构造一种适用于未知数据分布，可结合更宽泛的损失函数类型的广义线性模型估计方法. Zhang 和 Jiang 等人 [8,9] 将一类称之为布雷格曼散度 (Bregman Divergence, BD) 的损失函数和 Lasso, SCAD 等惩罚函数进行结合，提出了一种高维广义线性模型的变量选择方法. 受此启发，本文将 BD 损失引入到了带组

结构的稀疏模型中，结合组 Lasso, 分层 Lasso, 自适应分层 Lasso 和稀疏组 Lasso 惩罚进行模型参数估计和变量选择.

本文提出的带组结构的稀疏模型的参数估计和变量选择方法不局限于特定模型或特定的损失函数，该方法不仅可以用于平方损失和适用于指数簇的对数似然损失等常见损失函数，还可以用于观测值的分布未知或部分未知的情形. 同时，该方法可以在选择单个重要变量的同时识别出多个变量间的组结构信息，在实际问题的变量选择和结果预测中提供更多的数据信息和更好的拟合结果. 本文还以逻辑斯蒂回归 (Logistic Regression) 为例，详细阐述了当解释变量带组结构时模型的估计方法，并给出了相应的迭代算法. 经过模拟计算比较文中提到的 8 种惩罚函数，自适应分层 Lasso 和稀疏组 Lasso 惩罚方法能同时进行组内和组间的变量选择且参数估计和变量选择结果更为精确. 同时，稀疏组 Lasso 惩罚方法的预测结果更为准确，其误分率 (Misclassifacation Rate) 在 8 种方法中最低. 本文将带有稀疏组 Lasso 惩罚的逻辑斯蒂模型应用于骨关节炎 (Osteoarthritis, OA) 患者的外周血单核细胞 (PBMC) 基因表达水平的分析，选出了包含有 136 个基因的 9 个基因集与骨关节炎有关. 该发现可用于相关基因的后续生物医学研究，进一步阐明骨关节炎的发病机制并研发靶向药物.

2　方法

2.1　模型

令 $\{(\boldsymbol{X}_1, Y_1), \cdots, (\boldsymbol{X}_n, Y_n)\}$ 为基于总体 $(\boldsymbol{X}, Y)$ 的一组独立观测，其中 $\boldsymbol{X} = (X_1, \cdots, X_p)^\top \in \mathbb{R}^p$ 为解释变量组成的向量, Y 是被解释变量.考虑广义线性模型 $m(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x}) = F^{-1}(\beta_0 + \boldsymbol{x}^\top\boldsymbol{\beta})$, 其中 β_0 为截距项， $F(\cdot)$ 为连接函数. 当解释变量有组结构时，根据组结构信息将其分为 G 个组，每个组包含有 p_g 个变量 $(p = p_1 + \cdots + p_G)$, 于是模型变为：

$$m(\boldsymbol{x}) = F^{-1}\Big(\beta_0 + \sum_{g=1}^{G} \boldsymbol{x}^{(g)\top}\boldsymbol{\beta}^{(g)}\Big), \tag{1}$$

其中 $\boldsymbol{x} = (\boldsymbol{x}^{(1)\top}, \cdots, \boldsymbol{x}^{(G)\top})^\top$, $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)\top}, \cdots, \boldsymbol{\beta}^{(G)\top})^\top$.

在本文考虑的基因表达水平模型中，被解释变量 Y 是一个二元变量，即 $Y_i \in \{0, 1\}$. 此时 (1) 式中的 $m(\boldsymbol{x})$ 就是分类条件概率 $Pr(\boldsymbol{x}) = Pr(Y = 1|\boldsymbol{x})$. 考虑对数连接函数 $F(\theta) = \log\{\theta/(1-\theta)\}$, 可构建如下逻辑斯蒂回归模型：

$$F\{m(\boldsymbol{x})\} = \beta_0 + \sum_{g=1}^{G} \boldsymbol{x}^{(g)\top}\boldsymbol{\beta}^{(g)}. \tag{2}$$

2.2　损失函数

布雷格曼散度 (Bregman Divergence, BD) 是一种广义距离测度簇，其形式如下：

$$Q(\nu, \mu) = -q(\nu) + q(\mu) + (\nu - \mu)q'(\mu),$$

其中 $q(\cdot)$ 被称 $Q(\cdot,\cdot)$ 的生成函数. 假设 $q(\cdot)$ 是凹函数且有一阶导数 $q'(\cdot)$, 则 $q(\cdot)$ 的凹性确保了 $Q(\cdot,\cdot)$ 是一个非负函数. 当 $q(\cdot)$ 严凹时, 当且仅当 $\nu=\mu$ 时 $Q(\nu,\mu)=0$. 因此, 当我们分别用 ν 和 μ 代表观测和估计时, $Q(\nu,\mu)$ 可具备损失函数的优良性质.

使用不同的生成函数 $q(\cdot)$, 布雷格曼散度可变换为多种误差损失 [10], 包括许多常见的损失函数. 例如, 当 $q(\mu)=a\mu-\mu^2$ 且 a 为常数时, $Q(y,\mu)=(y-\mu)^2$ 是二次损失函数; 当 $q(\mu)=-2\{\mu\log(\mu)+(1-\mu)\log(1-\mu)\}$ 时, $Q(y,\mu)=-\{y\log(\mu)+(1-y)\log(1-\mu)\}$ 为伯努利距离损失函数; 当 $q(\mu)=2\{\mu(1-\mu)\}^{1/2}$ 时, $Q(y,\mu)=\exp\{-(y-1/2)\log(\mu/(1-\mu))\}$ 为指数损失函数 [11]; 当 $q(\mu)=\mu-\mu\log(\mu)$ 时, $Q(y,\mu)=y\{\log(y)-\log(\mu)\}-(y-\mu)$ 为拟似然损失函数. 特别地, 布雷格曼散度还包含了许多稳健的损失函数, 例如常见的 Check function, 根据 Zhang 和 Jiang 等 [8] 的定理三可知, 若损失函数 $Q(Y,\mu)$ 在 $j=1,\cdots,K$ 时, $Q(Y,\mu_j+)=\lim\limits_{\mu\downarrow\mu_j}Q(Y,\mu)$ 和 $Q(Y,\mu_j+)=\lim\limits_{\mu\uparrow\mu_j}Q(Y,\mu)$ 存在且有限, 且满足当 $Y\neq\mu_j$ 时, $\frac{Q(Y,\mu_j+)-Q(Y,\mu_j-)}{Y-\mu_j}\leq 0$ 并且与 Y 无关, 于是存在一个凹函数 q, 可以作为生成函数, 使得损失函数 $Q(Y,\mu)$ 为一种布雷格曼散度. 对于 Check function, $Q(Y,\mu)=\rho_\tau(Y-\mu)=(Y-\mu)[\tau-I(Y<\mu)]$, 容易计算 $Q(Y,\mu_j+)=I(Y<\mu_j)-\tau$, $Q(Y,\mu_j-)=\tau-I(Y<\mu_j)$, 此时的 $Q(Y,\mu)$ 满足上述条件, 于是可将 Check function 看作是一种布雷格曼散度. 另外, Zhang 和 Guo 等人 [12] 提出了一类稳健的 BD 函数, 包含有 Huber 损失函数, 稳健拟似然损失函数等; Hennequin 和 David 等人 [13] 证明了 β 散度是一种布雷格曼散度, 而 Mihoko 和 Eguchi[14] 基于 β 散度构建了一种稳健的估计方法.

在本文考虑的基因表达水平分析中, 被解释变量 Y 是一个二元变量, 因此, 伯努利距离损失 $Q(y,\mu)=-\{y\log(\mu)+(1-y)\log(1-\mu)\}$ 和指数损失 $Q(y,\mu)=\exp\{-(y-1/2)\log(\mu/(1-\mu))\}$ 都可被看做 BD 损失. Zhang 和 Jiang 等人 [8] 指出, 两种损失函数在回归模型估计中对分类结果的影响可忽略不计, 实际应用中选择其中任意一种损失函数即可.

2.3 惩罚函数

惩罚函数的使用可以减少庞大, 冗杂的数据信息对稀疏模型估计的干扰, 并达到变量选择的效果. 近年来, 统计学家们提出了许多经典的惩罚方法, 包括岭回归, Lasso, 自适应 Lasso, SCAD 等方法. 本文将对此类方法进行简单介绍, 以便与带组结构的 Lasso 方法进行比较. 当被解释变量包含有组结构信息时, 经典方法不能充分利用数据信息, 估计出变量的组结构形式. 此时, 就需要引入一些带组结构的惩罚方法, 例如组 Lasso, 分层 Lasso, 自适应分层 Lasso 和稀疏组 Lasso.

2.3.1 岭回归 (Ridge)

Hoerl 和 Kennard[2] 提出了一种以损失无偏性来换取数值高的参数的稳定性, 从而减弱多重共线性的方法. 岭回归是一种正则化模型, 对参数 $\boldsymbol{\beta}$ 施加了 L_2 惩罚. 但该方法对回归系数的惩罚并不能使其压缩至零, 因而不能产生稀疏解. 岭回归的惩罚函数如

下：

$$P_{\lambda(\boldsymbol{\beta})} = \lambda||\boldsymbol{\beta}||_2^2,$$

其中，$||\cdot||_2$ 表示一个向量的 L_2 范数.

2.3.2 Lasso

Tibshirani[1] 提出了一种通过对回归系数的绝对值之和 (即回归系数向量的 L_1 范数) 进行惩罚的方法来压缩回归系数的大小，使绝对值较小的回归系数自动被压缩为 0，从而产生稀疏解和实现变量选择. Lasso 的惩罚函数形式为：$P_\lambda(\boldsymbol{\beta}) = \lambda||\boldsymbol{\beta}||_1$, 其中，$||\cdot||_1$ 表示一个向量的 L_1 范数.

2.3.3 自适应 Lasso(AdLasso)

由于 Lasso 方法的估计是有偏的，为了克服此缺点，Zou[3] 提出了一种让各个变量对应的回归系数所受到的惩罚程度具有自适应的性质的惩罚函数. 自适应 Lasso 使用不同的自适应权重，使惩罚函数中对应目标变量的回归系数的惩罚较大，而对应噪声变量的回归系数的惩罚较小. 自适应 Lasso 具有 Oracle 性质，即其模型选择具有稀疏性和相合性，且参数 $\boldsymbol{\beta}$ 的估计满足渐进正态性. 自适应 Lasso 的惩罚函数形式如下：$P_{\lambda,\omega}(\boldsymbol{\beta}) = \lambda\boldsymbol{\omega}^T||\boldsymbol{\beta}||_1$, 其中，$\boldsymbol{\omega} = (\omega_1, \cdots, \omega_p)$ 为权重向量. 在实际应用中，可以选用 $\omega_j = \frac{1}{|\widehat{\beta}_j^{\mathrm{OLS}}|^\gamma}$, 其中 $\widehat{\beta}_j^{OLS}$ 为第 j 个回归系数的普通最小二乘估计值，$\gamma > 0$; 当模型维数很高时，也可选用岭回归的参数估计值构造权函数，即 $\omega_j = \frac{1}{|\widehat{\beta}_j^{\mathrm{Ridge}}|^\gamma}$，$\widehat{\beta}_j^{\mathrm{Ridge}}$ 为第 j 个回归系数的岭回归估计，$\gamma > 0$.

2.3.4 光滑截断绝对差，SCAD

Fan 和 Li[4] 提出的光滑截断绝对差 (Smoothly Clipped Absolute Deviation, SCAD) 方法和自适应 Lasso 的思想类似，通过对各个变量对应的回归系数施加不同程度的惩罚而达到参数估计和变量选择的一致性. SCAD 惩罚将绝对值小于 λ 的回归系数 (即噪声变量的回归系数) 压缩至零，而对绝对值在 $[\lambda, \alpha\lambda]$ 区间的回归系数 (即目标变量的回归系数) 随着回归系数绝对值的增大而减小压缩的程度，对于绝对值大于 $\alpha\lambda$ 的回归系数 (即目标变量的回归系数) 不进行压缩. SCAD 的惩罚函数形式为：

$$\begin{aligned} P_{\lambda,\alpha}(\boldsymbol{\beta}) =& \lambda|\beta_j|I(|\beta_j| \le \lambda) \\ &+ \frac{(2\alpha\lambda|\beta_j| - |\beta_j|^2 - \lambda^2)}{2(\alpha-1)} I(\lambda < |\beta_j| \le \alpha\lambda) + \frac{(\alpha+1)\lambda^2}{2} I(|\beta_j| > \alpha\lambda). \end{aligned}$$

2.3.5 组 Lasso(GLasso)

当解释变量具有组结构时，Yuan 和 Lin[5] 提出的组 Lasso(Group Lasso) 方法可以将具有某些共同特征的一组变量作为一个整体被同时选中参与模型构建或同时从模型中移除. 组 Lasso 的模型稀疏性是特征变量组水平上的，那些未被选中的组中的特征变

量系数均为 0. 组 Lasso 的惩罚函数为：

$$P_{\lambda(\boldsymbol{\beta})} = \lambda \sum_{g=1}^{G} \sqrt{p_g} ||\boldsymbol{\beta}^{(g)}||_2.$$

2.3.6 分层 Lasso(HLasso)

组 Lasso 是一种 “all-in-all-out” 的变量选择方法，即若组中的一个重要变量被选中后，整个组的其他变量都会被选中. 为了实现在移除不重要的组的同时，也能保持在组内仅选取重要变量的灵活性， Zhou 和 Zhu[6] 提出了分层组 Lasso (Hierarchical Lasso). 分层组 Lasso 将每个参数分解为两个层次结构：组分量和个体分量. 第 g 个组的第 j 个参数 β_{gj} 可被写为： $\beta_{gj} = d_g \alpha_{gj}$, 其中 $d_g > 0$ 是组分量， α_{gj} 是个体分量. 分层组 Lasso 的惩罚函数可写为：

$$P_\lambda(\boldsymbol{\beta}) = \sum_{g=1}^{G} \Big(d_g + \lambda \sum_{j=1}^{p_g} |\alpha_{gj}| \Big) = \lambda \sum_{g=1}^{G} \sqrt{||\boldsymbol{\beta}^{(g)}||_1}.$$

2.3.7 自适应分层 Lasso(AdHLasso)

由于普通分层 Lasso 对于参数 $\boldsymbol{\beta}$ 的每一个分量的惩罚程度是一致的，这样的惩罚程度对某些较大的 β_j 而言会显得过大，从而使得估计值 $\widehat{\boldsymbol{\beta}}$ 有偏. Zhou 和 Zhu[6] 还提出了自适应分层 Lasso(Hierarchica Lasso) 的方法，自适应分层 Lasso 通过对不同组的变量施加不同的权重从而施加不同的惩罚，使得对参数 $\boldsymbol{\beta}$ 的惩罚程度具有 “自适应” 的性质，即惩罚程度随着变量对模型影响程度的增大而减小. 自适应分层 Lasso 的惩罚函数为：

$$P_{\lambda,\omega}(\boldsymbol{\beta}) = \lambda \sum_{g=1}^{G} \sqrt{\boldsymbol{\omega}_g{}^T ||\boldsymbol{\beta}^{(g)}||_1},$$

其中， $\boldsymbol{\omega} = (\omega_1 1, \cdots, \omega_g p_g)$ 为权重向量. 同样地，可以选用普通最小二乘估计值 $\widehat{\beta}_{gj}^{\mathrm{OLS}}$ 或岭回归估计 $\widehat{\beta}_{gj}^{\mathrm{Ridge}}$ 构造权函数 $\omega_j = \frac{1}{|\widehat{\beta}_{gj}^{\mathrm{OLS}}|^\gamma}$ 或 $\omega_j = \frac{1}{|\widehat{\beta}_{gj}^{\mathrm{Ridge}}|^\gamma}$， $\gamma > 0$.

2.3.8 稀疏组 Lasso(SGLasso)

为了同时实现组间特征选择和组内的特征选择， Simon 和 Friedman 等人 [7] 在普通组 Lasso 中引入 L_1 范数，提出了稀疏组 Lasso(Sparse Group Lasso) 的方法. 该方法可以同时在组水平和个体水平对参数 $\boldsymbol{\beta}$ 进行不同惩罚，同时实现组水平和个体水平的稀疏性，从而进行变量选择. 稀疏组 Lasso 的惩罚函数形式为：

$$P_{\lambda,\alpha}(\boldsymbol{\beta}) = (1-\alpha)\lambda \sum_{g=1}^{G} \sqrt{p_g} ||\boldsymbol{\beta}^{(g)}||_2 + \alpha\lambda ||\boldsymbol{\beta}||_1.$$

3　算法

坐标轴下降算法 (Coordinate Descent, CD) 是一种对一个可微的多变量凸函数 $F(\boldsymbol{X})$ 每次沿一个坐标方向优化来获取最小值的迭代算法. 在每次迭代时，固定一个坐标方向进行优化并更新变量值，然后选择下一个坐标方向进行优化，直到向量 $\boldsymbol{X}$ 的各个方向上的优化都达到收敛时，损失函数值最小，此时的参数估计值即为我们要求的结果. 很多学者都已经分别从理论和模拟试验的角度说明了坐标轴下降算法的优良性质，并应用于变量选择问题中. Friedman 和 Hastie 等人 [15], Wu 和 Lange[16] 将该方法引入到了带有惩罚的线性回归问题， Zhang 和 Zhang 等人 [10] 又将该算法引入到了带有惩罚的广义线性模型中. 本文介绍的各种惩罚方法都可以结合不同的 BD 损失函数使用坐标轴下降算法求解. 特别地，当解释变量维数很高时，使用 CD 算法求解含稀疏组 Lasso 惩罚的广义线性模型，算法的收敛速度会很慢. 于是，我们引入了 Zhang 和 Chai 等人 [17] 提出的加速全梯度更新算法 (Accelerated Full Gradient Update, AFGU) 来求解含稀疏组 Lasso 惩罚的模型， Zhang 和 Chai 等人 [17] 讨论了 AFGU 算法的 Q- 二次收敛性，说明了 AFGU 算法的加速效果. AFGU 算法每次只需计算参数似然的全梯度以极小化目标函数，而 CD 算法每次需要计算每个坐标方向上的梯度来求得极小值；同时，在每次迭代中，当更新了全梯度后， AFGU 使用牛顿迭代法对参数进行更新，此时，参数空间下降至非零参数的维数，这在稀疏模型中就能大大加快运算速度. 值得注意的是，由于加速全梯度更新算法的第一步是对似然函数或拟似然函数进行全梯度更新，因此，该方法需要知道随机变量的分布，写出其似然函数；或者知道随机变量的期望和方差之间的联系函数，构建拟似然函数. 下面，我们将以分层 Lasso 为例给出坐标轴下降算法的具体过程，并给出稀疏组 Lasso 的加速全梯度更新算法过程.

3.1　分层 Lasso 的坐标轴下降算法

受 Zhang 和 Jiang 等人 [10] 的启发，本文将惩罚函数为 Lasso 的广义线性模型的坐标轴下降算法推广至惩罚函数为分层 Lasso 的情形. 在广义线性模型中，带分层 Lasso 惩罚的参数 $\boldsymbol{\beta}$ 的 BD 估计可通过极小化如下目标函数求得：

$$T(\boldsymbol{d},\boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^{n} Q(Y_i, m(\boldsymbol{X}_i)) + \sum_{g=1}^{G}(d_g + \lambda\sum_{j=1}^{p_g}|\alpha_{gj}|). \tag{3}$$

定义 $q^j(y,\theta) = (\partial^j/\partial\theta^j)Q(y, F^{-1}(\theta))$， $j = 0, 1, \cdots$， $\widetilde{\beta}_0$, 对 BD 损失函数 $Q(Y_i, m(\boldsymbol{X}_i)) = Q(Y_i, F^{-1}(\beta_0 + \boldsymbol{x}_i^\top\boldsymbol{\beta}))$ 进行泰勒展开后可得其近似计算值：

$$\frac{1}{2}\sum_{i=1}^{n} t_i(Z_i - \beta_0 - x_i^\top\boldsymbol{\beta})^2,$$

其中， $t_i = q^2(y_i, \widetilde{\beta}_0 + x_i^\top\widetilde{\boldsymbol{\beta}})/n$， $Z_i = (\widetilde{\beta}_0 + x_i^\top\boldsymbol{\beta}) - q^1(y_i, \widetilde{\beta}_0 + x_i^\top\widetilde{\boldsymbol{\beta}})/q^2(y_i, \widetilde{\beta_0} + x_i^\top\widetilde{\boldsymbol{\beta}})$， $\widetilde{\beta}_0$ 和 $\widetilde{\boldsymbol{\beta}}$ 分别为 β_0 和 $\boldsymbol{\beta}$ 的初值. 因此，分层 Lasso 的坐标轴下降算法可通过如下过程实现：

1. 初始化 d_g, α_{gj} 和 $\beta_{gj}=d_g\alpha_{gj}$.

2. 根据上述公式计算 Z_i 和 t_i. 标准化 Z_i 和 x_i, 即 $\widetilde{Z}_i = Z_i - \big(\sum_{i=1}^n t_iZ_i\big)/\big(\sum_{i=1}^n t_i\big)$, $\widetilde{x}_i = x_i - \big(\sum_{i=1}^n t_ix_i\big)/\big(\sum_{i=1}^n t_i\big)$, 以消除模型中的截距项. 此时, 迭代问题变为:

$$\min_{d_g,\alpha_{gj}} \Big\{\frac{1}{2}\sum_{i=1}^n t_i(\widetilde{Z}_i-\widetilde{x}_i^\top\boldsymbol{\beta})^2+\sum_{g=1}^G(d_g+\lambda\sum_{g=1}^G\sum_{j=1}^{p_g}\omega_{gj}|\alpha_{gj}|)\Big\},$$

其中 ω_{gj} 为自适应分层 Lasso 的权重, 普通分层 Lasso 的 ω_{gj} 为 1.

3. 使 $x^d_{i,gj}=d_gx_{i,gj}$, $g=1,\cdots,G$, $j=1,\cdots,p_g$. 计算 α_{gj}:

   (a) 根据以下公式更新 α_{gj},

$$\alpha_{gj}^{\text{new}}=\arg\min_{\alpha_{gj}}\Big\{\frac{1}{2}\sum_{i=1}^n t_i\Big(\widetilde{Z}_i-\sum_{g,j}\widetilde{x}^d_{i,gj}\alpha_{gj}\Big)^2+\lambda\sum_{g=1}^G\omega_{gj}\sum_{j=1}^{p_g}|\alpha_{gj}|\Big\}.$$

   (b) 当 $\max(|\alpha_{gj}^{\text{new}}-\alpha_{gj}|)$ 足够小时, 令 $\alpha_{gj}=\alpha_{gj}^{\text{new}}$ 并终止迭代. 否则, 令 $\alpha_{gj}=\alpha_{gj}^{\text{new}}$ 并回到步骤 3(a).

4. 令 $\widetilde{x}^\alpha_{i,g}=\sum_{j=1}^{p_g}\alpha_{gj}x_{i,gj}$, $g=1,\cdots,G$.

   计算 d_g:

   (a) 根据下式更新 d_g

$$d_g^{\text{new}}=\arg\min_{d_g\geq 0}\Big\{\frac{1}{2}\sum_{i=1}^n t_i\Big(\widetilde{Z}_i-\sum_{g=1}^G\widetilde{x}^\alpha_{i,g}d_g\Big)^2+\sum_{g=1}^G d_g\Big\}.$$

   (b) 当 $\max(|d_g^{\text{new}}-d_g|)$ 足够小时, 令 $d_g=d_g^{\text{new}}$ 并终止迭代. 否则, 令 $d_g=d_g^{\text{new}}$ 并返回步骤 4(a).

5. 令 $\beta_{gj}^{\text{new}}=d_g\alpha_{gj}$. 若 $\|\boldsymbol{\beta}^{\text{new}}-\boldsymbol{\beta}\|_1$ 足够小, 则终止迭代. 否则, 令 $\boldsymbol{\beta}=\boldsymbol{\beta}^{\text{new}}$ 并返回步骤 2.

3.2 稀疏组 Lasso 的加速全梯度更新算法

加速全梯度更新算法 (AFGU) 大大加快了求解含稀疏组 Lasso 惩罚的模型的运算速度, 同时, 该方法还可推广至其他惩罚形式. 在广义线性模型中, 带稀疏组 Lasso 惩罚的参数 $\boldsymbol{\beta}$ 的 BD 估计可通过极小化如下目标函数求得:

$$T(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^n Q(Y_i,m(\boldsymbol{X}_i))+(1-\alpha)\lambda\sum_{g=1}^G\sqrt{p_g}\|\boldsymbol{\beta}^{(g)}\|_2+\alpha\lambda\|\boldsymbol{\beta}\|_1. \tag{4}$$

于是, 加速全梯度更新算法可通过以下过程实现:

1. 初始化 β_0 和 β_j, 得到 $\boldsymbol{\beta}$ 的初值 $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}^{(1)}, \cdots, \widetilde{\beta}^{(G)})$.

2. 组间循环更新每个 $\boldsymbol{\beta}^{(g)}$, 对于每个组，运行步骤 3.

3. (a) 基于全梯度更新的思想估计 $\widetilde{\boldsymbol{\beta}}^{(g)}$, 即极小化以下目标函数：

$$\frac{\rho}{2}||\widetilde{\boldsymbol{\beta}}^{(g)} - \widetilde{\boldsymbol{\beta}}^{(g)\text{old}}||_2^2 + \{\nabla Q(\widetilde{\boldsymbol{\beta}}^{(g)\text{old}})\}(\widetilde{\boldsymbol{\beta}}^{(g)} - \widetilde{\boldsymbol{\beta}}^{(g)\text{old}}) - P_{\lambda,\alpha}(\boldsymbol{\beta}^{(g)})$$

其中， $\nabla Q(\widetilde{\boldsymbol{\beta}}^{(g)})$ 为关于 $\widetilde{\boldsymbol{\beta}}^{(g)}$ 的梯度值； ρ 为一个很小的正常数，用于控制梯度更新步长. 相同于 Simon 和 Friedman 等人 [7] 的做法，更新每个组的 $\widetilde{\boldsymbol{\beta}}^{(g)}$,

$$\widetilde{\boldsymbol{\beta}}^{(g)\text{new}} = \Big(1 - \frac{\rho(1-\alpha)\lambda}{||S(\widetilde{\boldsymbol{\beta}}^{(g)} - \rho\nabla Q_{(-g)}(\widetilde{\boldsymbol{\beta}}^{(g)}, \rho\alpha\lambda))||_2}\Big)_+ S(\widetilde{\boldsymbol{\beta}}^{(g)} - \rho\nabla Q_{(-g)}(\widetilde{\boldsymbol{\beta}}^{(g)}), \rho\alpha\lambda),$$

其中 $Q_{(-g)}(\cdot)$ 为不含 g 组的损失函数值， $S(\cdot)$ 为软阈值函数 (Soft Thresholding), 即 $S(z_j, \alpha\lambda) = \text{sign}(z_j)(|z_j| - \alpha\lambda)_+$.

(b) 使用牛顿算法进行加速, 用 $\mathcal{A}$ 代表 $\widetilde{\boldsymbol{\beta}}^{(g)\text{new}}$ 中非零项的下标集合, 用 $\nabla Q_{\mathcal{A}}(\widetilde{\boldsymbol{\beta}}^{(g)\text{new}})$ 和 $\nabla^2 Q_{\mathcal{A}}(\widetilde{\boldsymbol{\beta}}^{(g)\text{new}})$ 分别表示函数 $Q(\widetilde{\boldsymbol{\beta}}^{(g)\text{new}})$ 关于 $\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}^{(g)\text{new}}$ 的既约梯度和海森矩阵. 于是 $\widetilde{\boldsymbol{\beta}}^{(g)}$ 可由下式更新获得：

$$\widetilde{\boldsymbol{\beta}}^{(g)} = \widetilde{\boldsymbol{\beta}}_{\mathcal{A}}^{(g)\text{new}} - \{\nabla^2 Q_{\mathcal{A}}(\widetilde{\boldsymbol{\beta}}^{(g)\text{new}})\}^{-1}\nabla Q_{\mathcal{A}}(\widetilde{\boldsymbol{\beta}}^{(g)\text{new}})$$

(c) 当 $\|\boldsymbol{\beta}^{(g)\text{new}} - \boldsymbol{\beta}^{(g)}\|_1$ 足够小时，终止迭代. 否则，令 $\widetilde{\boldsymbol{\beta}}^{(g)\text{old}} = \widetilde{\boldsymbol{\beta}}^{(g)\text{new}}$ 并返回步骤 2.

尽管已有研究已经讨论了坐标轴下降算法和加速全梯度更新算法的收敛性和适用性，但我们在现阶段的研究中暂时还无法给出二者在本文框架下严格的理论推导，我们将在下一章模拟研究中给出分层 Lasso 坐标轴下降算法和稀疏组 Lasso 加速全梯度更新算法的解路径图，以讨论算法的收敛性.

4　模拟计算

为了比较本文讨论的不同惩罚函数和变量分布对变量选择效果的影响，并且选择适用于本文关注的基因表达水平模型的惩罚函数, 本章节设计了如下两个模拟试验. 首先，构造含有组结构的解释变量. 为了模拟一般分组时，各组间具有不同特征，组内具有相似特征的组结构信息. 我们独立生成 325 个组变量，其中有 50 个组的 p_g 为 25, 有 275 个组的 p_g 为 50, 共计 $p = 15000$ 个解释变量. 为了模拟模型的稀疏性，我们设计了 10 个显著组共 80 个显著变量，其中，有 2 个含 5 个显著变量的组以及 3 个含 10 个显著变量的组来自于总变量个数为 50 的组；同样地，有 2 个含 5 个显著变量的组以及 3 个含 10 个显著变量的组来自于总变量个数为 20 的组. 在模拟一中，每个组内的变量 $\boldsymbol{X}^{(g)}$ 取自多元正态分布 $N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{p_g}})$, 其中 $\Sigma_{p_g} = \rho\mathbf{1}_{\mathbf{p_g}}\mathbf{1}_{\mathbf{p_g}}^{\top} + (\mathbf{1}-\rho)\mathbf{I}_{\mathbf{p_g}}$, $\mathbf{g} = \mathbf{1}, \cdots, \mathbf{G}$, $\rho = \mathbf{0.8}$.

为了考察自变量服从厚尾分布时，文中方法的表现，我们在模拟二中，使每个组内的变量 $\boldsymbol{X}^{(g)}$ 服从自由度为 3 的多元 t 分布 $t(3,\mathbf{0},\boldsymbol{\Sigma}_{\mathbf{p_g}})$, 其中， Σ_{p_g} 同模拟一. 模拟一和模拟二的被解释变量 Y 均服从 Bernoulli 分布，可根据以下模型生成：

$$m(\boldsymbol{x}) = E(Y|\boldsymbol{X}=\boldsymbol{x}) = F^{-1}(\beta_0 + \boldsymbol{x}^{\top}\boldsymbol{\beta}),$$

其中， $\theta = \beta_0 + \boldsymbol{X}^{\top}\boldsymbol{\beta}$ 且 $F(\theta) = \log\{\theta/(1-\theta)\}$. 在模拟一和模拟二中，分别重复上述步骤，随机生成一个样本量 $n = 140$ 的训练集，用来估计模型参数和变量选择. 同样地，随机生成一个样本量 $n = 10000$ 的测试集，用来测试模型变量选择的和分类预测的效果. 我们用 Oracle 代表当我们已知准确组结构信息及相应的非零重要变量的理想状态模型，分别用上文提到的 8 种不同惩罚函数构建带有组结构的稀疏模型进行参数估计. 每个模拟实验重复 100 次，比较以下指标以检验模型的估计效果：

- MR: 误分率，用训练集估计的参数拟合的模型预测测试集样本时分类错误的比率
- CSZ: 正确选中的回归系数为零的变量个数
- CSNZ: 正确选中的回归系数不为零的变量个数
- CSG: 正确选中显著组的个数
- WSG: 错误选择非显著组作为显著的个数
- Time: 计算机进行单次模型估计所需要的时间, 单位为秒. 计算机配置为：Intel(R) Core(TM) i7-4790 CPU@3.60GHz, RAM 8GB

在调节参数的选择方面，常用的方法主要有 AIC, BIC, EBIC, GCV 和最小误分率法. 用 $c = 1/2$ 代表相等的预测分类错误损失，则众所周知最优的分类方法为贝叶斯准则 $Y_B(\boldsymbol{X}) = I[m(\boldsymbol{x}) > 1/2]$, 于是误分率可表示为 $\mathrm{MR} = \frac{\#(Y \neq Y_B)}{\#Y}$, 于是可以选取使误分率 MR 最小的调节参数 λ_R 和 α_R 作为最终的模型调节参数. 对于稀疏组 Lasso 惩罚，为了减轻计算负荷, 我们采用 BIC 准则来选择调节参数. 一方面，BIC 方法和最小误分率方法选择的调节参数相似, 但 BIC 方法计算简单, 耗时短; 另一方面, 尽管当解释变量维数较高时，用 BIC 选择可能会使变量较多的模型比变量少的模型接受到更高的选择概率，从而使得变量选择的假阳性 (False discovery rate) 增高，且 Chen 和 Chen[18] 提出了可以较好的控制假阳性率的 EBIC 方法，即 $\mathrm{EBIC}_\gamma(s) = -2\log L_n\{\widehat{\beta}(s)\} + \nu(s)\log n + 2\gamma\log\tau(S_j)$, $0 \leq \gamma \leq 1$. 然而，该方法在实际应用中，会引入讨厌参数 γ, 这又带来了新的计算负担和挑战. 由于本文关注的实际问题是选出与疾病相关的基因，以便进行生物学通路的探究和开展后续的生物学验证试验，我们认为略高的假阳性率是可接受的. 因此，我们采用 BIC 方法选择稀疏组 Lasso 惩罚的调节参数：

$$\mathrm{BIC}_g(\boldsymbol{\beta}_g) = 2l_g(\boldsymbol{\beta}_g) + df(\boldsymbol{\beta}_g)\log(n)/n,$$

其中 $l_g(\cdot)$ 是 g 组的负对数似然函数， $df(\boldsymbol{\beta}_g)$ 是 g 组所有非零 β 的数量.

模拟一的计算结果如表 1 所示，括号内的值为标准误差．Oracle 方法展示的真实模型有 14920 个变量系数为 0, 80 个系数非零的显著变量，共 10 个显著组．由于岭回归惩罚方法无法将系数压缩为 0, 该方法将所有组的所有变量系数均估计为非 0, 未能识别出模型的稀疏性；同时，也导致岭回归在 8 种方法中预测误分率最高．光滑截断绝对差，Lasso, 自适应 Lasso 方法可以正确估计出部分参数，但由于他们无法利用组结构信息，估计结果对应真实组信息时的变量选择错误率较高，且光滑截断绝对差方法的预测误分率较高．组 Lasso 方法可有较低的预测误分率，但其稀疏性保证较差，组 Lasso 方法在正确估计出部分显著组和显著变量时，也错误的选择了较多的显著组和显著变量，变量选择结果较差．分层 Lasso 方法的变量选择结果相对于组 Lasso 方法有所改进，错误估计的显著组和显著变量大大减少，但其预测误分率较高．自适应分层 Lasso 通过权重的自适应调整，一方面保持了分层 Lasso 方法的稀疏性识别优势，对显著组和显著变量的系数估计更加准确，特别地，在本例中未错误估计出显著组；另一方面大大降低了预测的误分率，更准确的变量选择结果也带来了更精确的预测结果．稀疏组 Lasso 方法是所有方法中预测误分率最低的，同时也是正确估计出最多显著变量的，其错误估计出的组结构也较少．在计算耗时方面，普通 Lasso 方法由于忽略了组结构信息，运行速度较快．而利用了组结构信息的惩罚方法中，分层 Lasso 方法运行最快，自适应分层 Lasso 进行加权后速度稍稍减慢；稀疏组 Lasso 方法由于惩罚结构相对复杂，既含有 L_1 惩罚又含有 L_2 惩罚，运行速度较慢，但使用加速全梯度下降算法提速后，也仅仅只需要自适应分层 Lasso 方法的大约 5 倍左右时间；而组 Lasso 方法仅含有 L_2 惩罚，使用普通坐标下降方法需耗时近 4 小时，几乎是使用 AFGU 算法的稀疏组 Lasso 方法运行时间的 30 倍.

表 1　自变量来自正态分布时，不同惩罚函数模型模拟计算结果

Method	MR (std)	CSZ (std)	CSNZ (std)	CSG (std)	WSG (std)	Time (std)
Oracle	0.187 (0.021)	14920 (0.0)	80 (0.0)	10 (0.0)	0 (0.0)	2.3 (0.5)
Ridge	0.403 (0.009)	0 (0.0)	80 (0.0)	10 (0.0)	315 (0.0)	0.8 (0.0)
SCAD	0.251 (0.016)	14851 (4.9)	13 (2.5)	8 (1.1)	57 (4.4)	14.1 (0.1)
Lasso	0.185 (0.018)	14891 (16.7)	8 (2.3)	5 (1.0)	15 (12.2)	9.9 (0.3)
AdLasso	0.181 (0.018)	14903 (9.7)	7 (2.4)	5 (1.0)	7 (6.6)	57.0 (6.9)
GLasso	0.172 (0.019)	14347 (195.7)	34 (5.3)	6 (1.1)	26 (12.4)	13786.6 (549.4)
HLasso	0.303 (0.048)	14887 (9.0)	10 (3.1)	6 (1.3)	9 (3.1)	32.8 (2.1)
AdHLasso	0.199 (0.126)	14907 (14.4)	7 (4.4)	4 (1.7)	0 (1.1)	96.2 (17.8)
SGLasso	0.167 (0.045)	14785 (61.3)	40 (7.7)	4 (1.0)	1 (1.1)	507.9 (187.8)

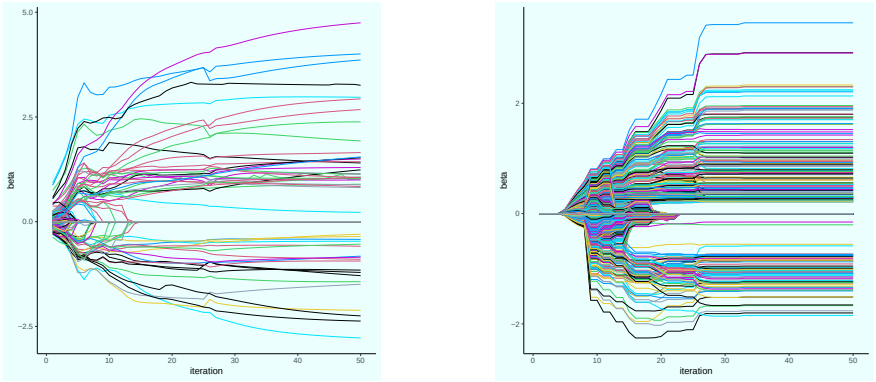
模拟二的计算结果如表 2 所示．当自变量 $\boldsymbol{X}$ 的分布为 t 分布时，文中所有方法的表现均比模拟一的略差，但各方法间的差别和趋势类似．岭回归惩罚方法无法识别出模型的稀疏性，且预测误分率最高；光滑截断绝对差， Lasso, 自适应 Lasso 方法无法利用组结构信息，造成组结构估计错误较多；组 Lasso 方法可以利用组结构信息，但其稀疏性保证较差，错误估计出了较多的显著组和显著变量．分层 Lasso 的稀疏性保证较好，但预测误分率较高，自适应分层 Lasso 通过权重的自适应调整使得组结构的错误识别率

和预测误分率均大大降低. 稀疏组 Lasso 方法的预测误分率仍较低, 但高于 Lasso, 自适应 Lasso 和组 Lasso 方法, 稀疏组 Lasso 方法是正确估计出最多显著变量和最少错误估计出组结构的方法. 在计算耗时方面, 各方法的运行时间差异与模拟一类似.

表 2 自变量来自 t 分布时, 不同惩罚函数模型模拟计算结果

Method	MR (std)	CSZ (std)	CSNZ (std)	CSG (std)	WSG (std)	Time (std)
Oracle	0.198 (0.023)	14920 (0.0)	80 (0.0)	10 (0.0)	0 (0.0)	4 (1.4)
Ridge	0.420 (0.010)	0 (0.0)	80 (0.0)	10 (0.0)	315 (0.0)	1 (0.2)
SCAD	0.286 (0.022)	14848 (4.5)	12 (2.3)	8 (1.0)	58 (4.0)	17 (1.4)
Lasso	0.209 (0.028)	14890 (19.0)	7 (2.4)	5 (1.1)	16 (13.2)	12 (0.9)
AdLasso	0.207 (0.054)	14903 (9.4)	6 (2.7)	5 (1.0)	8 (6.5)	55 (10.8)
GLasso	0.201 (0.029)	14277 (235.8)	33 (4.9)	6 (1.2)	29 (13.9)	29638 (2068.4)
HLasso	0.357 (0.054)	14884 (33.7)	10 (4.9)	5 (1.3)	10 (3.3)	31 (4.5)
AdHLasso	0.263 (0.121)	14905 (14.2)	6 (4.2)	3 (1.7)	1 (1.4)	84 (15.1)
SGLasso	0.228 (0.107)	14806 (85.4)	34 (14.0)	3 (1.5)	1 (1.3)	731 (216.2)

综上所述, 自适应分层 Lasso 和稀疏组 Lasso 惩罚方法在模拟一和模拟二中的表现均优于其它方法, 它们可以较好的识别出组结构和选择显著变量, 且稀疏性保证较好. 这两种惩罚方法可应用于本文关注的基因表达水平与疾病关联的问题, 更准确的筛选出与疾病相关的生物学通路以及各通路中起主要影响作用的基因. 特别是稀疏组 Lasso 方法, 该方法在模拟计算中表现出的低误分率可以更好的保证筛选出的基因和疾病的高关联性.



(a) 分层 Lasso (b) 稀疏组 Lasso

图 1 分层 Lasso 坐标轴下降算法和稀疏组 Lasso 加速全梯度更新算法的解路径图

为了说明第三章中的两种算法的收敛性, 我们以分层 Lasso 和稀疏组 Lasso 为例, 从模拟一中随机抽取了一组数据, 分别用分层 Lasso 的坐标轴下降算法和稀疏组 Lasso 的加速全梯度更新算法迭代 50 次求解, 绘制了两种算法的解路径图, 如图 1 所示. 可

以看出，分层 Lasso 的坐标轴下降算法在迭代到大约 15 次时收敛，非显著变量的系数被压缩至 0; 显著变量的参数估计值在小范围内波动．稀疏组 Lasso 的加速全梯度更新算法在迭代到大约 25 次时收敛，非显著变量的系数被压缩至 0；显著变量的参数估计值保持不变．在本文的计算中，为了在保证算法的收敛性的同时减轻计算负担，我们设置了最大迭代次数为 100 次，参数估计收敛准则为 1×10^{-3}, 即当参数两次迭代计算结果的差值小于 1×10^{-3} 时，我们认为算法收敛，不再继续计算．

5　应用

骨关节炎 (Osteoarthritis, OA) 是一种中老年人的常见病，发病者会出现关节疼痛，僵硬及关节功能性障碍，致使病人生活质量降低，甚至引发残疾．然而，由于骨关节炎的发病机制十分复杂且尚未明确，虽然临床治疗方法众多，但仍未有特别有效和治愈疾病的方法．随着分子生物学，细胞生物学等相关学科的发展和交叉渗透，基因表达谱分析和基因治疗成为了探究发病机制与病理进程，靶向治愈疾病的有效途径． Ramos 和 Bos 等人 [19] 开展了一项遗传性骨关节炎和其进展 (GARP) 的研究，他们用受试对象的外周血制备基因芯片．通过外周血单核细胞 (PBMC) 表达谱对比分析，从基因水平分析与骨关节炎发病密切相关的差异表达基因与信号通路，以便进一步研究蛋白水平和细胞水平的表达，从而阐释骨关节炎发病机制，寻找相应的骨关节炎治疗靶点．此项研究采集了 106 名遗传性骨关节炎患者和 33 名健康对照者共 139 例样本的外周血，对每个样本均进行基因表达谱分析，每个样本得到超过 40000 个基因的表达水平数据．染色体作为基因的载体，其临近的基因组元件之间存在密切的功能联系，共同参与通路或生物学过程．因此，可根据染色体位置对基因进行分组，分析单个基因以及基因集合的表达变化．本文根据 GSEA 官网的 MSigDB 基因集 [20] 给出的 C1 位置基因集合将基因芯片检测到的基因分成 326 组，共有 14839 个有效基因．我们分别使用模拟计算选出的自适应分层 Lasso 和稀疏组 Lasso 惩罚函数构造带组结构的逻辑斯蒂回归模型，并对模型进行估计．具体地，我们先对样本数据做 10 次五折交叉验证，选取调节参数 λ 和 α；再从样本中随机抽取 16 名健康对照者和 53 名骨关节炎患者作为训练集，余下样本作为测试集，以进行模型估计和评估，结果如表 3 所示．

表 3　骨关节炎相关基因选择结果

	误分率	显著基因个数	显著组个数
adHLasso	0.3000	23	8
SGLasso	0.1429	136	9

由表 3 可以看出，自适应分层 Lasso 选出了 8 个基因集合共 23 个显著基因，但其误分率较高，为 0.3. 稀疏组 Lasso 选出了 9 个基因集合共 136 个显著基因，且误分率仅为 0.1429. 因此，我们最终选择带有稀疏组 Lasso 惩罚的逻辑斯蒂模型进行骨关节炎的基因水平表达分析．我们发现，该方法选出的 136 个基因中，有三个基因和 Ramos 和 Bos 等人 [19] 的研究结果相同，它们为： KCNJ2, ADRB2 和 SNORD13. 现有研究发

现，由该 KCNJ2 编码的蛋白质在骨骼肌中表现活跃，且有可能参与骨发育. Asmar 和 Barrettjolley 等人 [21] 发现 KCNJ2 基因在胎儿和青少年的软骨细胞中存在差异表达，提示该基因可能和骨关节炎发病相关. Jiao 和 Niu 等人 [22] 利用 β- 拮抗剂和激发剂研究 ADRB2 在骨的间充质干细胞的信号传导，发现其介导下的软骨骨质流失可能和骨关节炎发病相关. SNORD13 属于小核仁 RNA, 在人类细胞的基因表达水平调控中有充当着重要角色，小核仁 RNA 水平与癌症，退行性疾病等相关 [23]. 由此可见，我们的模型变量选择结果具有实际应用意义，我们选出了与骨关节炎密切相关的基因，其中部分基因的作用机制已被证实. 因此，其余未见报道的基因具有极大的研究价值，特别是研究这些基因所在基因集合的整体功能和作用，进一步研究阐释其在细胞的生命周期，生化调控路径，蛋白质交互作用的作用机制，有可能进一步揭示骨关节炎的发病机制.

参 考 文 献

[1] Tibshirani R. Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996, 58: 267–288

[2] Hoerl A E., Kennard R W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 2000, 42(1): 80–86

[3] Zou H. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 2006, 101(476): 1418–1429

[4] Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 2001, 96(456): 1348–1360

[5] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 2006, 68(1): 49–67

[6] Zhou N, Zhu J. Group variable selection via a hierarchical lasso and its oracle property. *Statistics and Its Interface*, 2010, 3(4): 557–574

[7] Simon N, Friedman J H, Hastie T, Tibshirani R. A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 2017, 22(2): 231–245

[8] Zhang C, Jiang Y, Shang Z. New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation. *Canadian Journal of Statistics*, 2009, 37(1): 119–139

[9] Zhang C, Jiang Y, Chai Y. Penalized Bregman divergence for large-dimensional regression and classification. *Biometrika*, 2010, 97(3): 551–566

[10] Zhang C, Zhang Z, Chai Y. Penalized Bregman Divergence Estimation via Coordinate Descent. *Journal of the Iranian Statistical Society*, 2011, 10(2): 125–140

[11] Hastie T, Tibshirani R, Friedman J H. The elements of statistical learning: data mining, inference, and prediction. *The Mathematical Intelligencer*, 2001, 27(2): 83–85

[12] Zhang C, Guo X, Cheng C, Zhang Z. Robust-BD estimation and inference for varying-dimensional general linear models. *Statistica Sinica*, 2014, 653–673

[13] Hennequin R, David B, Badeau R. Beta-divergence as a subclass of Bregman divergence. *IEEE*

Signal Processing Letters, 2011, 18(2): 83–86

[14] Mihoko M, Eguchi S. Robust blind source separation by beta divergence. *Neural computation*, 2002, 14(8): 1859–1886

[15] Friedman J H, Hastie T, Hofling H, Tibshirani R. Pathewise coordinate optimization. *The Annals of Applied Statistics*, 2007, 1(2): 302–332

[16] Wu T T, Lange K. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2008, 2(1): 224–244

[17] Zhang C, Chai Y, Guo X, Gao M, Devilbiss D M, Zhang Z. Statistical Learning of Neuronal Functional Connectivity. *Technometrics*, 2016, 58(3): 350–359

[18] Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 2008, 95(3): 759–771

[19] Ramos Y F, Bos S D, Lakenberg N, Bohringer S, Den Hollander W, Kloppenburg M, Slagboom P E, Meulenbelt I. Genes expressed in blood link osteoarthritis with apoptotic pathways. *Annals of the Rheumatic Diseases*, 2013, 73(10): 1844–1853

[20] Subramanian A, Tamayo P, Mootha V K, Mukherjee S, Ebert B L, Gillette M A, Paulovich A, Pomeroy S L, Golub T R, Lander, E S, Mesirov J P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 2005, 102(43): 15545–15550

[21] Asmar A, Barrettjolley R, Werner A, Kelly R E, Stacey M. Membrane channel gene expression in human costal and articular chondrocytes. *Organogenesis,* 2016, 12(2): 94–107

[22] Jiao K, Niu L N, Li Q H, Ren G T, Zhao C M, Liu Y D, Tay F R, Wang M Q. β 2-adrenergic signal transduction plays a detrimental role in subchondral bone loss of temporomandibular joint in osteoarthritis. *Scientific Reports*, 2015, 5: 12593, doi: 10.1038/srep12593

[23] Stepanov G A, Filippova J A, Komissarov A B, Kuligina E V, Richter V A, Semenov D V. Regulatory Role of Small Nucleolar RNAs in Human Diseases. *BioMed Research International*, 2015, 206849, doi: 10.1155/2015/206849

Estimation and Variable Selection on Sparse Model with Group Structure

ZHANG Yunqi　　ZHANG Chunming†

(*Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, Kunming* 650091, *China*)

(*Department of Statistics, University of Wisconsin-Madison, Madison* 53705, *USA*)

(† *E-mail: cmzhang@stat.wisc.edu*)

TANG Niansheng

(*Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, Kunming* 650091, *China*)

Abstract We introduce the Bregman divergence as a general loss function for the generalized linear sparse model with group structures so that the parameter estimation and variable selection methods are not limited to a specific model or a specific loss function. We compare the characteristics of eight kinds of penalty functions, such as Ridge, SACD, Lasso, Adaptive Lasso, Group Lasso, Hierarchical Lasso, Adaptive Hierarchical Lasso and Sparse Group Lasso, and the methods of parameter estimation and variable selection with these penalties. The Coordinate Descent algorithm for Hierarchical Lasso and the Accelerated Full Gradient Update algorithm for Sparse Group Lasso are also detailed. The simulation study shows that the Group Lasso, Hierarchical Lasso, Adaptive Hierarchical Lasso, and Sparse Group Lasso can better utilize the group structure information of the data, Adaptive Hierarchical Lasso and Sparse Group Lasso in terms of variable selection accuracy and parameter estimation accuracy. Compared with other methods, the Sparse Group Lasso is optimal in model prediction accuracy. As an empirical example, we apply a logistic model with Sparse Group Lasso penalty to the analysis of gene expression levels in peripheral blood mononuclear cells of patients with osteoarthritis and selected 136 genes in 9 gene sets which affect osteoarthritis, in order to have a certain guiding value for the follow-up biomedical research.

Key words Lasso; Bregman divergence; group structure; generalized linear model; sparse model

MR(2000) Subject Classification 62J12; 62P10

Chinese Library Classification O212.1