第 32 卷第 11 期
2015 年 11 月
统计研究
Statistical Research
Vol. 32, No. 11
Nov. 2015

大数据的整合分析方法*

马双鸽　王小燕　方匡南

内容提要: 大数据具有数据来源差异性、高维性及稀疏性等特点,如何挖掘数据集间的异质性和共同性并降维去噪是大数据分析的目标与挑战之一。整合分析(Integrative Analysis)同时分析多个独立数据集,避免因地域、时间等因素造成的样本差异而引起模型不稳定,是研究大数据差异性的有效方法。它的特点是将每个解释变量在所有数据集中的系数视为一组,通过惩罚函数对系数组进行压缩,研究变量间的关联性并实现降维。本文从同构数据整合分析、异构数据整合分析以及考虑网络结构的整合分析三方面梳理了惩罚整合分析方法的原理、算法和研究现状。统计模拟发现,在弱相关、一般相关和强相关三种情形下,L_1 Group Bridge、L_1 Group MCP、Composite MCP 都表现良好,其中 L_1 Group Bridge 的假阳数最低且最稳定。最后,将整合分析用于研究具有来源差异性的新农合家庭医疗支出,以及具有超高维、小样本等大数据典型特征的癌症基因数据,得到了一些有意义的结论。

关键词: 大数据; 整合分析; 变量选择; 医疗支出; 癌症基因
中图分类号: C829.2　　**文献标识码:** A　　**文章编号:** 1002 - 4565(2015)11 - 0003 - 09

Integrative Analysis for Big Data

Ma Shuangge　Wang Xiaoyan　Fang Kuangnan

Abstract: The difference of data source, high dimensionality and sparsity are the main characteristics of big data. How to mining the heterogeneity and association of different datasets and achieve dimension reduction is one of goals and challenges of big data analysis. Integrative analysis provides an effective way of analyzing big data. It simultaneously analyzes multiple datasets, avoiding the model instability from individual variations caused by regional and time factor and so on. The coefficients of each covariate across all datasets are treated as a group and penalty function is used to shrinkage these groups of coefficients to achieve variable selection. In this paper, we review the existing research of penalized integrative analysis from three aspects of homogeneity integrative analysis, heterogeneity integrative analysis and network integrative analysis. Three simulations are conducted to verify the performance of integrative analysis, including weak, moderate and strong correlations. It shows that L_1 Group Bridge、L_1 Group MCP、Composite MCP perform well, and L_1 Group Bridge has the lowest false positive and is most stable. Finally, integrative analysis is applied to analyze the new rural cooperative medical expenditure data with source difference, as well as cancer genetics data with typical characteristics of big data such as super high dimension and small sample.

Key words: Big Data; Integrative Analysis; Variable Selection; Medical Expenditure; Cancer Genetics Data

一、引言

21 世纪是信息爆炸的时代,计算机技术的飞速发展,极大地方便了数据的获取和储存,使得很多部门每天都有大量的数据产生。大数据通常是由来源、主体或格式不同的数据合并而成,例如来自不同地区的调查数据,来自不同市场的金融数据,来自不同实验室的基因数据等。这种基于多个数据集的建模十分常见,了解不同子样本间的异质性(heterogeneity or difference)和同质性(homogeneity or similarity)是大数据分析的两个重要目标[1]。但它的建模比较

* 本文获国家统计局重大项目"大数据的统计方法研究"(2012LD001)、国家统计局重点项目"大数据线性、理论及处理技术的发展和创新研究"(2013LZ53)、国家社会科学基金重大项目"大数据与统计学理论的发展研究"(13&ZD148)、国家社会科学基金青年项目"大数据的高维变量选择方法及其应用研究"(13CTJ001)以及国家自然科学基金面上项目"广义线性模型的组变量选择及其在信用评分中的应用"(71471152)资助。

特殊，一方面，由于不同来源的数据存在差异，各不同数据源的同一变量的系数显著性和估计值可能存在差异，传统的处理方法是简单合并所有样本，建立统一模型，但是这种方法过于笼统，忽略了数据间的异质性(heterogeneity) ; 另一方面，也不能分开各自建立模型，因为这样会忽略各个数据集间的关联性。整合分析(Integrative Analysis) 方法同时兼顾这两方面，通过目标函数综合不同地区的数据，从统计角度考虑数据的异质性和同质性，以多个变量为研究目标，充分考虑了不同地区间相互影响，同时求解多个模型。整合分析方法起源于 20 世纪 60 年代，把不同来源、格式、特点性质的数据集中起来，相对于单一数据集模型，整合了更多的原始信息，能解决因地域、时间等因素造成的样本差异而引起的建模不稳定，在模型解释性和预测方面都具有显著优势。

整合分析也是解决“大 p 小 n”问题的有效方法。它综合多个数据集而增加了样本量，是解决小样本问题的有效途径。该问题在大数据中亦十分常见，一方面源于大数据的稀疏性、价值密度低，即信息的边际价值并未随数据量增加而提升; 另一方面是大数据的高维性突出[1]，互联网和云计算为数据的获得和存储带来便利，与研究现象相关的微小因素都可能被收集起来，维度自然会很高，“去噪提纯”是亟待解决的问题。基于惩罚方法的整合分析(Penalized Integrative Analysis) 是将惩罚变量选择方法与整合分析结合，是降维和提取信息的有效方式，不仅能对模型进行选择，还能分析数据集间的关联性，以便更好地识别信号和噪音。鉴于大数据的来源差异性、高维性、稀疏性等特点，如何对其充分利用和综合分析比新技术更为重要，因此非常有必要在大数据时代下研究不同数据集的整合分析。

在单数据集变量选择中，惩罚方法是最为广泛使用的一类方法，它通过对未知参数的值进行压缩，同时实现变量选择和参数估计，具有降低估计偏差、提高预测精度和模型可解释性的优点。其研究可追溯到 Lasso(Tibshirani，1996) 的提出，它颠覆了逐步回归、最优子集、模型选择等贪婪方法，以压缩的角度实现自动识别。此后，学者提出了多种基于惩罚的变量选择方法，根据其特点可分为 4 类: 只能选择单个变量的单变量选择方法(Individual Variable Selection) ，如 Lasso(Tibshirani，1996) 、SCAD(Fan 和 Li，2001) 、MCP (Zhang，2007) 、Bridge (Frank 和 Friedman，1993) ; 高度相关数据的变量选择方法，如弹性网(Zou 和 Hastie，2005) 、Mnet (Huang 等，2010) ，在一定程度上能解决共线性问题; 组选择方法，Group Lasso(Yuan 和 Lin，2006) [2]、CAP(Zhao 等，2009) 等，对以组形式出现的变量进行选择; 双层选择方法如 Sparse Group Lasso (Simon 等，2013) [3]，L_1 Group Bridge(Huang 等，2009) [4] 等，在变量组内和组间实现双层选择。

整合分析依旧借鉴单数据集变量选择的思想，特殊之处在于整合分析中解释变量的回归系数不再是一个而是一组，不仅要筛选出显著的变量，还要识别出它在哪些数据集中显著，问题变得更加复杂。本文是作者在该领域多年的研究成果基础上，对整合分析从函数构成上进行归纳、梳理，将整合分析分为同构数据、异构数据以及考虑网络结构下的整合分析，通过统计模拟，对各种方法进行了比较，并将这些方法应用到我国家庭医疗支出调查分析和癌症基因分析中。

本文剩余部分安排如下: 第二部分介绍模型的基本结构; 第三部分介绍同构数据、异构数据下的惩罚整合分析方法，并对两者都适用的网络惩罚方法进行原理分析; 第四部分介绍了整合分析的常用算法——组坐标下降法的思路和流程，并对调整参数选择的常用方法做了概述; 第五部分做了 3 个模拟分析，对各种方法进行了比较; 第六部分将整合分析应用于两个实际问题中，分析来源差异性的家庭医疗支出数据，以及具有超高维、小样本等特征的基因数据，并从预测角度验证模型的有效性; 第七部分总结全文。

二、模型基本形式

整合分析不仅适合分析多个独立的数据集，还能分析具有多元互相关联因变量的单一数据集。研究思路大同小异，本文以前者为例展开分析。

假设有 M 个数据集，p 个解释变量。第 m 个数据集的样本量为 $n^{(m)}$，因变量 $y^{(m)}$ 为 $n^{(m)} \times 1$ 向量，连续型和离散型均可，解释变量 $X^{(m)}$ 为 $n^{(m)} \times p$ 矩阵，并假设数据已被标准化。为了阐述方便，本文设因变量为连续型变量，考虑最简单的线性回归，对第 m 个数据集建立如下模型:

$$y^{(m)} = X^{(m)}\beta^{(m)} + \varepsilon^{(m)} \quad (1)$$

其中，$\beta^{(m)} = (\beta_1^{(m)}, \cdots, \beta_p^{(m)})^T$ 为回归系数;

$\varepsilon^{(m)}$ 为随机项,满足 $E(\varepsilon^{(m)}=0)$、$\mathrm{var}(\varepsilon^{(m)})=\sigma^2_{(m)}$。记解释变量 X_j 在所有数据集中的回归系数为 $\beta_j=(\beta_j^{(1)},\cdots,\beta_j^{(m)})$。与单数据集模型相比,这 M 个模型变量的显著性有其特殊之处: 每个变量具有 M 个回归系数,它们归属于同一解释变量故会存在某种关联性或相似性,无法分别做参数估计和变量选择,否则会忽略这种关联; 它们的显著性不尽相同,亦不能简单地综合做估计。惩罚整合分析正是充分利用了这种特殊性来研究数据的差异,模型一般形式为:

$$\beta=\arg\min_\beta\{L(X,y;\beta)+P(\beta;\lambda)\} \quad (2)$$

其中 $y=(y^{(1)\prime},\cdots,y^{(M)\prime})'$ 是 $\sum_{m=1}^{M}n^{(m)}\times 1$ 因变量,$X=diag(X^{(1)},\cdots,X^{(M)})$ 是 $\sum_{m=1}^{M}n^{(m)}\times Mp$ 设计矩阵,$\beta=(\beta^{(1)\prime},\cdots,\beta^{(M)\prime})'$ 是 Mp 维未知参数向量。$L(X,y;\beta)$ 是建立在所有数据集上的损失函数,通常可表示为 $L(X,y;\beta)=\sum_{i=1}^{M}L(X^{(m)},y^{(m)};\beta^{(m)})$,$L(\cdot)$ 可取对数函数的负向变换、最小二乘函数等,下文分析以最小二次函数为例,即 $L(X,y;\beta)=(y-X\beta)'(y-X\beta)$。$P(\lambda;\beta)$ 是惩罚函数,通过调整参数 λ 的值平衡模型的拟合度和复杂度,估计参数并同步实现变量选择。λ 越大,$P(\lambda;\beta)$ 的值越大,参数 β 被压缩得越严重,估计为零的参数也就越多; 反之,λ 值越小,惩罚函数不足以将回归系数压缩为零,估计的参数非零的也就越多。因此如何合理地确定 λ 的值极为重要。

三、惩罚整合分析方法

根据数据产生背景中蕴含的先验信息,数据集可分为同构型(homogeneity)和异构型(heterogeneity),本文将分别介绍这两类数据的惩罚整合分析方法,同时概述了两者在考虑网络结构(network)关系下的惩罚方法。整合分析的回归系数具有两层含义: 第一是变量层面,这与普通的单数据集模型一致; 其次是数据集层面,同一个解释变量具有 M 个回归系数,各数据集的关联正是通过这些回归系数连接。这也是整合分析的特殊之处,变量的显著性不再是针对一个回归系数,而是一组回归系数,因此需要特殊的变量选择方法。

(一) 同构数据的整合分析

同构数据模型中,解释变量在 M 个模型中的显著性是一致的,每个模型具有相同的显著变量,即若 X_j 在数据集 m 中显著,则它在所有数据集中都显著。同构数据常见于调查问卷相同、实验设计相同等数据收集方式一致的情形中,在这种先验信息下,建立的同构模型显然会减少未知参数个数,降低计算量,模型结构也将更简洁。同构模型的性质可表示为:

$$I(\beta_j^{(1)}=0)=\cdots I(\beta_j^{(M)}=0),\ j=1,\cdots,p \quad (3)$$

从式(3)可知,向量 β_j 中各元素要么全为 0,要么全非 0。若将同一变量的 M 个系数视为一组参数,那么同构模型的变量选择为整组选择,只需组间选择,无需组内选择,具有"all-in-all-out"的特点。

同构数据的惩罚整合分析思想与单个数据集下的组选择类似,包含两层嵌套的惩罚函数,由组间惩罚 P_{outer} 和组内惩罚 P_{inner} 构成,具体形式为:

$$P(\beta;\lambda)=P_{outer}\Big(\sum_{k=1}^{p_j}P_{inner}(|\beta_k^{(j)}|;\lambda)\Big) \quad (4)$$

该惩罚函数的特点之一是组间 P_{outer} 惩罚函数具有变量选择功能,特点之二是组内 P_{inner} 只能压缩而无选择变量功能,通常组内 P_{inner} 用 Ridge 惩罚函数(Hoerl 和 Kennard,1970),利用它无法将系数压缩至 0 的特点,保证了同组回归系数同时非 0。这两个特点也是实现整组选择而不在组内选择的原理。与单数据集的不同之处在于整合分析的组是同一个变量在不同数据集上的多个回归系数,每组仅对应一个解释变量,而后者的组由多个虚拟变量或者解释变量群构成。常用方法有 L_2 Group Bridge、L_2 Group MCP 等。

1. L_2 Group Bridge。

Ma 等(2011a)[5] 在 Logistic 回归中提出复合型方法 L_2 Group Bridge,建立同构数据模型。惩罚函数为组内 Ridge、组间 Bridge,形式为:

$$P(\beta;\lambda,\gamma)=\lambda\sum_{j=1}^{p}\|\beta_j\|^\gamma=\lambda\sum_{j=1}^{p}\Big(\Big(\sum_{i=1}^{M}(\beta_j^{(i)})^2\Big)^{1/2}\Big)^\gamma$$

其中 $0<\gamma<1$。文中以 Group Lasso 估计作为初始值进行迭代估计,并从理论上证明了 Group Lasso 会选择过多,但 L_2 Group Bridge 满足选择一致性[6]。Ma 等(2012)[7] 又将 L_2 Group Bridge 用到了 AFT(Accelerated Failure Time)模型,并从理论上证明了选择一致性。

2. L_2 Group MCP。

L_2 Group MCP 最早用于单数据集中连续型因变

量建模[8,9]，解决以组形式出现的变量选择问题。Ma 等(2011b)[10]首次将其用于整合分析，分析复杂的删失生存数据。它的惩罚函数结构为组内 Ridge、组间 MCP，形式为:

$$P(\beta;\lambda,\alpha) = \sum_{j=1}^{p} P_{MCP}(\|\beta_j\|;\lambda,a)$$

其中 $P_{MCP}(\cdot)$ 为 *MCP* 惩罚，属于二次样条型惩罚，形式为:

$$P_{MCP}(\theta;\lambda,\alpha) = \begin{cases} \lambda\theta - \dfrac{\theta^2}{2a}, & \theta \leqslant a\lambda \\ \dfrac{a\lambda^2}{2}, & \theta > a\lambda \end{cases}$$

$$P'_{MCP}(\theta;\lambda,\alpha) = \begin{cases} \lambda - \dfrac{\theta}{a} & \theta \leqslant a\lambda \\ 0 & \theta > a\lambda \end{cases}$$

其中 a 是正则化参数，用于控制函数的凹性。MCP 计算简单因而在单数据集分析中备受欢迎。Liu 等(2014)[11]的研究中再次提到了同构模型下的 L_2 Group MCP，并将其作为模拟分析中的主要方法。

3. Group Lasso。

Group Lasso 是单数据集中最早的群组变量选择方法。它也适合同构模型的整合分析，但未得到系统研究，仅在 Zhang 等(2015)[12]的研究中，有简单的分析和理论论证。惩罚函数形式为:

$$P(\beta;\lambda) = \lambda\sum_{j=1}^{P} \|\beta_j\|$$

该文并未提出新的方法，而是从理论上证明了已有方法的性质，证明了在一定条件下 Group Lasso、L_2 Group SCAD、L_2 Group MCP 满足选择一致性。

总结同构模型方法，先验信息确定了同一解释变量在所有数据集中显著性一致，故将它的 M 个回归系数视为一组，它不再是鉴别变量组，而是识别在所有数据集中都显著的单个解释变量。因此 L_2 Group SCAD、CAP、adaptive Group Lasso(Wang 和 Leng，2006)等在单数据集中具有组选择功能的方法预期也是适用的。

(二)异构数据的整合分析

与同构数据模型不同的是，异构数据模型中解释变量在 M 个数据集中的显著性不一定相同，即对给定的 j，$I(\beta_j^{(m)} = 0)\ (m = 1,\cdots,M)$ 可以不全相等。异构数据模型更一般化，同构数据模型可以看作是异构数据模型的特殊情形。这类模型中变量显著性不一致通常有两方面的原因: 一是各数据集的产生方式(或环境因素)引起的变量显著性差异，如不同地区、不同时间点的数据集; 二是研究问题的细分，如同种疾病的不同子类别数据。异构模型的变量选择不仅仅要考虑解释变量是否显著，还要考虑它在哪些模型中显著，因此涉及到双层选择。已有的方法可分为复合惩罚类和稀疏组惩罚类。

1. 复合惩罚类。

复合惩罚函数形式如式(4)，与同构数据不同的是，此处组内和组间函数都具有单变量选择效果，组内不再是诸如 Ridge 等不能选择变量的函数。如 L_1 Group MCP[11]:

$$P(\beta;\lambda,a) = \sum_{j=1}^{p} P_{MCP}(\sum_{m=1}^{M} |\beta_j^{(m)}|;\lambda,a)$$

组内是 Lasso，组间是 MCP 函数。Lasso 形式简单，计算易实现，但是在单数据集变量选择中，它倾向选择过多的变量，理论上不满足 Oracle 性质(Fan 和 Li，2001)，效果不如 MCP。因此 Liu 等(2014a)又提出了 Composite MCP 惩罚，它的组内、组间都是 MCP 函数，惩罚函数为:

$$P(\beta;\lambda,a,b) = \sum_{j=1}^{p} P_{MCP}(\sum_{m=1}^{M} P_{MCP}(|\beta_j^{(m)}|;\lambda,a);\lambda,b)$$

Composite MCP 的理论性质比 L_1 Group MCP 更好，Zhang 等(2015)证明了在一定条件下，Composite MCP 在组内和组间均满足选择一致性，而 L_1 Group MCP 只满足组选择一致性。

在单数据集的双层选择中，L_1 Group Bridge(Huang 等，2009)是最早的方法，而将其用于整合分析是在 Shi 等(2014)[13]的研究中。L_1 Group Bridge 组内是 Lasso 函数、组间是 Bridge 函数，因此实现了两层选择。惩罚函数为:

$$P(\beta;\lambda) = \lambda\sum_{j=1}^{p} p_j \|\beta_j\|_l^{\gamma}$$

2. 稀疏组惩罚类。

稀疏组惩罚是两个惩罚函数的线性组合，一个具有组选择功能，另一个具有单变量选择功能，两者共同实现两层选择。一般形式为:

$$P(\beta;\lambda_1,\lambda_2) = \lambda_1\sum_{j=1}^{p} P_1(\|\beta_j\|) + \lambda_2\sum_{j=1}^{p}\sum_{m=1}^{M} P_2(|\beta_j^{(m)}|)$$

其中，函数 $P_1(\cdot)$ 作用在系数组上，具有组选择功能，无法在组内选择，因而能够选择重要的解释变

量; 函数 $P_2(\cdot)$ 作用在每一个系数上,能够进行单个系数选择,故能识别解释变量在哪些数据集中显著。Zhang 等(2015) 从理论上证明了这类方法的选择一致性。并建立 Sparse Group MCP 函数($P_1(\cdot)$ 和 $P_2(\cdot)$ 均为 MCP 惩罚),模拟分析了它的整合分析效果。

在单数据集分析中,已有学者提出了稀疏组惩罚方法 Sparse Group Lasso(SGL)(Simon 等,2013) 和 adaptive Sparse Group Lasso(adSGL)(Fang 等,2014)[14]。这两者的惩罚函数形式分别为:

$$P_{SGL}(\beta;\lambda_1,\lambda_2) = \lambda_1\sum_{j=1}^{p}\|\beta_j\| + \lambda_2\|\beta\|_1$$

$$P_{adSGL}(\beta;\lambda_1,\lambda_2) = \lambda_1\sum_{j=1}^{p} w_j\|\beta_j\|_2 + \lambda_2\xi^T|\beta|$$

SGL 是 Lasso 和 Group Lasso 的线性组合,两者在理论上都不满足 Oracle 性质,预期 SGL 也不满足,因此 Fang 等(2014) 提出了更一般化的 adSGL,通过引入组权重 w 和单个系数权重 ξ,改进选择一致性和估计一致性。两个权重都由数据本身决定,与系数的真实值成反比,真实值越大,权重越小,压缩越轻,估计越接近真实值。SGL 和 adSGL 都是 Lasso 型惩罚,形式简单,计算易实现,可直接用组坐标下降法求解。这两种方法尚未用于异构数据的整合分析,但预期也是可行的。

(三) 考虑网络结构关系的整合分析

传统的计量建模中通常假设各观察项是相互独立的,但是在大数据时代各个变量间往往相互关联,变量或回归系数两两之间会存在相互影响,形成一张网络结构图。以上方法考虑了变量在不同数据集中的显著性关系,并未考虑回归系数之间的关联。同一数据集中不同解释变量可能会相互作用,表现为它们在同一数据集中的系数具有某种关系,这称为数据集内部结构(within-dataset structure)。不同数据集具有相同的解释变量甚至因变量,因此有理由相信,同一解释变量在不同数据集中的系数存在某种相似性,称为跨数据集结构(across-dataset structure)。

Liu 等(2013)[15] 建立了数据集内部结构下的网络结构惩罚方法,惩罚函数为:

$$P(\beta;\lambda) = \lambda\sum_{1\leqslant j,k\leqslant p} a_{jk}\left(\frac{\|\beta_j\|_2}{\sqrt{M_j}} - \frac{\|\beta_k\|_2}{\sqrt{M_m}}\right)^2 \quad (5)$$

该惩罚函数针对数据集内部结构,将解释变量的 M 个系数作为一个整体,惩罚其 L_2 范数差。其中 a_{jk} 为权重,若变量 X_j 与 X_k 越相似,则惩罚越重,那么 β_j 与 β_k 的 L_2 范数差越小,它们的估计值越相近。Liu 等(2013) 将提出的惩罚与 L_2 Group MCP 结合,用于同构数据的建模。

Shi 等(2014) 研究跨数据结构,提出了 Contrast 惩罚,通过对回归系数的差进行惩罚,解决系数相似性问题。Contrast 惩罚函数为:

$$P_C(\beta) = \lambda\sum_{j=1}^{p}\sum_{k\neq l} a_j^{(kl)}(\beta_j^{(k)} - \beta_j^{(l)})^2 \quad (6)$$

它惩罚同一变量在不同数据集中的系数值之差,式(6) 中 $a_j^{(kl)} = I(\mathrm{sgn}(\beta_j^{(k)}) = \mathrm{sgn}(\beta_j^{(l)}))$,若 $\mathrm{sgn}(\beta_j^{(k)}) = \mathrm{sgn}(\beta_j^{(l)})$,则变量 X_j 在数据集 k 和 l 中的系数越相似; 若 $\mathrm{sgn}(\beta_j^{(k)}) \neq \mathrm{sgn}(\beta_j^{(l)})$,$X_j$ 在这两个数据集中的系数符号相反,因此不存在相似性,相应的 Contrast 惩罚值为零。估计 $\mathrm{sgn}(\beta_l^{(k)})$ 的方法可有多种,具体可参见文献 Shi 等(2014)。Contrast 惩罚与 L_2 Group Bridge、L_1 Group Bridge 组合,可分别用于同构数据和异构数据的建模。

四、计算

(一) 算法

对于惩罚整合分析的计算,最常用的优化方法是组坐标下降法(Group Coordinate Descent,GCD)(Yuan 和 Lin,2006)。GCD 是坐标下降法(Coordinate Descent,CD)(Fu 等,1998) 在组结构下的扩展,它的思想是在固定其他参数的情形下,每次迭代只优化一组参数,直到所有参数收敛到给定精度。GCD 在单数据集组变量选择方法中十分常用,最早出现在线性模型的 Group Lasso 求解,Meier 等(2008) 也用该算法求解 Logistic 回归下的 Group Lasso,其中损失函数用二次函数逼近。在最小二乘框架下,其基本流程如下(Zhao 等,2015)[16]:

步骤 1: 给定初始值 $\beta^{[0]} = (\beta_0^{(1)T},\cdots,\beta_0^{(J)T}$ 和收敛精度,记已循环次数 $s = 0$,计算当前残差 $r = y - X\beta^{[0]}$。

步骤 2: 对每个 $j \in (1,\cdots,p)$,固定 $\beta_k^{[s]}$ $(k \neq j)$,对 $\beta_j = (\beta_j^{(1)},\cdots,\beta_j^{(M)})'$ 进行估计。

① 计算 $z_j = \frac{X_j'r}{n} + \beta_j^{[s]}$,其中 X_j 是设计阵中与 β_j 有关的子矩阵;

② 更新 $\beta_j^{[s+1]} \leftarrow F(z_j,\lambda)$,$F(z_j,\lambda)$ 是由目标函

数求解而得的优化式，例如由偏导数为零得到参数更新方程；

③ 更新当前残差：$r \leftarrow r - X_j(\beta_j^{[s+1]} - \beta_j^{[s]})$。

步骤 3：更新 s 为 $s+1$。

步骤 4：重复步骤 2、步骤 3 直到收敛。

该算法的收敛性在 Tseng(2001) 中有严格的论证。当目标函数为严格凸函数时，显然会得到全局最优解。而以上方法的目标函数并不满足凸性，只有损失函数满足该性质，因此 Tseng 证明了即便如此，只要目标函数的不可微部分(惩罚函数)是可分的，算法就会收敛。以 Group Lasso 为例，最小二次函数作为损失函数时，$L(\beta; y, X)$ 为严格凸函数，而惩罚函数 $P(\beta;\lambda)$ 不可微，但是它在组之间是可分的，即可拆分为 $P(\beta;\lambda)\sum_{j=1}^{p} f_\lambda(\beta_j)$，因此 GCD 算法在该问题中是收敛的。

(二) 调整参数的选择

调整参数 λ 连接损失函数和惩罚函数，其取值直接影响建模效果。在选择最优值之前，通常要确定 λ 的大致范围，以减少计算成本并提高建模准确率。第一步确定最大值 λ_{max}，此时所有参数 $\beta = 0$，满足这一条件的 λ 非常多，但是会存在一个下确界，该下确界可作为 λ_{max}。第二步确定最小值 λ_{min}，通常取接近 0 的数，或者取 λ_{max} 的很小比例，如 $\lambda_{min} = 0.001\lambda_{max}$。粗略确定取值范围 $[\lambda_{min}, \lambda_{max}]$ 后，接着基于模型选择的思想确定最优 λ。

模型选择中，常用的评价准则有交叉验证(Cross Validation，CV)、广义交叉验证(GCV)、广义信息准则(GIC)、AIC、BIC、风险膨胀准则(RIC)、C_p 准则等。鉴于 CV 的思想简单且非常流行，而且现有整合分析方法(Ma 等，2011a、2011b、2012) 发现其他准则的效果不如它，本文只介绍 k 倍 CV 的基本思想：①构建评价指标，例如预测误差平方和，将样本随机划分为等量 k 份；②$k-1$ 份作为训练集，用于模型建立、模型估计，余下的样本作为测试集，用于检验模型，计算测试集上的评价指标值；③循环第(2) 步，直到所有样本都被作为测试集一次且仅一次；④对于每个 λ，计算它们的预测指标值，该值最小时的 λ 即为最优值。

五、模拟分析

由于异构数据模型更具一般性，在实际应用中更常见，并且同构数据模型在某种角度上可以看作是异构数据模型的特例。由于篇幅限制，本文对多种异构模型进行统计模拟分析。每个模拟都包含 3 个数据集，样本量都是 80，解释变量 1000 个：

模拟 1 中各数据集分别有 5、6、7 个显著变量，模型共有 18 个显著系数。前两个数据集的共同变量有 3 个，后两者之间也是 3 个，而第 1 和第 3 个数据集无共同变量。

模拟 2 中各数据集分别有 7、8、9 个显著变量，3 个数据集共同的显著变量有 5 个，且各有 2、3、4 个特有变量，共 24 个显著系数。

模拟 3 中 3 个数据集各有 10 个显著变量，且两两无重叠。显著系数也是 30 个。以上三种情况包含部分重叠、完全重叠和不重叠。

参考相关文献(Zhang 等，2015；Liu 等，2014)，数据产生方式如下：① X 服从多元正态分布，满足 $\mathrm{cov}(X_i, X_j) = \rho^{|i-j|}$，$\rho$ 为 X_i 与 X_j 的相关系数；② 非 0 系数从 $U[0.5, 1] \cup U[-1, -0.5]$ 中随机产生，误差项的标准差 $\sigma = 0.5$。

模拟方法包含 MCP、L_1 Group MCP、L_1 Group Bridge 以及 Composite MCP，MCP 作为代表性的单变量选择方法用于对比，后三者是典型的双层选择方法，都能用于异构数据整合分析。评价指标有两个：P 表示选择的非 0 系数数目，TP 是正确选择的个数。模拟 100 次的平均结果如表 1 所示。

从模拟结果得出：①随着相关系数增大，4 种方法的 P 指标值越接近 TP，说明假阳数随之减少。② 3 种异构数据方法在各例子中能较准确地识别非零系数，尤其在弱相关和一般相关情形中能接近 100% 识别。③3 种异构数据方法在强相关数据中，明显比 MCP 好；在相关系数为 0.2 或 0.5 时，平均来说 MCP 的假阴数是最高的，其他 3 种方法的假阴数基本为零。④3 种异构数据方法中，L_1 Group Bridge 的效果最好，它在所有模拟结果中假阳数都最低，且接近零，方差也最小，因而最稳定；Composite MCP 的假阳数平均来说比 L_1 Group MCP 要少。

六、应用分析

本文将惩罚整合分析方法应用到两个具有来源差异性的实际数据中，一个是来自不同地区的新农合家庭医疗支出调查数据，可以研究农村医疗支出的地区差异性问题；另一个是具有超高维、小样本等典型大数据特征的癌症基因数据，通过惩罚整合分

表 1　　模拟结果(括号内数据为标准差)

	ρ	MCP		L1 Group MCP		L1 Group Bridge		Composite MCP	
		P	TP	P	TP	P	TP	P	TP
模拟 1	0. 2	21. 91(3. 49)	18(0)	30. 12(13. 08)	18(0)	19. 14(0. 80)	18(0)	29. 94(10. 80)	18(0)
	0. 5	25. 86(5. 45)	17. 93(0. 26)	27. 87(9. 57)	18(0)	18. 62(0. 72)	18(0)	26. 25(8. 00)	18(0)
	0. 8	16. 90(4. 92)	10. 57(1. 13)	19. 00(2. 34)	16. 97(0. 91)	18. 19(1. 10)	17. 05(0. 91)	19. 02(2. 40)	16. 95(0. 92)
模拟 2	0. 2	30. 34(5. 44)	23. 97(0. 30)	44. 14(14. 11)	24(0)	25. 30(0. 95)	24(0)	42. 48(12. 87)	24(0)
	0. 5	28. 17(4. 05)	23. 90(0. 33)	34. 32(7. 64)	24(0)	24. 40(1. 01)	23. 76(0. 45)	34. 14(6. 27)	24(0)
	0. 8	20. 10(4. 70)	12. 47(1. 35)	24. 66(2. 36)	22. 10(1. 53)	21. 79(1. 46)	21. 49(1. 27)	24. 72(2. 26)	22. 09(1. 53)
模拟 3	0. 2	35. 30(5. 93)	29. 98(0. 20)	48. 23(11. 76)	30(0)	31. 09(1. 06)	29. 98(0. 14)	48. 10(11. 98)	30(0)
	0. 5	32. 18(2. 85)	29. 19(1. 46)	38. 02(4. 46)	30(0)	30. 22(0. 80)	29. 79(0. 41)	37. 86(4. 24)	30(0)
	0. 8	22. 31(4. 71)	14. 19(1. 51)	25. 72(2. 47)	24. 15(2. 54)	27. 69(1. 11)	27. 37(1. 05)	25. 68(2. 47)	24. 13(2. 50)

析综合不同研究机构的临床数据,从数万个基因中筛选出对癌症有显著作用的少数基因。

(一) 新农合家庭医疗支出分析

新型农村合作医疗制度(简称新农合) 是政府解决农民基本医疗卫生问题的大规模医疗保障制度。新农会的地区差异性致使医疗支出也存在地域上的区别。本文研究的数据由厦门大学数据挖掘中心于 2012 年 7 月至 9 月的农村入户调查得到,调查范围包括福州、龙岩、三明、南平、漳州 5 个地级市。经数据预处理后得到有效样本 688 份,5 个地区各含 87、58、296、59、188 份。因变量为家庭过去一年的农村家庭实际医疗支出,也就是指医保报销后的家庭实际现金支出。自变量分为 3 类: 一是基本信息,包括家庭人数、65 岁以上人数、户主年龄、户主教育、户主婚姻,共 5 个变量; 二是经济指标,包括家庭总收入、家庭基本支出、家庭储蓄、农业支出、烟酒支出,共 5 个变量; 三是健康相关指标,包含参保人数、健康自评、住院次数、门诊次数等共 8 个变量。其中婚姻、教育、参合因素是多水平分类变量,通过虚拟变量处理后,最终得到 24 个解释变量。由于每个地区对新农合的投入、实施情况不同,而且每个调查地区的经济情况、生活水平、文化观念等也有所不同,并且每个地区的调查是由不同的调查小组完成的,数据集的调查误差也略有不同。如果简单地合并所有数据进行分析,很可能会忽略数据集间的关联性等信息。整合分析能有效分析来自不同地区的数据集,因此本文用异构数据模型分析新农合政策下医疗支出影响因素的地区差异。

由于模拟分析中 L_1 Group Bridge 综合表现最好,本文建立该方法下关于医疗支出的异构数据模型,估计结果如表 2 所示。可看出: ①5 个地区对家庭医疗支出的影响因素都是不一样的,这也进一步验证了如果简单地合并所有数据集再进行分析,很容易忽略了地区间的差异性和关联性信息。②5 个数据集共有 15 个显著变量,其中“住院次数”为共同显著变量,且在 5 个地区中对医疗支出都成正向影响,即住院次数越多,医疗支出越高。③“住院意愿是否改变”在南平外的 4 个地区都是显著的,且在其中 3 个地区成正向影响,即选择更好的医院治疗。④“慢性病人数”在福州、三明、龙岩都是正向影响,家庭的慢性病人数越多,医疗支出越高。⑤“门诊次数”、“医院收费合理性”在两个地区显著,且门诊次数越多,支出就越高,而医院收费是否合理对医疗支出的影响方向在不同地区是不同的。⑥ 4 个经济指标显著且呈正向影响,其中收入、农业支出仅在三明市显著,基本支出和储蓄在龙岩市显著。以上结论比较符合现实意义,也与已有的研究成果(Mcbride,2005; Fang 等,2012) 在不同程度上吻合。

表 2　　医疗支出数据的估计结果

地区	显著变量	系数值	地区	显著变量	系数值
福州	健康自评	-0. 065	龙岩	基本支出	0. 145
	慢性病人数	0. 191		储蓄	0. 435
	住院次数	0. 477		慢性病人数	0. 259
	住院意愿是否改变	-0. 028		住院次数	0. 414
	门诊次数	0. 164		住院意愿是否改变	0. 015
三明	教育(大学)	-0. 050	南平	参合人数	-0. 044
	收入	0. 083		参合因素(老人)	0. 423
	慢性病人数	0. 172		参合因素(成年人)	0. 348
	住院次数	0. 493		住院次数	0. 206
	住院意愿是否改变	0. 067	漳州	住院次数	0. 237
	婚姻(离婚)	0. 042		住院意愿是否改变	0. 057
	农业支出	0. 118		门诊次数	0. 018
	医院收费合理性	0. 009		医院收费合理性	-0. 011

尽管上述模型估计的结果较为合理,但为了更进一步验证异构模型在本实证分析中的有效性,本文从预测角度将其与传统模型进行比较。包含三个

模型: L_1 Group Bridge 惩罚异构数据模型,合并 5 个数据集建立 MCP 惩罚线性模型,5 个数据集分别建立 MCP 惩罚模型。后两者代表单数据集模型,之所以选择 MCP 惩罚,是因为该惩罚在单变量选择中综合效果最好。5 个数据集都按 3∶1 随机划分为训练集和测试集,基于训练集建立模型,测试集上构建预测指标 $MSE = \frac{(y_{test} - \hat{y}_{test})^T (y_{test} - \hat{y}_{test})}{n_{test}}$,并分地区计算了预测指标值。运算 100 次的平均结果如表 3 所示,可得出不论在总体还是各地区中,L_1 Group Bridge 异构模型的预测效果都比分开的 MCP 模型好。再与合并的 MCP 模型进行对比,除三明市外,其他地区 L_1 Group Bridge 异构模型的预测指标值都要低。整体来看,异构模型的预测效果显然比两个单数据模型好。两个 MCP 模型进行比较时,合并数据集时效果更好,这可能是样本量较高的缘故。

表 3 预测结果(括号内数值为标准差)

数据	L_1 Group Bridge	合并的 MCP 模型	分开的 MCP 模型
总体	0.733(0.070)	0.742(0.071)	0.786(0.067)
福州	0.398(0.105)	0.626(0.089)	0.529(0.141)
龙岩	0.925(0.232)	0.739(0.121)	1.026(0.215)
南平	0.639(0.110)	0.647(0.123)	0.672(0.115)
三明	0.942(0.345)	0.819(0.282)	0.972(0.273)
漳州	0.912(0.077)	0.922(0.105)	0.948(0.068)

(二) 肺癌基因筛选分析

自 1985 年起,肺癌已成为全球最常见的恶性肿瘤之一,肺癌的死亡率排在我国恶性肿瘤的第一位。基因分析在肺癌诊断研究中广泛使用,通过搜寻与症状相关的基因以辅助临床治疗和诊断。基因数据存在典型的高维性,基因数目常常成千上万,同时数据获取的途径特殊、成本高且不具再现性,故存在高维小样本的特点。基于传统的单数据集的分析结果往往不尽人意(Liu 和 Ma,2014),需要整合不同医院或者地区的数据以增大样本量,但由于不同来源的数据具有异质性,又不能简单地合并,因此整合分析方法在此具有显著的优势,此外,在其他癌症的诊断中也十分常用(Liu 和 Ma,2014; Shi 等,2013; Liu 等,2013)。

肺癌基因数据共有 3 个独立的数据集,来自 3 个不同的研究机构,解释变量(被测基因)共 22283 个,总的有效样本数 336,3 个数据集的有效样本数分别为 175、79、82; 其中在研究过程中死亡的样本数分别为 102、60、35,共计 197 个。显然高维性、小样本、来源差异性特征都很明显,故非常适合用整合分析来筛选变量。同时,由于数据来自 3 个不同的、相互独立的研究,数据集间的异质性不能忽略,因此基于两种异构数据整合分析方法 L_1 Group MCP 和 Composite MCP 展开分析,以 AFT(accelerated failure time)模型为基础。

基因选择和参数估计结果限于篇幅不再给出,从结果可得出: ①L_1 GroupMCP 从 22283 个基因中筛选出 25 个显著基因作为解释变量,只有两个基因(SOD1、PTMA)出现在两个数据集中,其他 23 个都仅出现在一个数据集中。②CompositeMCP 筛选出 16 个基因,且不同数据集中不存在交叉基因,该特点与已有研究一致,在 Liu 和 Ma(2014)中,该方法筛选出的 5 个基因在不同数据集中也不存在交叉。③Composite MCP 筛选出的所有基因都被 L_1 Group MCP 识别出来,且每个基因在两种方法下系数估计值的符号一致,甚至估计值相等,或者数量级相同。④从两种方法的分析结果发现,不同数据集具有不同的显著基因,这在一定程度上解释了已有研究中不同数据集下鉴别的基因无法统一的原因。

由于临床数据的不可重现性,要对上述基因选择的准确度进行验证是很难的。因此,本文采用交叉验证(Cross-Validation)的预测评价方式来验证(Huang 和 Ma,2010; Ma 等,2009)。数据按 3∶1 随机分为训练集和测试集,基于 log-rank 统计量对预测结果进行考察。根据重复 100 次的预测结果取中位数,得到 L_1 Group MCP 的 log-rank 统计量为 4.77,Composite MCP 为 3.70。且 L_1 Group MCP 能显著地将因变量的两类分割开来(p 值为 0.029)。根据该预测结果,我们认为对于肺癌基因数据,L_1 Group MCP 的分析结果更理想。

七、总结

大数据往往由来源不同的数据集构成,且呈现出高维性和稀疏性的特点。如何建立合适的统计方法挖掘数据集间的关联性与差异性并实现降维去噪是大数据时代统计建模面临的重大挑战之一。基于惩罚方法的整合分析将变量选择思想与整合分析相结合,能同时分析多个数据集,利用原始数据信息分析数据集间的异同,避免数据来源差异引起的建模不稳健问题,是实现大数据分析目标的有效方法。整合分析将同一解释变量在所有数据集中的回归系数视为一个组,惩罚函数对系数组进行压缩。与单

个数据集的群组惩罚不同之处在于,整合分析的组由同一变量的所有系数构成,而后者的组是多个解释变量的系数。

根据同一解释变量在不同数据集中的显著性是否相同,数据可分为同构型和异构型,本文分别探讨了这两类数据的整合分析方法,从惩罚函数的内部结构来总结它们的原理与特点。同构数据整合分析引入组变量选择方法,确保同一变量在所有数据集中具有相同的显著性。异构数据整合分析更为一般化,解释变量在不同数据集中可有不同的显著性,利用双层变量选择方法来实现变量筛选。此外,考虑到实际应用中数据集之间以及变量之间往往普遍存在复杂的网络结构,本文对同构数据和异构数据中网络结构的惩罚方式分别进行了梳理。统计模拟分析了 3 个异构数据的整合分析,考虑了弱相关、一般相关和强相关情形,发现各种整合分析方法都能较好地识别非 0 系数,但都存在假阳性,而 L_1 Group Bridge 的假阳数最低、方差最小。最后,将整合分析方法应用到两类具有来源差异性的代表数据。首先,利用 L_1 Group Bridge 异构数据模型分析了新农合政策下家庭医疗支出影响因素的地区差异性,模型估计结果比较符合现实意义,不论在总体还是各地区中 L_1 Group Bridge 的预测效果都比单数据集 MCP 要好。然后,将 L_1 GroupMCP 和 Composite MCP 构建异构 AFT 模型分析了具有超高维、小样本等特征的癌症基因数据,两种方法筛选出的基因不完全一致,预测结果显示 L_1 Group MCP 的分析结果更为理想。综上所述,整合分析方法在分析具有来源差异性、高维的数据集时具有很好的效果。

参考文献

[1] Fan J, Han F, Liu H. Challenges of Big Data analysis [J]. National Science Review, 2014, 1(2):293 – 314.

[2] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables [J]. Journal of the Royal Statistical Society: Series B, 2006, 68: 49 – 67.

[3] Simon N, Friedman J, Hastie T and Tibshirani R. A sparse Group lasso [J]. Journal of Computational and Graphical Statistics, 2013, 22(2):231 – 245.

[4] Huang J, Ma S, Xie H and Zhang C.-H. A group bridge approach for variable selection [J]. Biometrika, 2009, 96:339 – 355.

[5] Ma S, Huang J, Song X. Integrative analysis and variable selection with multiple high-dimensional data sets [J]. Biostatistics, 2011a, 12(4): 763 – 775.

[6] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties [J]. Journal of the American Statistical Association, 2001, 96:1348 – 1360.

[7] Ma S, Dai Y, Huang J and Xie Y. Identification of breast cancer prognosis markers via integrative analysis [J]. Computational statistics and data analysis, 2012, 56(9): 2718 – 2728.

[8] Huang J, Wei F, Ma S. Consistent group selection and estimation via normed minimax concave penalty, 2010. Unpublished manuscript.

[9] Huang J, Breheny P, Ma S. A selective review of group selection in high-dimensional models [J]. Statistical Science, 2012, 27(4): 481 – 499.

[10] Ma S, Huang J, Wei F, et al. Integrative analysis of multiple cancer prognosis studies with gene expression measurements [J]. Statistics in medicine, 2011b, 30(28): 3361 – 3371.

[11] Liu J, Ma S, Huang J. Integrative analysis of cancer diagnosis studies with composite penalization [J]. Scandinavian Journal of Statistics, 2014, 41(1): 87 – 103.

[12] Zhang Q, Zhang S, Liu J, et al. Penalized integrative analysis under the accelerated failure time model [J]. arXiv preprint arXiv: 1501. 02458, 2015.

[13] Shi X, Liu J, Huang J, et al. Integrative Analysis of High - throughput Cancer Studies With Contrasted Penalization [J]. Genetic epidemiology, 2014, 38(2): 144 – 151.

[14] Fang K, Wang X, Zhang S, et al. Bi-level variable selection via adaptive sparse group Lasso [J]. Journal of Statistical Computation and Simulation, 2014. DOI: 10. 1080/00949655. 2014. 938241.

[15] Liu J, Huang J, Ma S. Incorporating Network Structure in Integrative Analysis of Cancer Prognosis Data [J]. Genetic Epidemiology, 2013, 37(2):173 – 183.

[16] Zhao Q, Shi X, Huang J, et al. Integrative Analysis of "-Omics" Data Using Penalty Functions [J]. Wiley Interdisciplinary Reviews Computational Statistics, 2015, 7: 99 – 108.

作者简介

马双鸽,男,1978 年生,内蒙古呼伦贝尔人,2004 年 7 月获美国威斯康辛大学统计博士学位,现任美国耶鲁大学生物统计系副教授,厦门大学经济学院讲座教授,厦门大学数据挖掘研究中心副主任。研究方向为数理统计、数据挖掘、生物统计。

王小燕,女,1987 年生,湖南娄底人,2015 年毕业于厦门大学经济学院统计系,获经济学博士学位,现为湖南大学金融与统计学院助理教授。研究方向为数据挖掘、计量经济学。

方匡南,男,1983 年生,浙江台州人,2010 年毕业于厦门大学统计系,获经济学博士学位,现为厦门大学经济学院教授、博士生导师。研究方向为数据挖掘、计量经济学。

(责任编辑: 方　原)