*Communication on Contemporary Anthropology*, 2012, 6,26-29/e4
doi:10.4236/coca.2012.61004 Published Online January 17, 2012

REVIEW

# Variable selection and dimensional reduction in genetic analysis

**Kelin Xu**

*MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China*

Email: xukelin0202@gmail.com

Received: Dec.7, 2011; Revised: Dec. 11, 2011; Accepted: Jan. 17, 2012.

**Abstract:** In the context of the development of next generation sequencing, a large amount of genetic data is accumulated. We usually focus on the extraction of meaningful information from these data, considering the problem of computation cost and the challenges in traditional statistical method.Consequently, variable selection and dimension reduction play important roles in genomics. Here, I reviewed some popular methods referring to that, classifying them to three main types along with their properties and application fields.

**Keywords:**genetic data, high dimension hazard, variable selection, dimension reduction

现代人类学通讯 2012 年第六卷第 26-29 页**专题综述**

# 遗传数据中的变量选择与降维

徐珂琳

复旦大学生命科学学院现代人类学教育部重点实验室, 上海 200433

**摘要**：随着新一代测序技术与芯片杂交技术的发展，海量高维数据涌现在研究者们的面前，如何从这些高维数据中提取有效信息成为摆在人们面前的一大难题。在这种高维问题的背景下，许多基于低纬度的统计结论不再成立；另外，庞大的数据量对计算速度提出了很高的要求。于是，在这种数据驱动的研究背景下，变量选择与降维成为主要的研究方向。本文从当下遗传数据的特点出发，回顾了当今几种主流的变量选择与降维方法，如主成分分析、偏最小二乘回归、切片逆回归、LASSO 等，并就这几种方法的性质与适用范围展开讨论。

**关键词**：遗传数据；高维问题；变量选择；降维

随着基因组测序技术的发展，特别是以 Roche 公司的 454 技术、Illumina 公司的 Solexa 技术和 ABI 公司的 SOLiD 技术为标志的第二代测序技术的产生，使得测序通量得到显著提升。与第一代测序技术相比，第二代测序技术测序在保证高准确率的同时提高了测序速度，降低了测序成本，使得研究者可以以较低的成本更全面深入的对人类基因组进行研究[1]，催生出了海量的测序数据。

然而，从这些测序得出的海量信息中选出真正与性状相关的位点是非常困难的。一些研究人员倾向于选择编码区的变异位点进行研究，即采用外显子捕获技术，但是这种技术从产生之初就具有着不可忽视的局限：只能获得获得外显子区域内部及边界的变异信息，对大部分发生在非编码区的遗传变异束手无策，并且不能检测到基因组内较大结构性的变异及基因拷贝数变化[2]。由此，我们只能返回以数据为导向，大规模、工业化的研究模式，以求在数据处理方法上得到突破。

另外，在基因表达芯片数据中也存在着类似的问题。随着芯片杂交技术的不断发展成熟，用十几个芯片得到上千个基因的表达数据成为现实。但是这种数据中存在着大量的未知相关关系结构的信息[3]，又由于变量的维数 p 远大于样本的个数 n，传统的统计方法此时不再适用。

在这些高维数据中，不同基因间的相关关系所造成的复共线性也将对模型结果造成很大影响[4]，且会使得系数的估计值远大于真实值[5]。当所选基因之间存在相关性，如

线性相关时，若采用传统的线性回归模型，设此时的表达量矩阵为 X，则 X'X 至少有一个特征值很小，由

$$E(\hat{\beta}^2) = \beta^2 + \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}$$

可知此时系数估计值过大，需要对$\beta^2$做惩罚。

另外，无关基因的选入也将会使模型的噪声增大，且使检验的功效降低[6]。Fan 在其 1996 年的文章中用数值模拟的方法得出：当模型的维度非常高时，变量之间存在着许多假的相关关系，也就是说，即使使用计算随机模拟出的线性无关变量间计算它们之间的相关系数，当这些变量的数量非常大时，它们之间的相关系数也会显著的偏离零——这个结果显然违背了人们对线性无关变量性质的预期。而针对实际数据，他得出，当不加选择的加入所有变量时，模型的结果与随机猜想的结果是没有差别的[6]。由此，在不损失疾病相关基因的前提下选择合适的变量、降低模型的维度是非常重要的。

现今，降低模型维度的方法主要分为以下几种：1、按一定的惩罚规则筛选出目的基因，利用传统的统计方法进行分析，如 t 检验法与秩和检验[7-8]；2、对多个基因的信息进行组合后，利用约束统计量等进一步分析，如主成分分析、因子分析、偏最小二乘与切片逆回归[9-13]；3、基于稀疏性假设的降维预测方法，如 LASSO (Least Absolute Shrinkage and Selection Operator) [14-15]等。

## 筛选法

在利用高维基因表达芯片数据进行复杂疾病关联分析研究时，由于疾病相关位点的复杂性，传统的关联分析方法此时不再有效。另外，在测序数据处理中，由于位点数量庞大且相关关系纷繁复杂，逐个位点进行相关性筛选显然是无效的[16]。由此，我们考虑最简单直接的方法——筛选出与疾病密切相关的基因表达数据集，从而降低变量的维数，之后再用传统的统计学方法进行关联分析。这种筛选方法的一般思路是按照一定的顺序在模型中添加删除变量，直到约束统计量达到最优值。在这种方法中，常用的约束统计量有 Akaike 在 1973 年提出的 AIC 准则[17]，即最小化：

$$-\ell_n(\hat{\theta}) + \sum_{j=1}^{p} I(\hat{\theta}_j \neq 0)$$

其中，$\ell_n(\hat{\theta})$表示对数似然函数。这种原则保证了在缩小变量集合的前提下最大化模拟参数出现的概率。类似的约束统计量还有 Schwartz 在 1978 年提出的 BIC 准则[18]等。

另外一种筛选基因的方法是单纯排序方法：通过对基因与疾病的相关系数等统计量的排序，选出与疾病最相关的几个基因，然后建立回归模型。此外还可以考虑逐步计算模型的功效函数值，取此值达到最优时的基因数据集。这种基于变量选择筛选方法的思路比较直观简单，但由于会重复建立模型，计算量很大，当数据维数过高时便不太适应；另外，这种筛选方法基于检验统计量的值随模型维度改变能取到最优值的假设，且这种最优值是向前或向后回归中的局部最优，故而在许多情形下并不能有效降维。

## 信息组合方法

信息组合方法是通过对变量(测序位点、基因等)计算加权线性值建立新的变量代替原始变量，从而达到降维的目的。在多元统计中，最经典的组合降维方法是主成分分析方法(PCA)。在基因表达芯片分析中，主成分分析方法利用各基因表达数据的线性组合信息，根据累计贡献率等统计量的大小，选取所含信息最多的多基因组合变量[19]。另一种组合降维方法——因子分析的基本出发点与主成分分析有异曲同工之处，相较于主成分分析的由原始基因表达数据出发计算出几个有代表性的线性组合量，因子分析从假设存在的因子出发，根据需要的性质计算出这些因子的表示形式[20]，于是，在方法与性质方面，因子分析比主成分分析有更大的灵活性。

然而，无论是主成分分析还是因子分析都存在着一个致命的弱点，即主成分或因子在实际问题中代表的含义很难解释。另外，在主成分的选取方法中，并没有考虑综合变

量与疾病间的相关关系，由这种方法选出的变量可能与所研究的疾病性质并没有很大的相关性。

Xiong 在 2002 年提出了一种基于 Hotelling's $T^2$ 检验统计量的综合多基因位点信息的统计方法[21]，这种方法在综合了多个基因位点的表达信息的同时考虑了它们之间的互作，但这种互作只是局限在一个单倍型中，对于更高程度的相互作用这种方法是无能为力的。另外，由于 Hotelling's $T^2$ 检验统计量的自由度等于标记的个数，当同一区域能的标记间的存在着很强的连锁关系时，就会造成该统计量中一些自由度的浪费，导致检验功效的降低[16]。

另外，在这种综合多位点信息处理高维芯片数据的方法中，偏最小二乘是最近比较常用的一种。这种方法通过计算芯片表达数据与目标函数之间的协方差矩阵，充分利用了变量的先验信息，可以看做是一种先验的 PCA 方法，而且相较于 PCA 方法有更好的功效[22]。此外，这种计算方法的效率很高，灵活性很强，有很多不同的算法可以选择[11]，但是当芯片数据维数过高，远大于 n 时，这种方法的效力是有限的，可以考虑运用稀疏性方法与偏最小二乘的组合[23]。

另外一种比较常用的、基于目标因素考虑的组合信息降维方法是切片逆回归方法[24-26]。与偏最小二乘相比，这种方法不受线性模型的限制，即在表达数据与研究目标之间存在非线性关系时仍然适用[27]，但也存在着它本身的弊端：当表达数据本身存在着很强的相关性时效果很差[28]，这时我们可以考虑采用切片逆回归与偏最小二乘的一种结合方法——部分逆回归[29]。此外，切片逆回归算法非常灵活，还可以根据实际需要与变量选择方法或稀疏性假设方法综合使用[30-31]。

## 基于稀疏性假设的几种方法

高维基因表达数据的一个主要特点就是所要研究的基因数远大于样本的个数，但是，真正的疾病相关基因往往是比较少的，这样，我们可以考虑对这些数据运用稀疏性假设，即认为大部分基因与疾病无关，它们的回归系数为零。

在稀疏性假设的前提下，最直接的方法是对非零系数的个数设定约束后运用传统的线性回归建立模型。但是这种基于 $L_0$ 范数的方法在具体算法上是 NP 难问题，计算代价太高，于是基于 $L_1$ 范数的 LASSO 方法应运而生。由于 LASSO 方法在将无关变量的系数归零方面的卓越性质，统计学家们对其的发展热情可谓经久不衰[15, 32-33]。此外，统计学家还在对惩罚函数 $p_\lambda(\beta)$ 的选择上进行了一系列的尝试：

Fan 和 Li 在 2001 年总结出惩罚函数选择的三个条件[34]：

1、稀疏性条件模型将小的估计参数归零；

2、无偏性条件模型的估计参数是无偏的；此条件特别强调在估计参数比较大时成立；

3、连续性条件模型的估计参数是连续的，即在芯片数据有微小变动时，模型结果不会有很大改变。

针对以上三个条件，他们推导出了惩罚方程在各条件下需满足的性质，其中，针对范数惩罚方程 Lq，稀疏性条件要求 q 值大于零小于 1，而连续性条件要求 q 值大于 1，这显然是矛盾的，于是他们提出了 SCAD (Smoothly Clipped Absolute Deviation)方法，惩罚方程的导函数可化为如下形式：

$$p'_\lambda(t) = \lambda\left\{I(t \le \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda)\right\}$$

for some a>2

具有这种导函数的惩罚方程所建立的模型是满足 Fan 与 Li 所提出的三个条件的。类似的，Zhang 在 2009 年推导出满足：

$$p'_\lambda(t) = \frac{(a\lambda - t)_+}{a}$$

的惩罚方程也是满足如上三条性质的，由此建立了 MCP (Minimax Concave Penalty)方法[35]。

另外，对于超高维数据，Fan 与 Lv 提出了一种 ISIS (Iterative Sure Independence Screening)数据筛选方法，即在独立性假设下，交替进行粗筛与精筛，从而达到降低模型维度的目的。当将此方法适用于分类问题时，还衍生出了 FAIR (Features Annealed Independence)方法。

## 讨论与展望

在当下这种实验技术逐步成熟，各种数据海量涌现的大背景下，以数据为导向的研究模式成为一种趋势。针对这些海量高维数据，各种统计方法也不断发展起来。针对高维数据处理方法中的问题，我们有两种主要研究方向，即变量筛选与降维(在最少的损失所得变量信息的前提下)，而稀疏性假设下的模型建立可以看做是这两种方向的综合。然而，除稀疏性假设下的 ISIS 方法外，其余方法都局限在普通的高维问题($p = Cn^{\alpha}$)中,而对现在所面临的超高维问题有些“鞭长莫及”，我想，如何更好的处理这些问题将成为以后数据处理方法研究中的热点。

**参考文献：**


1.Quackenbush J(2001)Computational analysis of microarray data. Nat Rev Genet 2: 418-427.

2.杨旭,焦睿,杨琳,吴莉萍,李英睿,王俊(2011)基于新一代高通量技术的人类疾病组学研究策略. 遗传 33: 829-846.

3.Wang H, van der Laan MJ(2011)Dimension reduction with gene expression data using targeted variable importance measurement. BMC Bioinformatics 12:312.

4.Fan JQ, Lv JC(2008)Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Series B Stat Methodol 70: 849-883.

5.王松桂,陈敏,陈丽萍(1999)线性统计模型. 北京: 高等教育出版社. 61.

6.Fan JQ(1996)Test of significance based on wavelet thresholding and Neyman's truncation. J Am Stat Asso 91: 674-688.

7.Hedenfalk I, Duggan D, Chen Y(2002)Gene-expression profiles in hereditary breast cancer. Advances in Anatomic Pathology 9(1): 1-4.

8.Dettling M, Buhlmann P(2003)Boosting for tumor classification with gene expression data. Bioinformatics 19: 1061-1069.

9.Ghosh D(2002)Singular value decomposition regression modeling for classification of tumors from microarray experiments. Pac Symp Biocomputi 2002: 18-29.

10.Meng J(2011)Uncover cooperative gene regulations by microRNAs and transcription factors in glioblastoma using a nonnegative hybrid factor model.In International Conference on Acoustics, Speech and Signal Processing.

11.Nguyen DV(2005)Partial least squares dimension reduction for microarray gene expression data with a censored response. Math Biosci 193:119-137.

12.Chun H,Keles S(2009)Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. Genetics 182: 79-90.

13.Antoniadis A, Lambert-Lacroix S, Leblanc F(2003)Effective dimension reduction methods for tumor classification using gene expression data. Bioinformatics 19: 563-570.

14.Tibshirani R(1996)Regression shrinkage and selection via the Lasso. J R Stat Soc Series B Stat Methodol 58(1): 267-288.

15.Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K(2005)Sparsity and smoothness via the fused lasso. J R Stat Soc Series B Stat Methodol 67: 91-108.

16.沈炎峰(2010)多变量数据遗传分析方法的研究. 浙江大学博士论文.

17.Akaike H(1973)Information theory and an extension of the maximum likelihood principle. In Second International Symposium on Information Theory 267-281.

18.Schwartz G(1978)Estimating the dimension of a model. Ann Statist 6: 461-464.

19.于秀林,任雪松(2007)多元统计分析. 北京:中国统计出版社.

20.何晓群(1998)现代统计分析方法与应用. 北京:中国人民大学出版社.

21.Xiong MM, Zhao JY, Boerwinkle E(2002)Generalized T-2 test for genome association studies. Am J HumGenet 70: 1257-1268.

22.Rocke DM, Nguyen DV(2002)Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 18: 39-50.

23.Chun H, Keles S(2010)Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J R Stat Soc Series B Stat Methodol 72: 3-25.

24.Li KC(1991)Sliced inverse regression for dimmension reduction. Journal of the American Statistical Association 86: 316-327.

25.Li KC(1992)On principal hessian directions for data visualization and dimension reduction-another application of steins lemma. J Am Stat Asso 87:1025-1039.

26.Cook RD(2000)SAVE: A method for dimension reduction and graphics in regression. Commun Stat Theory Methods 29: 2109-2121.

27.杨乐(2010)现代数学基础丛书. 北京:科学出版社.

28.Naik P, Tsai CL(2000)Partial least squares estimator for single-index models. J R Stat Soc Series B Stat Methodol 62: 763-771.

29.Li LX,Cook D, Tsai CL(2007)Partial inverse regression. Biometrika 94: 615-625.

30.Li LX, Cook RD, Nachtsheim CJ(2005)Model-free variable selection. J R Stat Soc Series B Stat Methodol 67: 285-299.

31.Wang Q, Yin XR(2008)A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE. Computational Statistics & Data Analysis 52: 4512-4520.

32.Lu Y, Zhou Y, Qu W, Deng M, Zhang C(2011)A Lasso regression model for the construction of microRNA-target regulatory networks. Bioinformatics 27: 2406-2413.

33.Tibshirani RJ, Taylor J(2011)The Solution Path of the Generalized Lasso. Ann Statist 39: 1335-1371.

34.Fan JQ, Li RZ(2001)Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Asso 96: 1348-1360.

35.Zhang CH(2010)Penalized linear unbiased selection. Ann Statist38: 894- 942.