Communication on Contemporary Anthropology, 2012, 6,39-44/e7
doi:10.4236/coca.2012.61007 Published Online January 17, 2012

REVIEW

Application of Partial Least Squares in high-dimensional genomic data analysis

Panpan Wang

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

Email: catherine64278@163.com

Received: Dec.7, 2011; Revised: Dec. 14, 2011; Accepted: Jan. 17, 2012.

Abstract: Partial Least Squares (PLS) is a statistical regression technology which could perform well on the analysis of high-dimensional genomic data, such as the microarray data, SNP data from GWAS, and proteomic data. In this article, we review the challenges that are faced by the classical linear regression, and lead to the advantages of PLS. PLS can not only solve the problem of co-linearity through dimension reduction but also the problem of regression singularity in the condition of small sample size and high dimensional predictive variables. We also provide some modified algorithms of PLS incorporate with the application in the real biological data analysis. For example, sparse partial least squares can simultaneously realize dimension reduction and variable selection, and the combination of PLS with cluster analysis or general linear regression can deal with diverse problems of data analysis.

Key words: partial least squares, high-dimensional genomic data, dimension reduction, variable selection

现代人类学通讯 2012 年第六卷第 39-44 页**专题综述**

偏最小二乘在高维基因组数据分析中的应用

王盼盼

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：偏最小二乘是一个非常高效的统计回归技术，它能很好的应用于高维的基因组数据的分析中，如基因表达的芯片数据，全基因组关联分析的 SNP 数据，甚至蛋白质组数据等。在本文中，我们将从最初的线性回归讲起，引出偏最小二乘回归在高维数据分析中的优势。它不仅能通过降维解决预测变量的共线性问题，也能解决样本数目偏少的回归奇异性问题。并结合偏最小二乘在实际生物数据中的应用，给出修正的算法。如稀疏的偏最小二乘方法能在降维的同时实现变量选择，偏最小二乘与聚类分析或广义线性回归结合能更多的应用于各种不同的数据分析问题。

关键词：偏最小二乘；高维基因组数据；降维；变量选择

线性回归是应用最为普遍的统计回归方法，其中回归系数的求解是使用最小二乘方法，而且回归方程及系数显著性检验也有相应的 F 分布检验和 t 分布检验。但是最为一种理论完善的回归方法，线性回归并不能直接应用于现在的很多生物技术所产生的数据，如基因组芯片数据，高通量序列数据，GWAS 数据等。这些数据的特点是维度很大，表达芯片数据一般包括成千上外的基因(维度 p，基因)，GWAS 数据则包括更多的 SNP 数据，同时样本量又较小。

面对这种高维数据，降维与变量选择是非常重要的一个数据处理环节[1]。而且，在对高维的变量进行回归分析时，预测变量之间经常是相关的，比如基因之间有相互作用，SNP 位点之间有连锁不平衡。这种变量之间的相关性就称为共线性。主成分分析(Principle Component Analysis)和偏最小二乘方法(Partial Least Squares)是两种重要的对回归降维的方法[1]，能通过降维解决变量的共

线性问题。相比与主成分分析，偏最小二乘是一种有监督的降维方法，从而能更高效地实现降维。而且，偏最小二乘具有通用性，它能够很容易的应用于多种任务，如分类，生存期分析，遗传网络分析[2]等。

本文主要介绍偏最小二乘的原理算法，并结合它在生物数据分析中的应用给出更多的基于偏最小二乘的方法。

1. 偏最小二乘回归

偏最小二乘(Partial Least Squares，PLS)是 Herman Wold 在 20 世纪六七十年代提出来的，它是基于连续使用最小二乘来进行通路模型的建模。较近的两篇文章回顾了 PLS 从 Herman Wold 的通路模型以来的发展过程[3, 4]。以前大部分关于 PLS 的文章都是关于化学计量学的，随着现代生物技术产生的高维数据，相信 PLS 也会在生物数据分析领域起到重要作用。尽管很多统计学家从统计学的观点研究了 PLS 回归，但是一般来说，PLS 被认为是基于数据的分析方法，因为它缺少一套完善的概率模型[5]作为统计理论。为了方便理解 PLS 作为一个回归降维的方法，我们试图把它与最基本的最小二乘线性回归和主成分分析联系起来。

线性回归模型

$$\mathrm{Y} = \mathrm{X}\beta + \mathrm{E}$$

$X_{n\times p}$:预测变量矩阵，也称设计矩阵，在高维数据的情况下，p 很大；

$Y_{n\times q}$: 响应变量矩阵，q 是相应变量的维数，大多数情况 q=1,Y 是向量；

$E_{n\times q}$: 误差矩阵；

在正常情况下，设计矩阵 X 满足 rk(X)=p,此时观测数目$\mathrm{n} \geq \mathrm{p}$, 且各预测变量$X_i$，$\mathrm{i} = 1, \cdots, \mathrm{p}$互相独立，没有相关性。此时系数的 LS 估计为：

$$\hat{\beta} = (X'\mathrm{X})^{-1}X'Y$$

由上面可以看出，线性回归的 LS 估计有很强的要求条件，当预测变量的维数很大，且各项之间有相关性时，主成分分析是很重要的一种降维解决共线性的方法[1]。

主成分分析

σ_{ij}: 预测变量X_i与X_j之间的协方差；

$\sum$：协方差矩阵$((\sigma_{ij}))$；

$\mathrm{rank}(X) = r \leq p$

在主成分分析(PCA)中，就是去寻找原始预测变量X_i，$\mathrm{i} = 1, \cdots, \mathrm{p}$的线性组合$\sum \alpha_i X_i$来作为新的预测变量。

由于协方差矩阵$\sum$ 是对称矩阵，它的 Jordan 分解可以写为：

$$\Gamma^T \Sigma \Gamma = \Lambda = diag\left(\lambda_1, \lambda_2, \cdots, \lambda_p\right)$$

其中λ_i是第 i 个特征向量，并且满足$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$。这样，通过变换

$$\mathrm{z}_{1\times p} = (x - \mu)_{1\times p}\Gamma_{p\times p}$$

原始预测变量$x_{1\times p}$中心化之后，经过线性变换，得到新的预测变量$\mathrm{z}_{1\times p}$。这时$z_{1\times p} = \left(z_1, z_2, \cdots, z_p\right)$中，$z_i$是对应于特征值$\lambda_i$的主成分，一般情况下，只取前几个较大的特征值对应的主成分来进行接下来的线性回归。主成分之间相互独立，且使用较少的新变量就可以去的较好的回归估计。

也就是说，主成分分析通过寻找原始变量的线性组合找到对应的主成分，不仅能用很少的变量捕捉到原来 p 个变量的大部分信息，而且解决的变量之间的共线性问题。

但是，协方差阵$\sum$ 只是包含了预测变量 X 的信息，没有考虑预测变量与响应变量之间的相关性，因而主成分分析是一种无监督的降维方法。如果我们考虑预测变量与相应变量之间的协方差阵，就引出了 PLS 方法。

PLS 模型

PLS 回归建立在假设 X 和 Y 矩阵都有潜变量的分解[2]如下：

$$\mathrm{X} = \mathrm{T}P^T + E$$

$$\mathrm{Y} = \mathrm{T}Q^T + \mathrm{F}$$

$\mathrm{T} \in R^{n\times c}$ 是对 n 个观测值的潜变量矩阵，c 是潜变量的个数；

$\mathrm{P} \in R^{p\times c}$和$\mathrm{Q} \in R^{q\times c}$是系数矩阵；

$\mathrm{E} \in R^{n\times p}$和$\mathrm{F} \in R^{n\times q}$是随机误差阵；

如果有矩阵$\mathrm{T}, \mathrm{P}, \mathrm{Q}$，满足上面的分解式，那么对任何一个可逆矩阵$\mathrm{M}$,$\tilde{T} = TM, \tilde{P} = P(M^{-1})^T$和$\tilde{Q} = Q(M^{-1})^T$也满足分解式，也就是说由T的列扩张成的空间要比矩阵T本身更

重要。而且，像主成分分析一样，潜变量 T 也可以写成原始预测变量 X 的线性变换：

$$X = TW$$

矩阵W是权重系数矩阵，而W 的列 $W = (w_1, \cdots, w_k)$可以通过优化问题来得到。

就像我们前面提到的，PLS 试图寻找潜变量$T = (T_1, T_2, \cdots, T_c)$,并且$T_i$能最大的捕捉预测变量与响应变量之间的协方差信息。

假设 X 和 Y 都是已经中心化的数据，并在此考虑单变量的响应变量 Y，那么

$$\widehat{\mathrm{Cov}}(Y, T_i) = \frac{1}{n} T_i{}^T Y = \frac{1}{n} w_i{}^T X^T Y$$

目标函数：单变量的响应变量(PLS1)

$$\text{For i} = 1, \cdots, \text{c}$$

$$w_i = \underset{w}{\operatorname{argmax}}\, w^T X^T Y Y^T X w \quad (1)$$

$$\text{subject to } w_i{}^T w_i = 1$$
$$and\ t_i{}^T t_j = w_i{}^T X^T X w_j = 0,$$
$$for\ j = 1, \cdots, i-1$$

其中 c 是潜变量的个数，一般由用户指定或者使用正交验证得到。权重向量 w 可以通过 Martens 和 Naes[6]提出的一个简单又快速的“algorithm with orthogonal scores”来实现，这个算法可以得到正交的潜变量t_i,因此得到广泛的应用。

对于多元响应变量的情况，主要有两种不同的约束下的优化问题[2]。下面我们给出两种表达式。

目标函数：(PLS2)

$$w_i = \underset{w}{\operatorname{argmax}}\, w^T X^T Y Y^T X w \qquad (2)$$

$$\text{subject to } w_i{}^T \left(I_p - W W^+\right) w_i = 1$$
$$and\ t_i{}^T t_j = w_i{}^T X^T X w_j = 0,$$
$$for\ j = 1, \cdots, i-1,$$

其中I_p是 p × p 的单位矩阵，

W^+是 W 的 More − Penrose 广义逆矩阵。

对这个优化问题，有两种重要的算法 NIPALS 和 Kernel-PLS 方法可以实现，在 R 的软件包里分别对应 pls 和 pls.pcr。

下面介绍第二种方法 SIMPLS(Statistically Inspired Modification of PLS)。

目标函数：SIMPLS

$$w_i = \underset{w}{\operatorname{argmax}}\, w^T X^T Y Y^T X w \quad (3)$$

$$\text{subject to } w_i{}^T w_i = 1$$
$$and\ t_i{}^T t_j = w_i{}^T X^T X w_j = 0,$$
$$for\ j = 1, \cdots, i-1$$

上面的目标函数中$w^T X^T Y Y^T X w$跟单相应变量是一样的，在多元的情况下，可以写成潜变量 T 与$Y_1, \cdots, Y_q$协方差的平方和。

$$w^T X^T Y Y^T X w = ((Xw)^T Y)^T ((Xw)^T Y)$$
$$= n^2 \sum_{j=1}^{q} \widehat{Cov}\left(T, Y_j\right)^2,$$

由此我们可以看出，SIMPLS 可以看为多元响应变量与单响应变量情况下的统一表达式。

求的权重矩阵 W 之后，我们就可以得到估计：

$$\hat{T} = X\widehat{W}$$
$$\hat{B} = \widehat{W}\hat{Q}^T$$

其中$\hat{Q}^T$是把$\hat{T}$带入到 Y 的表达式，使用最小二乘得到。

2. PLS 用于高维数据分析

PLS 用于回归分析

对连续变量，PLS 可以直接应用于回归分析中去。已知一些目标基因在不同条件下的芯片表达数据，以及转录因子跟基因见的连接信息(比如 CHIP 数据)，Boulesteix 和 Strimmer[7]假设表达数据与连接信息之间有以下的线性关系：

$$\text{Y} = \text{A} + \text{XB} + \text{E}$$

其中 Y 是n × q的数据矩阵，包含了 n 个基因在 q 种条件下的表达水平，X 是n × p的矩阵，包含的是 n 个基因对应 p 个转录因子的连接信息。A 是截距矩阵，E 是误差矩阵。要求解的 B 是一个p × q的矩阵，对应的是 p 个转录因子在 q 种条件下的活动水平。

这样，对转录因子的活动水平的估计，就可以通过一个简单的回归问题得到解决。

其他的也有很多应用，比如通过建立回归模型来构建基因之间的关联网络[8]，PLS 都能有很好的表现。

PLS 用于分类

前面我们讨论 PLS 应用于连续响应变量的回归分析，当响应变量为分类变量时，比如说 Y 有 K 个不同的取值(0 到 k-1)，一般会通过下面的变换来进行：

$$Y_j = \begin{cases} 1 & if\ Y = j-1 \\ 0 & otherwise \end{cases} \quad \text{for j} = 1, \cdots, \text{k}$$

在这样的变换下，Y 一开始先被当做连续变量做回归，得到系数的估计，然后用于后面的分类估计。这时得到的分类变量的值是连续的，再通过一定的变换把连数值对应回 0 或 1 的分类值。Musumarra 等人[9]使用多元 PLS，根据 8 种不同类型的一共 60 个癌症细胞系的 9605 个基因的表达数据进行回归，进而进行分类预测。结果证明 PLS 能取得很好的预测效果，尽管之前一直认为使用经典的线性回归并能很好的预测分类变量。

PLS 用于特征选择

与分类相关的一项工作就是特征选择，即对上面的分类问题，选择真正与分类相关的基因是非常重要的，这样就能够知道哪些基因能很好的用于区分不同类型的疾病，那么这些基因也就应该是对于疾病的发生起重要作用的基因。

一般情况下，对于响应变量是单变量(PLS1)时，会根据第一个权重向量 $w_1 = (w_{11}, \cdots, w_{p1})^T$ 来排列 p 个基因的重要程度。

我们根据 F 统计量在方差分析中的应用，记 F_j 是由 X 对第 j 个基因的计算结果，

$$F_j = (n-2)\frac{\sum_{k=0}^{1}\sum_{i:y_i=k}(\bar{x}_{kj} - \bar{x}_j)^2}{\sum_{k=0}^{1}\sum_{i:y_i=k}(\bar{x}_{ij} - \bar{x}_{kj})^2}$$

其中 $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$,

$$\bar{x}_{kj} = \frac{1}{n_k}\sum_{i:y_i=k} x_{ij}$$

n_k 是样本中属于第 k 类的数目。F_j 经常会被用来作为筛选基因的标准，根据它来排列基因对于分类问题的相关性。而 Boulesteix[10]也曾经证明对 PLS1 来说，F_j 是 ${w_{j1}}^2$ 的单调变换，由此看出使用权重向量来选择基因也就可行了。

当然，如果使用多个权重向量(对应多个潜变量)综合信息来排序就能利用更多的信息实现选择[11]。基因 j 对第 r 个 PLS 潜变量的影响因子 VIN_{rj} 定义为对应权重 ${w_{j1jr}}^2$ 以及第 r 个潜变量所能解释的平方和所占的比例的一个函数，那么变量 j 的重要因子 VIP_j 就是对应 c 个潜变量的 VIN_{rj} 之和，然后根据 VIP_j 对基因进行排序。

这种方法的优势在于它能够捕捉到 PLS c 个潜变量的信息，由此它可能就能识别出 F 统计量不能发现的的非线性模式。它的缺点就是 VIP 因子没有一个理论背景作为支撑，可能需要更深入了解它跟回归系数矩阵之间的关系。

PLS 还可以进行生存分析的应用[12]，在此我们就不详细介绍了。而且也有很多 PLS 与其他方法结合的算法[2]，这些各种各样的应用，是 PLS 作为一个通用的降维方法的最好证明。

3. SPLS 及其应用

通过对 PLS 估计的更深入了解，Chun 和 Keles[13]证明了 PLS 的估计在 p 非常大而样本又较少的情况下没有渐进一致性，也就是说这个估计此时不是一个好的估计。而且，PLS 所找出来的潜变量是所有原始预测变量的一个线性组合，而实际上在成千上万个预测变量中可能只有很少的一部分是真正与相应变量相关的。那么如何再降维的同时实现变量的选择就显得非常重要。

稀疏的偏最小二乘(Sparse partial least squares, SPLS)通过在 PLS 的优化目标中加入稀疏性的惩罚，实现在降维的过程中同时实现变量的选取。这样它所获得的潜变量就是原始预测变量的一个稀疏的线性组合。

SPLS 目标函数

$$\min_{w,c}\{-kw^TM + (1-k)(c-w)^TM(-w) + \lambda_1|c|_1 + \lambda_2{|c|_2}^2\} \quad (4)$$

$$\text{subject to } w^T w = 1$$

其中 $M = X^T Y Y^T X$，$\|c\|_1 = \sum_{i=1}^{p} |c_i|$，$\|w\|_2 = \sqrt{\sum_{i=1}^{p} {c_i}^2}$

由上面的目标函数可以看出，它是对权重 w 的一个替代 c 进行惩罚而不是 w 本身，并通过第二项来约束 c 跟 w 能够靠的足够近[13]。这样的设计能够实现在 w 稀疏的地方达到确实为 0 的特点。L_1惩罚是为了 w 的稀疏性，L_2惩罚是考虑到在求解 w 时M可能出现的奇异性。

对于求解上面的优化问题，可以依次对固定的 c 迭代求 w，然后再对固定的 w 求 c。具体算法可以参考 Chun 和 Keles 的文章[13]。

参数的设定

SPLS 的目标函数中需要四个参数，系数 k,λ_1, λ_2,还有要求的潜变量的个数K。一般λ_1和K可以通过正交验证的方法在算法运行之前预处理得到。当然这个过程相比与其他算法可能需要较长的时间。

而λ_2一般需要设定的比较大，因为当 Y 的维数 q 很小时，以防$M = X^T Y Y^T X$出现奇异性。而且这时，优化问题的解具有阈值函数的估计[13]。

系数 k 也是很重要的一个参数，当 $0.5 \le k \le 1$时，前面凹函数的部分$-kw^T Mw$占的比重较大，在求解最优点时容易陷入局部最优的僵局而走不出来。因此，一般取 $0 < k < 0.5$。

目标函数(4)被用于选取相关的预测变量并进行降维，我们设$\mathcal{A}$为相关变量的指标集，K 是潜变量的数目，$X_{\mathcal{A}}$是包含在$\mathcal{A}$里面的变量对应的设计矩阵，那么 SPLS 的算法可以总结如下[14]：

1) 设初始值 $\hat{\beta}^{\mathrm{PLS}} = 0$，$\mathcal{A} = \emptyset$，$k = 1, Y_1 = Y$;
2) 当$k \le K$，
   a) 通过解优化问题(4)求解得到估计$\hat{w}$，其中$M = X^T Y_1 {Y_1}^T X$;
   b) 更新指标集$\mathcal{A}$为

$$\{i: \widehat{w_i} \neq 0\} \cup \{i: \hat{\beta}_i^{\mathrm{PLS}} \neq 0\}.$$

   c) 运行 PLS 算法，使用设计矩阵$X_{\mathcal{A}}$，潜变量个数k;
   d) 使用新的 PLS 估计$\hat{\beta}^{\mathrm{PLS}}$，并且更新 $Y_1 \leftarrow Y - X\hat{\beta}^{\mathrm{PLS}}, k \leftarrow k + 1$.

SPLS 算法的实现已经由 Chun 和 Keles 给出，可以下载 R 的软件包 spls 进行使用。

SPLS 应用举例

eQTL 的定位是很典型的相应变量与预测变量都很高维的例子。响应变量是全基因组基因的转录表达性状，预测变量是全基因组的 SNP 标记，因此如何高效的实现标记位点的选择和降维对于解决这个问题就至关重要。

以前的很多对数量性状的研究方法都每次要么针对一个基因的转录信息要么针对一个标记位点的信息，很少能够实现预测变量与响应变量都是多元的统计分析，这种做法也提高了假阳性的概率，效果不好。

Chun 和 Keles[14]使用 SPLS 方法与聚类分析结合，应用多元响应变量的 SPLS 来解决了 eQTL 的定位问题。

第一步：对G × N的表达矩阵进行聚类。关于基因表达的聚类有很多方法，如 k-means, 层次聚类法等，方法可以根据实验的设计选择。

第二步：针对每一个聚类组 k 使用多元的 SPLS 进行回归分析，系数显著性的置信区间可以用 bootstrap 的方法获得。

$$Y_i^{(k)} = X_i B^{(k)} + E_i$$

$Y_i^{(k)}$是第 k 组G_k个基因的表达矩阵，i 代表针对 i 组人群的测量数据，X_i和$B^{(k)}$也都是对应的设计矩阵和系数矩阵。

通过他们的模拟实验及对小鼠肥胖数据的分析，说明了 SPLS 方法能取得很好的 qQTL 定位效果。

今年一篇使用 SPLS 处理 GWAS 数据[15]的文章也很有代表意义。通过 GWAS 数据来识别与疾病相关的位点一直是我们做关联分析的目标。但是很多 GWAS 的结果只能找出部分常见病的标记点，对于罕见病的位点或者复杂疾病的位点却一直没有明显的进展。

使用 SPLS 的方法，候选区域的 SNP 位点作为预测变量，响应变量是疾病的状况(生病与否)，建立偏最小二乘回归，可以在降维

找出潜变量的同时实现标记位点的选择。回归系数显著性检验可以使用经验分布检验，因为毕竟 SPLS 没有完善的统计理论，系数统计量也没有一个精确的分布进行描述。

如 Chun 等人[15]通过模拟及真实数据展示的一样，SPLS 能相对高效的(相对于一般的 PLS 或 PCA)识别出那些 MAF 小于 5%的位点，表现出很好的统计效率。

4. 讨论

偏最小二乘(PLS)回归是一个具有很好的通用性的降维方法，连续变量和离散的分类变量都可以适用。它通过捕捉预测变量与响应变量的相关性信息，实现了有监督的降维，相比无监督的主成分分析提高了效率。同时也解决了正常的回归分析最小二乘估计所不能处理的$n < p$的情况。

稀疏的偏最小二乘(SPLS)能在降维的同时实现变量选择，不仅保证了降维之后回归分析的正常进行，也筛选出与响应变量相关的变量，从而降低了假阴性，有更高的统计效率。但是 SPLS 相对来说计算复杂度要高，因为在算法开始之前，还要通过交叉验证来确定参数的值。

PLS 和 SPLS 都是一种基于数据的统计方法，并没有严格的统计理论支撑，在参数显著性检验时可以采用错误发现率(false discovery rate, fdr)[8],bootstrap[14]或者经验分布[15]检验的方法实现。

总之，在现在的数据时代，基于数据的偏最小二乘方法不失为一个顺应时代的有效方法，在高维数据分析中可以助一臂之力。

参考文献:

1.Maitra S,Yan J (2008) Principle Component Analysis and Partial Least Squares:Two Dimension Reduction Techniques for Regression. Casualty Actuarial Society:80-90.
2.Boulesteix AL,Strimmer K (2005) Partial Least Squares: A Versatile Tool for the Analysis of High-Dimensional Genomic Data. Seminar for Applied Stochastics.
3.Martens H (2001) Reliable and relevant modelling of real world data: a personal account of the development of PLS Regression. Chemometr Intell Lab 58(2):85-95.
4.Wold S (2001) Personal memories of the early PLS development. Chemometr Intell Lab 58(2):83-84.
5.Garthwaite PH (1994) An Interpretation of Partial Least-Squares. J Am Stat Assoc 89(425):122-127.
6.Martens H,Naes T (1989) Multivariate Calibration. New York: Wiley.
7. Boulesteix AL, Strimmer K (2005) Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach. Theor Biol Med Model 2:23.
8.Datta S, Pihur V,Datta S (2008) Reconstruction of genetic association networks from microarray data: a partial least squares approach. Bioinformatics 24:561-568.
9.Brown PO, Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JCE, Lashkari D, Shalon D, Myers TG, Weinstein JN,Botstein D (2000) Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 24:227-235.
10.Boulesteix AL (2004) PLS Dimension Reduction for Classification with Microarray Data. Statistical Applications in Genetics and Molecular Biology 3(1):A33.
11.Musumarra G, Barresi V, Condorelli DF,Scire S (2003) A bioinformatic approach to the identification of candidate genes for the development of new cancer diagnostics. Biol Chem 384:321-327.
12.Nguyen DV,Rocke DM (2002) Partial least squares proportional hazard regression for application to DNA microarray survival data. Bioinformatics 18:1625-1632.
13.Keles S,Chun H (2010) Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J Roy Stat Soc B 72:3-25.
14.Keles S,Chun H (2009) Expression Quantitative Trait Loci Mapping With Multivariate Sparse Partial Least Squares Regression. Genetics 182:79-90.
15.Chun HH, Ballard DH, Cho J,Zhao HY (2011) Identification of Association Between Disease and Multiple Markers Via Sparse Partial Least-Squares Regression. Genet Epidemiol 35:479-486.