

doi: 10. 11830/ISSN. 1000-5013. 201604027



偏最小二乘建模及其多重 共线抑制能力分析

杨春华, 杨玲

(保山学院 数学学院, 云南 保山 678000)

摘要: 首先,分析偏最小二乘法解决问题的思路,进而从数学角度刻画偏最小二乘法的四步建模过程.然后,利用数学归纳法证实偏最小二乘法对多重共线的抑制能力.最后,以某地区的供水能力评价为研究实例,证实偏最小二乘法的有效性.结果表明:偏最小二乘法完全适用于多变量复杂关系的求解.
关键词: 偏最小二乘; 数学归纳法; 多重共线; 回归分析
中图分类号: O 625. 63 **文献标志码:** A **文章编号:** 1000-5013(2016)04-0523-04

Partial Least Squares Modeling and Its Multiple Collinear Inhibition Capability Analysis

YANG Chunhua, YANG Ling

(School of Mathematics, Baoshan University, Baoshan 678000, China)

Abstract: Firstly, by analyzing the thinking route to solve the problem of the partial least square method, the authors describe four modeling steps to the partial least square method. Finally, we confirmed the inhibition ability of partial least squares method for multiple collinear by using the mathematical induction method. By evaluating the water supply capacity of an area as a case study, it really shows the validity of the partial least squares method. Results in this paper shows that partial least squares method is completely applicable to the solution of multi variable complex relationships.
Keywords: partial least squares; mathematical induction; multiple collinear; regression analysis

在参数估计和回归分析领域,多个自变量和多个因变量间的关系是一个非常复杂的问题^[1-3].采用最小二乘法等常规分析方法,难以达到预期的效果.这是因为多个自变量之间,多个因变量之间往往存在多重相关性,即多重共线性^[4-10].偏最小二乘回归分析利用信息综合筛选技术进行回归模型的构建,有效规避了原有变量的多重相关问题^[11-15].本文对偏最小二乘回归分析方法及其建模过程进行研究.

1 偏最小二乘法

假设分析的问题中,存在 m 个自变量,其集合可表述为 $\{p_1, p_2, \dots, p_m\}$;存在 n 个因变量,其集合可表述为 $\{q_1, q_2, \dots, q_n\}$.根据统计方法获取 m 个自变量和 n 个因变量的原始数据后,用 P, Q 对这些数据进行描述.其后,偏最小二乘法的执行,就是在数据对象 P 和 Q 上进行.
首先,在数据对象 P, Q 上各自提取 1 个主成分,分别用 α_1, β_1 表示.实际上, α_1 是集合 $\{p_1, p_2, \dots, p_m\}$ 中各个元素的 1 个线性组合,而 β_1 则是集合 $\{q_1, q_2, \dots, q_n\}$ 中各个元素的 1 个线性组合.在提取 $\alpha_1,$

β_1 时,需满足 2 个条件:第一, α_1, β_1 要尽可能多地表征数据对象 P, Q 的变异特征;第二, α_1, β_1 的关联水平可以达到最高. 然后,在提取 α_1, β_1 后,偏最小二乘法进一步对数据对象 P, Q 执行有关 α_1, β_1 的回归检验. 如果回归检验满足既定的精度要求,偏最小二乘法执行完毕;如果回归检验没有满足既定的精度要求,则需要根据 P, Q 被 α_1, β_1 描述后的剩余信息,再次执行成分提取,直至满足检验精度. 最后,偏最小二乘法会为数据对象 P 提取出 i 个成分,即 $\alpha_1, \alpha_2, \dots, \alpha_i$;偏最小二乘法为数据对象 Q 提取出 j 个成分,即 $\beta_1, \beta_2, \dots, \beta_j$. 多因变量集合 $\{q_1, q_2, \dots, q_n\}$ 中任一因变量,可描述为 $\alpha_1, \alpha_2, \dots, \alpha_i$ 的回归关系.

2 多变量问题的偏最小二乘建模

应用偏最小二乘法,对多变量问题进行建模求解时,有如下 4 个步骤.

步骤 1 对数据对象 P, Q 执行标准化处理,进一步得到自变量和因变量矩阵 P_0, Q_0 ,其过程为

$$p_{i,j}^* = \frac{p_{i,j} - \bar{p}_j}{s_j}, \quad p_{i,k}^* = \frac{p_{i,k} - \bar{q}_k}{s_k}, \tag{1}$$

$$P_0 = (p_{i,j}^*)_{m \times n}, \quad Q_0 = (q_{i,j}^*)_{m \times n}. \tag{2}$$

式(1),(2)中: \bar{p}_j, \bar{q}_k 为均值; s_j, s_k 为标准差.

步骤 2 从自变量矩阵 P_0 和因变量矩阵 Q_0 中提取第 1 个主成分,即

$$\alpha_1 = P_0 a_1, \quad \beta_1 = Q_0 b_1. \tag{3}$$

式(3)中: a_1 为 $P_0' Q_0 Q_0' P_0$ 的特征向量; b_1 为 $Q_0' P_0 P_0' Q_0$ 的特征向量.

数据对象 P, Q 和第 1 个主成分的回归关系,可以描述为

$$P_0 = \theta_1' \alpha_1 + P_1, \quad Q_0 = \vartheta_1' \beta_1 + Q_1. \tag{4}$$

式(4)中: θ_1, ϑ_1 为回归方程中的回归系数.

步骤 3 根据第 1 个主成分的回归方程,可以递推第 2 个主成分的回归方程,即

$$P_1 = \theta_2' \alpha_2 + P_2, \quad Q_1 = \vartheta_2' \beta_2 + Q_2. \tag{5}$$

以此类推,可以获得第 λ 个主成分的回归方程,即

$$P_{\lambda-1} = \theta_\lambda' \alpha_\lambda + P_\lambda, \quad Q_{\lambda-1} = \vartheta_\lambda' \beta_\lambda + Q_\lambda. \tag{6}$$

步骤 4 假设最终数据对象 P 的秩是 λ ,则有

$$P_0 = \theta_1' \alpha_1 + \theta_2' \alpha_2 + \dots + \theta_\lambda' \alpha_\lambda, \quad Q_0 = \vartheta_1' \beta_1 + \vartheta_2' \beta_2 + \dots + \vartheta_\lambda' \beta_\lambda + Q_\lambda. \tag{7}$$

最终,因变量 q^* 的有关自变量的偏最小二乘形式为

$$q^* = \sum_{i=1}^{\lambda} \vartheta_i \theta_{i,j} p_{i,j}^* + \dots + \sum_{i=m}^{m+\lambda} \vartheta_i \theta_{i,j} p_m^*. \tag{8}$$

3 偏最小二乘的多重共线抑制能力分析

对多重共线的抑制能力,是偏最小二乘法的重要特征. 为了证实偏最小二乘法在此方面的性能,只要证明偏最小二乘法提取的多个成分之间相互直交. 据此,考察如下命题是否成立.

命题 1 当 $h \neq l$ 时,偏最小二乘法获得的多个成分 $\alpha_1, \alpha_2, \dots, \alpha_\lambda$ 相互直交,即存在 $\alpha_h' \alpha_l = 0$.

证明 采用数学归纳法证明此命题.

首先,证明 α_1, α_2 之间是否是直交的,即是否存在 $\alpha_1' \alpha_2 = 0$.

$$\alpha_1' \alpha_2 = \alpha_1' P_1 a_2 = \alpha_1' (P_0 - \alpha_1 \theta_1') a_2 = [\alpha_1' P_0 - \alpha_1' \alpha_1 \theta_1'] a_2 = [\alpha_1' P_0 - \alpha_1' \alpha_1 \frac{\alpha_1' P_0}{\|\alpha_1\|^2}] a_2 = 0.$$

至此, α_1, α_2 之间的直交关系得到证实. 根据数学归纳法,只要假设在 $\alpha_1, \alpha_2, \dots, \alpha_h$ 直交的前提下,证实 $\alpha_1, \alpha_2, \dots, \alpha_{h+1}$ 也是直交的,命题中的结论就可以得到证实.

$$\alpha_h' \alpha_{h+1} = \alpha_h' P_h a_{h+1} = \alpha_h' (P_{h-1} - \alpha_h \theta_h') a_{h+1} = [\alpha_h' P_{h-1} - \alpha_h' \alpha_h \frac{\alpha_h' P_{h-1}}{\|\alpha_h\|^2}] a_{h+1} = 0,$$

$$\alpha_{h-1}' \alpha_{h+1} = \alpha_{h-1}' P_h a_{h+1} = \alpha_{h-1}' (P_{h-1} - \alpha_h \theta_h') a_{h+1} = [\alpha_{h-1}' P_{h-1} - \alpha_{h-1}' \alpha_h \frac{\alpha_h' P_{h-1}}{\|\alpha_h\|^2}] a_{h+1} = 0.$$

因 $\alpha_{h-1}' \alpha_h = 0$,有

$$\alpha'_{h-2}\alpha_{h+1} = \alpha'_{h-2}\mathbf{P}_h\mathbf{a}_{h+1} = \alpha'_{h-2}(\mathbf{P}_{h-1} - \alpha_h\theta'_h)\mathbf{a}_{h+1} = [\alpha'_{h-2}\mathbf{P}_{h-2} - \alpha_{h-1}\theta'_{h-1} - \alpha_h\theta'_h]\mathbf{a}_{h+1} = 0.$$

至此, $\alpha_1, \alpha_2, \cdots, \alpha_\lambda$ 之间的直交关系得到证实. 在原始问题的回归分析中, 那些变量都可以表征为 $\alpha_1, \alpha_2, \cdots, \alpha_\lambda$ 之间的回归组合, 而这些成分又是相互直交的, 这就不会存在多重共线问题.

4 偏最小二乘建模在实际问题中的应用

假设某地区供水能力的影响因素分别用 p_1, p_2, \cdots, p_n 表示, 从而构建 1 个多影响因素集合 $P = \{p_1, p_2, \cdots, p_n\}$. 假设某地区供水能力, 可以有多个指标表征, 如 q_1, q_2, \cdots, q_m , 从而构建一个多指标评价集合 $Q = \{q_1, q_2, \cdots, q_m\}$. 供水能力影响因素和供水能力评价指标, 可以分别得到 2 个观测矩阵, 即

$$\mathbf{P} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,n} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,n} \\ \vdots & \vdots & \cdots & \vdots \\ p_{n,1} & p_{n,2} & \cdots & p_{n,n} \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,m} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,m} \\ \vdots & \vdots & \cdots & \vdots \\ q_{m,1} & q_{m,2} & \cdots & q_{m,m} \end{bmatrix}. \tag{9}$$

因此, 某地区供水能力的分析、评价与预测, 就演变为 \mathbf{P}, \mathbf{Q} 的偏最小二乘模型求解.

某地区主要依靠 3 个水库进行供水, 按照地理位置, 分为东区水库、西区水库和南大坝水库. 为此, 设计如下指标: $q_1 \sim q_3$ 分别为该地区东部、西部、南部用水量需求; p_1 为该地区东部水库供水量; p_2 为该地区东部水库泄洪量; p_3 为该地区西部水库供水量; p_4 为该地区西部水库泄洪量; p_5 为该地区南部水库供水量; p_6 为该地区南部水库泄洪量. 上述参数主要来源于 2000—2015 年度统计数据, 如表 1 所示.

表 1 主要参数的年度数据
Tab. 1 Main parameters of the annual data 万 m³

年度	q_1	p_1	p_2	q_2	p_3	p_4	q_3	p_5	p_6
2000	882.224	1 123.678	2 024.569	524.271	876.529	1 796.883	437.989	933.486	1 853.527
2001	871.963	1 029.277	2 187.537	568.235	855.428	1 826.227	446.526	917.552	1 907.226
2002	869.858	1 108.576	2 088.696	553.421	893.217	1 766.459	431.203	903.449	1 911.428
2003	852.702	1 145.299	2 075.428	568.275	905.426	1 801.227	444.583	889.428	1 824.567
2004	839.699	1 084.323	2 100.572	529.331	845.581	1 822.520	472.115	926.455	1 799.431
2005	862.517	1 102.688	2 078.439	583.227	823.119	1 783.442	469.874	919.270	1 926.584
2006	853.427	1 251.417	2 148.272	596.542	918.607	1 809.117	475.634	1 054.353	1 931.406
2007	891.118	1 176.532	2 082.547	601.228	909.335	1 853.227	489.502	878.288	1 878.493
2008	902.207	1 128.579	2 069.457	600.589	926.384	1 822.417	481.443	928.445	1 823.765
2009	885.235	1 074.621	1 998.206	599.808	871.818	1 764.116	483.699	969.714	1 858.532
2010	896.768	1 048.885	2 086.867	605.417	888.281	1 787.535	490.471	983.527	2017.431
2011	914.223	1 271.457	2 139.587	606.293	875.363	1 810.268	493.224	900.631	1 972.546
2012	902.421	1 188.202	2 209.493	587.281	912.587	1 745.258	492.695	918.529	1 868.229
2013	903.568	1 285.219	2 225.112	594.552	903.442	1 835.812	486.218	893.458	1 953.526
2014	892.587	1 176.245	2 287.635	607.829	905.184	1 855.281	496.602	845.276	1 977.275
2015	917.335	1 128.564	2 108.527	618.243	827.458	1 794.325	491.777	997.135	1 898.587

将表 1 的数据, 代入供水能力偏最小二乘模型, 进而执行偏最小二乘分析, 回归系数如表 2 所示. 由表 2 可知: p_1, p_2 和 p_1 的关联程度最高; p_3, p_4 和 q_2 的关联程度最高; p_5, p_6 和 q_3 的关联程度最高. 该地区供水能力影响因素 P 和供水能力 Q 的复相关系数为 0.762 2, 这表明 P, Q 之间密切相关.

在上述模型下, 进一步以供水能力影响因素 ($p_1, p_2, p_3, p_4, p_5, p_6$) 年度环比值预测其在 2016—2018 年度的变化, 根据关联系数及偏最小二乘模型预测该地区供水能力 ($\bar{q}_1, \bar{q}_2, \bar{q}_3$) 在 2016—2018 年度的变化; 进而根据供水能力 (q_1, q_2, q_3) 年度环比预测其在 2016—2018 年度的变化, 算出该地

表 2 偏最小二乘得出的回归系数
Tab. 2 Regression coefficient obtained by using partial least squares

指标	q_1	q_2	q_3
p_1	0.098 8	0.012 2	0.009 4
p_2	0.065 7	0.005 4	0.006 3
p_3	0.014 5	0.085 4	0.004 8
p_4	0.009 8	0.057 7	0.002 7
p_5	0.010 1	0.007 5	0.079 2
p_6	0.006 3	0.004 2	0.061 8

区未来 3 年富余水量的情况($\Delta q_1, \Delta q_2, \Delta q_3$),结果如表 3 所示. 由表 3 可知:未来 3 年中,该地区的东部、西部、南部供水量都有盈余,能够满足当地供水的需求.

表 3 2016—2018 年度的预测结果
Tab. 3 Forecast results for 2016—2018

年度	q_1	q_2	q_3	\bar{q}_1	\bar{q}_2	\bar{q}_3	Δq_1	Δq_2	Δq_3
2016	909.587	609.339	490.202	945.386	648.229	555.281	35.799	38.890	65.079
2017	916.322	616.077	495.083	977.521	676.314	568.274	61.199	60.237	73.191
2018	920.531	622.218	501.688	1 082.527	685.523	571.358	161.996	63.305	69.670

5 结 束 语

梳理了偏最小二乘法的建模过程及其解决实际问题的思路,重点分析了其对多重共线的抑制. 以某地区的供水能力评价为研究实例,其结果充分证明偏最小二乘分析具有多重共线抑制能力,对于多个变量的复杂关系求解具有较强的适用性.

参考文献:

[1] ADRIANO D A G, SCHENONE A V. Unfolded partial least squares/residual bilinearization combined with the successive projections algorithm for interval selection: Enhanced excitation-emission fluorescence data modeling in the presence of the inner filter effect[J]. Analytical and Bioanalytical Chemistry, 2015, 22(5): 30-37.

[2] MOKHTARI A, KEYVANFARD M, EMAMI I. Simultaneous chemiluminescence determination of citric acid and oxalic acid using multi-way partial least squares regression[J]. RSC Advances, 2015, 37(5): 29214-29221.

[3] ELDEN L. Computing frechet derivatives in partial least squares regression[J]. Linear Algebra and Its Applications, 2015, 473(11): 316-338.

[4] ADJORLOLO C, MUTANGA O, CHO M A. Predicting C3 and C4 grass nutrient variability using in situ canopy reflectance and partial least squares regression[J]. International Journal of Remote Sensing, 2015, 36(6): 1743-1761.

[5] TZANAKAKIS V A, MAUROMOUSTAKOS A, ANGELAKIS A N. Prediction of biomass production and nutrient uptake in land application using partial least squares regression analysis[J]. Water, 2015, 7(1): 1-11.

[6] 彭卓华, 辛会敏. 矩阵方程 $\sum_{i=1}^l A_i X_i B_i = C$ 的最小二乘广义双对称解[J]. 湘潭大学学报(自然科学版), 2014, 36(1): 16-20.

[7] KUANG Boyan, TEKIN Y, MOUAZEN A M. Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon pH and clay content[J]. Soil and Tillage Research, 2015, 146(8): 243-252.

[8] 吴瑞红, 王亚丽, 张环冲, 等. 一种基于最小二乘支持向量机的葡萄酒品质评判模型[J]. 华侨大学学报(自然科学版), 2013, 34(1): 30-35.

[9] 魏引尚, 郑活勃, 王宁. 采空区自然“三带”特征的最小二乘法分析[J]. 西安科技大学学报, 2015, 35(2): 159-164.

[10] 胡德, 郭刚正. 最小二乘法、矩法和最大似然法的应用比较[J]. 统计与决策, 2015, 33(9): 20-24.

[11] 宋媛媛, 王萍, 张庆芳, 等. 基于最小二乘法的 TD-LTE 传播模型校正研究[J]. 电子测量技术, 2015, 38(1): 123-125.

[12] 李鑫, 张跃强, 刘进博, 等. 基于直线段对应的相机位姿估计直接最小二乘法[J]. 光学学报, 2015, 44(6): 203-213.

[13] 陈明晶, 方源敏, 陈杰. 最小二乘法和迭代法圆曲线拟合[J]. 测绘科学, 2016, 41(1): 194-197.

[14] 王鹏, 刁山菊, 张季谦. 基于最小二乘法的单摆实验数据处理[J]. 安庆师范学院学报(自然科学版), 2015, 36(1): 136-139.

[15] 张开远, 周孟然, 闫鹏程, 等. 基于最小二乘法的 pH 值温度补偿系统设计[J]. 传感器与微系统, 2015, 34(5): 109-111.

(责任编辑: 钱筠 英文审校: 黄心中)