Pure Mathematics 理论数学, 2018, 8(3), 230-238
Published Online May 2018 in Hans. http://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2018.83029

Hans汉斯

Mathematical Theory and Analogy Analysis of Least Square Method and Its Related Methods

Dongmei Xing

Department of Mathematics, Nanchang University, Nanchang Jiangxi
Email: 2404061160@qq.com

Received: Apr. 28th, 2018; accepted: May 4th, 2018; published: May 24th, 2018

Abstract

In the case of statistical analysis and modeling of various problems, the error analysis is often used by least square method or its related methods. In this paper, these mathematical theories are described in detail on least squares, partial least squares and principal component analysis. We sketch the applications of these methods. In the meantime, the case for which they are not applicable is explained. We outline the correlation among them and their different phases. At last, the parameter test in the regression equations is simply explained.

Keywords

Least Squares Method, Partial Least Squares Method, The Principal Components, Regression

最小二乘法及其相关方法的数学原理与类比分析

幸冬梅

南昌大学数学系，江西 南昌
Email: 2404061160@qq.com

收稿日期：2018年4月28日；录用日期：2018年5月4日；发布日期：2018年5月24日

摘 要

在对各种问题的统计分析和建模处理时，经常用到最小二乘法或其相关的方法进行误差分析处理。本文

文章引用: 幸冬梅. 最小二乘法及其相关方法的数学原理与类比分析[J]. 理论数学, 2018, 8(3): 230-238.
DOI: 10.12677/pm.2018.83029

详细叙述了最小二乘法、偏最小二乘法与主成分分析法的数学原理，简述了这些方法的应用场合，对于它们不适用的情况作了说明。描述了它们之间的关联性，概述了它们的相异性。最后，简单说明了回归方程中的参数检验。

关键词

最小二乘法，偏最小二乘法，主成分，回归

1. 引言

在对计量化学、生物信息学、食品研究、医学、药理学等方面的研究时，经常需要从海量的数据中寻找出内在的联系性。这些海量的数据或通过实验得到、或在实践中得到，然而从这些数据的表面上并不能得到很多的信息，必需对这些数据进行进一步的处理。数据通常可以分成两组向量：因变量的向量、自变量的向量，讨论它们之间的内在联系。最简单的情况是：每组向量中均只有一个元素。因变量的向量与自变量的向量可能存在一定的回归关系。这些回归关系一般是非线性的关系，如果不会引起很大的失真性，可以用线性回归的方式近似。最小二乘(Least Square (LS))回归是这些回归模型中常用的处理误差的方法。为了合理、快捷地搜寻数据中存在的内在关系，研究者对最小二乘回归进行了各种改进，较典型的方法有：主成分分析(Principal Component Analysis (PCA))法、核主成分分析(Kernel Principal Component Analysis (KPCA))法与偏最小二乘(Partial Least Squares (PLS))法。核主成分分析法是主成分分析法的变种；偏最小二乘方法是主成分分析法与最小二乘回归相结合得到的一种方法，常称为 PLS 方法。现有的文献比较多侧重于这些方法的使用，但分析它们的数学原理、类比这些方法的异同处，并通过事例进行讨论的不多见。下面介绍最小二乘法、主成分分析法和偏最小二乘方法的数学工作原理，讨论它们的联系性。

2. 最小二乘法、主成分分析法与偏最小二乘方法的数学原理

2.1. 最小二乘方法

最小二乘方法[1]能够对实际问题有比较好的解释，该方法的核心是所有估计值与被估计值之差的平方和达到最小。因变量向量用 y 表示，自变量向量为 x。若只有一个因变量，则此时的最小二乘方法即为最简单形式的最小二乘的曲线拟合，给定 m 个合适的函数 $r_1(x),\cdots,r_m(x)$ 及 n 组数据 $\{y(i),x(i)\}_{i=1}^{n}$，其中 $\{y(i),x(i)\}$ 表示因变量、自变量组 (y,x) 的第 i 组值。为了方便，记 $y(i)$ 为 y_i。选定适当的参数 $\beta_i=\bar{\beta}_i(i=0,1,\cdots,m)$，得到曲线

$$y=\bar{\beta}_0+\bar{\beta}_1 r_1(x)+\bar{\beta}_2 r_2(x)+\cdots+\bar{\beta}_n r_m(x),\ \text{s.t.}$$

$$L(\bar{\beta}_0,\bar{\beta}_1,\cdots,\bar{\beta}_m)\triangleq\min_{\beta_i,0\le i\le m}\sum_j\left(y(j)-\left(\beta_0+\beta_1 r_1(x(j))+\beta_2 r_2(x(j))+\cdots+\beta_m r_m(x(j))\right)\right)^2.$$

在没有其他限制条件下，$L(\beta_0,\beta_1,\cdots,\beta_m)$ 关于 $\beta_i(i=1,2,\cdots,m)$ 的一阶偏导为 0 是得到 $L(\beta_0,\beta_1,\cdots,\beta_m)$ 的极小值的必要条件。选取参数 $\beta_i=\bar{\beta}_i(i=0,1,\cdots,m)$ 的数学工作原理如下：

令 $Y=[y_1,y_2,\cdots,y_n]^{\mathrm{T}}$，$L(\beta_0,\cdots,\beta_m)=\sum_j\left(y_j-\left(\beta_0+\beta_1r_1(x_j)+\beta_2r_2(x_j)+\cdots+\beta_mr_m(x_j)\right)\right)^2$，分别求 L 关于 $\beta_i\,(i=0,\cdots,m)$ 的偏导数，并令其为 0，即

$$\frac{\partial L}{\partial\beta_0}=-2\sum_j\left(y_j-\left(\beta_0+\beta_1r_1(x(j))+\cdots+\beta_ir_i(x(j))+\cdots+\beta_mr_m(x(j))\right)\right)=0$$

$$\frac{\partial L}{\partial\beta_i}=-2\sum_j r_i(x(j))\left(y_j-\left(\beta_0+\beta_1r_1(x(j))+\cdots+\beta_ir_i(x(j))+\cdots+\beta_mr_m(x(j))\right)\right)=0\,(i=1,2,\cdots,m)$$

计算可得

$$\beta=(\bar{\beta}_0,\bar{\beta}_1,\cdots,\bar{\beta}_m)=\left(X^{\mathrm{T}}X\right)^{-1}\cdot X^{\mathrm{T}}Y \tag{1}$$

其中，$X=\begin{bmatrix}1 & r_1(x) & \cdots & r_m(x)\end{bmatrix}=\begin{bmatrix}1 & r_1(x(1)) & \cdots & r_m(x(1))\\ 1 & r_1(x(2)) & \cdots & r_m(x(2))\\ \vdots & \vdots & \ddots & \vdots\\ 1 & r_1(x(n)) & \cdots & r_m(x(n))\end{bmatrix}$。

选择合适的函数 $r_i(x)(i=1,2,\cdots,m)$ 并不容易，常简化曲线拟合为线性拟合或多项式拟合，最简单的模型为线性回归模型，即因变量 y 是自变量 $x=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$[1] 的线性组合：

$$m=n,\ y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n=\beta^{\mathrm{T}}\cdot(1,x^{\mathrm{T}}),\ \text{其中}\ \beta=(\beta_0,\beta_1,\cdots,\beta_n)^{\mathrm{T}}.$$

给定 n 组数据 $\left\{y_i,x(i)^{\mathrm{T}}\right\}_{i=1}^{n}$，$x(i)^{\mathrm{T}}=(x_{i1},\cdots,x_{in})$，待定的系数 $\beta_i\,(i=0,1,\cdots,n)$ 使得误差平方的叠加和最小，易得出

当 $\beta=\left(X^{\mathrm{T}}X\right)^{-1}\cdot X^{\mathrm{T}}Y\triangleq\bar{\beta}$ 时，$\sum_j\left(y_j-\left(\beta_0+\beta_1x_{j1}+\beta_2x_{j2}+\cdots+\beta_nx_{jn}\right)\right)^2$ 其达到最小值，此时

$$y=\bar{\beta}_0+\bar{\beta}_1x_1+\bar{\beta}_2x_2+\cdots+\bar{\beta}_nx_n,$$

其中 $X=\begin{bmatrix}1 & x_1 & \cdots & x_n\end{bmatrix}=\begin{bmatrix}1 & x_{11} & \cdots & x_{1n}\\ 1 & x_{21} & \cdots & x_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & \cdots & x_{nn}\end{bmatrix}$，$Y=[y_1,y_2,\cdots,y_n]^{\mathrm{T}}$。

线性回归模型 $y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n$ 是最简单的数据处理模型，不一定能够真正反映实际数据的内存规律；最小二乘方法涉及所有的变量，不同自变量数据之间可能存在全部或部分线性相关性关系，这时数据的有效信息可能不能直接体现出来。减少线性回归模型的变量的个数，但又不要丢弃太多的信息，主成分分析方法[2] [3]是可选方法之一。

2.2. 主成分分析法

针对不同自变量数据之间存在全部或部相关性关系，导致(1)式中 $X^{\mathrm{T}}X$ 成为病态矩阵，不再适合用最小二乘方法，W.F. Massy [3]提出了主成分回归(Principal Component Regression，简称 PCR)方法。主成分分析的核心在于降维，将原有的多个性能指标(即多个自变量)减少为只含有少数几个综合性能指标。这些综合性能指标是原来的性能指标的线性组合，它们相互独立且集中反映出原始变量的较多信息，在选择综合性能指标时，一般通过信息量与原有信息量的百分比(如 95%或更多、更少的百分比)来衡量。

[1] 如果只有一个自变量，则 $n=1$。
[2] 这里的 x 已经标准化，即已中心化、均值化。

设 x 是有 n 个随机变量组成的列向量[2]，研究它们所体现的信息，通常得计算出这 n 个随机变量的方差、协方差，找出这 n 个随机变量间有意义的相关性结构。然而若 n 较大，计算向量 x 的方差、协方差要花费比较多时间，除非 n 很小或者变量之间的相关性结构很简单，否则发现随机变量间的相关性结构比较困难。研究随机变量隐含的信息时，考虑研究一些维数远远低于 n 的派生变量，这些派生变量是原来变量的线性组合，它们保留了原来变量绝大部分信息，这样的派生变量称之为主成分。

主成分分析法中，寻找主成分的步骤如下[3]：

第一步：寻找第一个主成分，即找出 x 的一个线性组合 β_1，记为 $\beta_1 \triangleq \alpha_1^{\mathrm{T}} x$，其中 α_1 是一个 n 维列向量，$\alpha_1 = (\alpha_{11}, \alpha_{12}, \cdots, \alpha_{1n})^{\mathrm{T}}$，$x = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$，$\mathrm{var}(\beta_1)$ 为 β_1 的方差。

$$\beta_1 \triangleq \alpha_1^{\mathrm{T}} x = \alpha_{11} x_1 + \alpha_{12} x_2 + \cdots + \alpha_{1n} x_n,$$

s.t. $\mathrm{var}(\beta_1)$ 最大，且 $\alpha_1^{\mathrm{T}} \cdot \alpha_1 = 1$。

第二步：寻找第二个主成分，也即寻找 x 的第二个线性组合 β_2，记为 $\beta_2 \triangleq \alpha_2^{\mathrm{T}} x$，其中 α_2 是一个 n 维列向量，$\alpha_2 = (\alpha_{21}, \alpha_{22}, \cdots, \alpha_{2n})^{\mathrm{T}}$，$x = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$，

$$\beta_2 \triangleq \alpha_2^{\mathrm{T}} x = \alpha_{21} x_1 + \alpha_{22} x_2 + \cdots + \alpha_{2n} x_n,$$

s.t. $\mathrm{var}(\beta_2)$ 最大，且 $\alpha_2^{\mathrm{T}} \cdot \alpha_2 = 1$，$\beta_1$ 与 β_1 线性无关。

以此类推，直至找到线性关系 β_k，使其与 $\beta_1, \beta_2, \cdots, \beta_{k-1}$ 线性无关且具有最大方差。至此，k 个主成分找到了。

第三步：建立因变量与主成分的回归方程，其中最简单的是线性回归方程。可依据最小二乘回归得到数学模型。

获得主成分的数学推导过程描述如下。

两个随机变量具有线性无关性，是指两个随机变量之间的协方差为 0。在寻找主成分前，先对随机变量的数值进行规范化处理，使得随机变量的均值为 0，方差为 1。设随机变量的向量 x 的协方差矩阵为 Σ。

对于第一主成分 β_1：利用 Lagrange 乘子法，构造 Lagrange 函数

$$L(\alpha_1) \triangleq \mathrm{var}(\beta_1) - \lambda(\alpha_1^{\mathrm{T}} \alpha_1 - 1) = \alpha_1^{\mathrm{T}} \cdot \Sigma \cdot \alpha_1 - \lambda(\alpha_1^{\mathrm{T}} \alpha_1 - 1),$$

其中 λ 是 Lagrange 乘子，依据极值原理，$\dfrac{\partial L(\alpha_1)}{\partial \alpha_1} = 0$ 是在条件 $\alpha_1^{\mathrm{T}} \cdot \alpha_1 = 1$ 时 $\mathrm{var}(\beta_1)$ 的取得最大值的充要条件。上式两边对 α_1 求导，令其为 0，即

$$\frac{1}{2} \cdot \frac{\partial L(\alpha_1)}{\partial \alpha_1} = \Sigma \cdot \alpha_1 - \lambda \alpha_1 = (\Sigma - \lambda \cdot I_n) \cdot \alpha_1 = 0 \Rightarrow \Sigma \cdot \alpha_1 = \lambda \alpha_1$$

其中 I_n 是 n 阶单位矩阵。显然 λ 是 Σ 特征值，α_1 是相应的特征向量。由于 $\mathrm{var}(\beta_1) = \alpha_1^{\mathrm{T}} \cdot \Sigma \cdot \alpha_1 = \lambda \alpha_1^{\mathrm{T}} \alpha_1 = \lambda$，所以 $\max \mathrm{var}(\beta_1) = (\max \lambda) \alpha_1^{\mathrm{T}} \alpha_1 = (\max \lambda) \triangleq \lambda_1$。因此，第一主成分中出现的系数为协方差矩阵 Σ 的最大特征值 λ_1 的特征向量 α_1。

对于第二主成分 β_2：由于 β_1, β_2 线性无关，故 $\mathrm{cov}(\beta_1, \beta_2) = \alpha_1^{\mathrm{T}} \cdot \Sigma \cdot \alpha_2 = 0$，而 $\alpha_1^{\mathrm{T}} \cdot \Sigma \cdot \alpha_2 = \alpha_2^{\mathrm{T}} \cdot \Sigma \cdot \alpha_1 = \alpha_2^{\mathrm{T}} \cdot \lambda_1 \alpha_1 = \lambda_1 \alpha_2^{\mathrm{T}} \cdot \alpha_1 = \lambda_1 \alpha_1^{\mathrm{T}} \cdot \alpha_2 = 0$，从而有

$$\alpha_1^{\mathrm{T}} \cdot \Sigma \cdot \alpha_2 = \alpha_2^{\mathrm{T}} \cdot \Sigma \cdot \alpha_1 = 0 = \alpha_2^{\mathrm{T}} \cdot \alpha_1 = \alpha_1^{\mathrm{T}} \cdot \alpha_2$$

利用 Lagrange 乘子法，构造 Lagrange 函数

[3] 尽管主成分分析没有忽略协方差和相关性，但是更注重方差。

$$L(\alpha_2) \triangleq \mathrm{var}(\beta_2) - \lambda\left(\alpha_2^{\mathrm{T}}\alpha_2 - 1\right) - \gamma\left(\alpha_2^{\mathrm{T}}\alpha_1 - 0\right) = \alpha_2^{\mathrm{T}} \cdot \Sigma \cdot \alpha_2 - \lambda\left(\alpha_2^{\mathrm{T}}\alpha_2 - 1\right) - \gamma\left(\alpha_2^{\mathrm{T}}\alpha_1 - 0\right)$$

上式两边对 α_2 求导，令其为 0，即

$$\frac{1}{2} \cdot \frac{\partial L(\alpha_2)}{\partial \alpha_2} = \Sigma \cdot \alpha_2 - \lambda\alpha_2 - \gamma\alpha_1 = 0 \tag{2}$$

两边左乘 α_1^{T}，得

$$\begin{aligned} &\alpha_1^{\mathrm{T}} \cdot \Sigma \cdot \alpha_2 - \lambda\alpha_1^{\mathrm{T}} \cdot \alpha_2 - \gamma\alpha_1^{\mathrm{T}} \cdot \alpha_1 = 0 \\ &\Rightarrow \gamma\alpha_1^{\mathrm{T}} \cdot \alpha_1 = 0 \\ &\Rightarrow \gamma = 0 \end{aligned}$$

代入(2)，得 $\Sigma \cdot \alpha_2 - \lambda\alpha_2 = 0 \Rightarrow \Sigma \cdot \alpha_2 = \lambda\alpha_2$，显然 λ 是 Σ 的特征值，α_2 是相应的特征向量。由于 $\mathrm{var}(\beta_2) = \alpha_2^{\mathrm{T}} \cdot \Sigma \cdot \alpha_2 = \lambda\alpha_2^{\mathrm{T}}\alpha_2 = \lambda$，所以 $\max \mathrm{var}(\beta_2) = (\max \lambda)\alpha_2^{\mathrm{T}}\alpha_2 = (\max \lambda) \triangleq \lambda_2$。假设 Σ 没有重根，则 $\lambda_1 > \lambda_2$，λ_2 是 Σ 的第二大特征根，相应的特征向量 α_2 是第二主成分中出现的系数[4]。

类似地，可以依次从大到小求出 Σ 的其它特征根及特征向量，找出相应的主成分，通常依据 $\dfrac{\sum_{i=1}^{s}|\lambda_i|}{\sum_{i=1}^{t}|\lambda_i|} \geq 95\%$ or 90% 确定从 Σ 所有 t 个特征值 $\lambda_i (i = 1, \cdots, t)$ 中选取 s 个特征值 $\lambda_i (i = 1, \cdots, s)$。

2.3. 偏最小二乘法

偏最小二乘方法[4] [5] [6] [7] [8]是 20 世纪 80 年代从应用中总结出来的一种降维技术。偏最小二乘回归方法的一大优点是：考虑样本总体对预测值的影响程度时，也充分考虑了单个因素间的综合作用对预测的影响。在回归速率上，偏最小二乘方法比一般的多元回归方法更快一些，对样本的要求更加宽松。偏最小二乘回归方法可以看成是最小二乘回归、主成分方法及典型相关分析方法的综合体。

偏最小二乘方法的原理如下：

设 X[5] 是含有 n 个随机变量 $x_1, x_2, \cdots, x_n$ 组成的自变量的列向量，Y 是含有一个或多个随机变量 $y_1, y_2, \cdots, y_q (q \geq 1)$ 的因变量的列向量。找出 X 的 k 个线性组合 $\beta_i (i = 1, 2, \cdots, k)$，这些线性组合满足主成分分析的要求；在 Y 中找出 r 个线性组合 $\tau_i (i = 1, 2, \cdots, r)$，它们也满足主成分分析的要求。建立 β_i 与 τ_i 的回归关系，使它们之间的相关程度达到最大；并构造 β_i 与 X, Y 的回归关系，建立 Y 与 X 的回归关系。

偏最小二乘回归分析的步骤：

第一步：分别找出 X 与 Y 的第一主成分 β_1, τ_1，并且使 β_1, τ_1 的协方差 $\mathrm{cov}(\beta_1, \tau_1)$ 达到最大。

$$\beta_1 \triangleq X^{\mathrm{T}}\alpha_1 = \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1n}x_n,$$

$$\text{s.t.} \quad \max \mathrm{var}(\beta_1), \text{ 且 } \alpha_1^{\mathrm{T}} \cdot \alpha_1 = 1.$$

$$\tau_1 \triangleq Y^{\mathrm{T}}\mu_1 = \mu_{11}y_1 + \mu_{12}y_2 + \cdots + \mu_{1q}y_q,$$

$$\text{s.t.} \quad \max \mathrm{var}(\tau_1), \text{ 且 } \mu_1^{\mathrm{T}} \cdot \mu_1 = 1;$$

$$\max \mathrm{cov}(\beta_1, \tau_1).$$

上述的方法在推导过程中，作了如下假定：1) 各变量的特征维数不大；2) 过程不存在序列相关性；3) 过程是线性的。以上方法中要求 X 及 Y 均已经作了预处理，使 X 及 Y 中心化、均值化。

[4] 协方差矩阵 Σ 为实对称矩阵，依据线性代数相关知识，可得 Σ 中属于不同特征值的特征向量必定线性无关。

[5] 此处的 X、Y 与(1)式中的 X、Y 含义不同。

第二步：先建立 β_1 与 τ_1 的回归关系：$\tau_1=\beta_1 W^{\mathrm{T}}$，其中 W 为相应的权值；由 β_1 与 X 的线性回归方程及 τ 与 Y 的线性回归方程，得到 β_1 与 Y 的线性回归方程。

$$X=\beta_1 P^{\mathrm{T}}+E_1,$$
$$Y=\tau_1 Q^{\mathrm{T}}+F_1^{(0)},$$
$$Y=\beta_1 R_1^{\mathrm{T}}+F_1$$

当 $E_1,F_1^{(0)},F_1$ 均为 0 时，显然有 $P=\dfrac{X^{\mathrm{T}}\cdot\beta_1}{\beta_1^{\mathrm{T}}\cdot\beta_1}=\dfrac{X^{\mathrm{T}}\cdot\beta_1}{\|\beta_1\|^2}$，$Q=\dfrac{Y^{\mathrm{T}}\cdot\tau_1}{\tau_1^{\mathrm{T}}\cdot\tau_1}=\dfrac{X^{\mathrm{T}}\cdot\tau_1}{\|\tau_1\|^2}$，$R_1=\dfrac{Y^{\mathrm{T}}\cdot\beta_1}{\beta_1^{\mathrm{T}}\cdot\beta_1}=\dfrac{Y^{\mathrm{T}}\cdot\beta_1}{\|\beta_1\|^2}$；

一般地，$E_1,F_1^{(0)},F_1$ 分别是这些回归方程的残差矩阵。

第三步：分别用 E_1,F_1 替换 X,Y，重复第一步~第二步，依次寻找 X 的第二主成分 β_2。类似地，依次得到第三主成分 β_3，…，第 d 个主成分 β_d。

由第二步得到 Y 与 X 的主成分 β 的回归关系式：$Y=\beta_1 R_1^{\mathrm{T}}+\beta_2 R_2^{\mathrm{T}}+\beta_3 R_3^{\mathrm{T}}+\cdots+\beta_d R_d^{\mathrm{T}}+F_d$，利用第一步中的结果，容易得到 Y 与 X 的回归关系。

下面对**第一步**作详细分析，利用 Lagrange 乘子法求解 β_1,τ_1。

令 $L(\alpha,\mu)=\mathrm{cov}(\beta_1,\tau_1)-\lambda_1(\alpha^{\mathrm{T}}\alpha-1)-\lambda_2(\mu^{\mathrm{T}}\mu-1)$，

由于 $\mathrm{cov}(\beta_1,\tau_1)=\alpha_1^{\mathrm{T}}\,\mathrm{cov}(X,Y)\mu_1=\alpha_1^{\mathrm{T}}(X^{\mathrm{T}}\cdot Y)\mu_1$，

所以 $L(\alpha,\mu)=\alpha^{\mathrm{T}}(X^{\mathrm{T}}\cdot Y)\mu-\lambda_1(\alpha^{\mathrm{T}}\alpha-1)-\lambda_2(\mu^{\mathrm{T}}\mu-1)$，

依据极值原理，$L(\alpha,\mu)$ 关于 α,μ 在边界上取得极大值的充要条件为相应的偏导数为 0。下面分别对 α,μ 求偏导，得

$$\frac{\partial L}{\partial\alpha_1}=X^{\mathrm{T}}Y\mu_1-2\lambda_1\alpha_1=0\Rightarrow X^{\mathrm{T}}Y\mu_1=2\lambda_1\alpha_1,\tag{i}$$

$$\frac{\partial L}{\partial\mu_1}=Y^{\mathrm{T}}X\alpha_1-2\lambda_2\mu_1=0\Rightarrow Y^{\mathrm{T}}X\alpha_1=2\lambda_2\mu_1,\tag{ii}$$

所以由(i)，(ii)可得

$$(X^{\mathrm{T}}Y)(Y^{\mathrm{T}}X)\alpha_1=(X^{\mathrm{T}}Y)\cdot(2\lambda_2\mu_1)=2\lambda_2(X^{\mathrm{T}}Y)\cdot\mu_1=2\lambda_2(2\lambda_1\alpha_1)=2\lambda_2\cdot2\lambda_1\alpha_1$$
$$(Y^{\mathrm{T}}X)(X^{\mathrm{T}}Y)\mu_1=(Y^{\mathrm{T}}X)\cdot(2\lambda_1\alpha_1)=2\lambda_1(Y^{\mathrm{T}}X)\cdot\alpha_1=2\lambda_1(2\lambda_2\mu_1)=2\lambda_2\cdot2\lambda_1\mu_1$$
$$\mu_1^{\mathrm{T}}\cdot(Y^{\mathrm{T}}X)\alpha_1=\alpha_1^{\mathrm{T}}\cdot(2\lambda_1\alpha_1)=2\lambda_1(\alpha_1^{\mathrm{T}}\cdot\alpha_1)=2\lambda_1\cdot(Y^{\mathrm{T}}X)\alpha_1=\mu_1^{\mathrm{T}}\cdot(2\lambda_2\mu_1)=2\lambda_2\mu_1^{\mathrm{T}}\cdot\mu_1=2\lambda_2$$
$$\alpha_1^{\mathrm{T}}\cdot(X^{\mathrm{T}}Y)\mu_1=\alpha_1^{\mathrm{T}}\cdot(2\lambda_1\alpha_1)=2\lambda_1(\alpha_1^{\mathrm{T}}\cdot\alpha_1)=2\lambda_1$$

注意到 $\mu_1^{\mathrm{T}}\cdot(Y^{\mathrm{T}}X)\alpha_1=(\alpha_1^{\mathrm{T}}\cdot(X^{\mathrm{T}}Y)\mu_1)^{\mathrm{T}}\Rightarrow2\lambda_2=(2\lambda_1)^{\mathrm{T}}=2\lambda_1$。

由此可得出，α_1 是矩阵 $(X^{\mathrm{T}}Y)(Y^{\mathrm{T}}X)$ 的最大特征值所对应的特征向量，也是 X 的第一主成分 β_1 的系数部分；μ_1 是矩阵 $(Y^{\mathrm{T}}X)(X^{\mathrm{T}}Y)$ 的最大的特征值所对应的特征向量，也是 Y 的第一主成分 τ_1 的系数。

采用 PLS 进行计算机处理数据时，常常通过迭代方式进行。在 n 维空间中任意选取一个非零列向量 u (与列向量 Y 有相同的维数的向量)，PLS 回归迭代过程包括 6 个步骤[3]：

1) $w=X^{\mathrm{T}}u/(u^{\mathrm{T}}u)$　　4) $c=Y^{\mathrm{T}}t/(t^{\mathrm{T}}t)$

2) $\|w\|\to1(w=w/\|w\|)$　　5) $\|c\|\to1(c=c/\|c\|)$

3) $t=Xw$　　6) $u=Yc$

如果 Y 是一维向量，我们用 y 表示 Y，此时有 $u=y$；当满足精确度要求时，迭代步骤停止。可以证明 w 是 α_1 的近似值，c 为 μ_1 的近似值，说明如下：

设最初选取的非零向量为 $u^{(0)}$，第一次利用 PLS 的 6 个步骤的 1) 3) 4) 6)得到的为 $w^{(1)},t^{(1)},c^{(1)},u^{(1)}$；由 $u^{(k)}$ 利用 PLS 的 6 个步骤的 1) 3) 4) 6)得到的为 $w^{(k+1)},t^{(k+1)},c^{(k+1)},u^{(k+1)}$

$$\begin{aligned}w^{(k+1)} &= X^{\mathrm{T}}u^{(k)}\Big/\left(u^{(k)\mathrm{T}}u^{(k)}\right)\\ &= X^{\mathrm{T}}Yc^{(k)}\Big/\left(\left(u^{(k)T}u^{(k)}\right)\cdot\left\|c^{(k)}\right\|\right)\\ &= X^{\mathrm{T}}YY^{\mathrm{T}}t^{(k)}\Big/\left(\left(u^{(k)\mathrm{T}}u^{(k)}\right)\cdot\left\|c^{(k)}\right\|\cdot\left(t^{(k)\mathrm{T}}t^{(k)}\right)\right)\\ &= X^{\mathrm{T}}YY^{\mathrm{T}}Xw^{(k)}\Big/\left(\left(u^{(k)\mathrm{T}}u^{(k)}\right)\cdot\left\|c^{(k)}\right\|\cdot\left(t^{(k)\mathrm{T}}t^{(k)}\right)\cdot\left\|w^{(k)}\right\|\right)\end{aligned}$$

所以

$$w^{(k+1)} = X^{\mathrm{T}}YY^{\mathrm{T}}X\cdot w^{(k)}\Big/\left(\left(u^{(k)\mathrm{T}}u^{(k)}\right)\cdot\left\|c^{(k)}\right\|\cdot\left(t^{(k)\mathrm{T}}t^{(k)}\right)\cdot\left\|w^{(k)}\right\|\right)$$

$$\left(X^{\mathrm{T}}YY^{\mathrm{T}}X\right)\cdot w^{(k)} = \left(\left(u^{(k)\mathrm{T}}u^{(k)}\right)\cdot\left\|c^{(k)}\right\|\cdot\left(t^{(k)\mathrm{T}}t^{(k)}\right)\cdot\left\|w^{(k)}\right\|\right)\cdot w^{(k+1)}$$

类似地，有以下式子成立

$$c^{(k+1)} = Y^{\mathrm{T}}XX^{\mathrm{T}}Yc^{(k)}\Big/\left(\left(t^{(k+1)\mathrm{T}}t^{(k+1)}\right)\cdot\left\|w^{(k+1)}\right\|\cdot\left(u^{(k+1)\mathrm{T}}u^{(k+1)}\right)\cdot\left\|c^{(k)}\right\|\right)$$

$$\therefore\quad c^{(k+1)} = Y^{\mathrm{T}}XX^{\mathrm{T}}Yc^{(k)}\Big/\left(\left(t^{(k+1)\mathrm{T}}t^{(k+1)}\right)\cdot\left\|w^{(k+1)}\right\|\cdot\left(u^{(k)\mathrm{T}}u^{(k)}\right)\cdot\left\|c^{(k)}\right\|\right)$$

$$\left(Y^{\mathrm{T}}X\cdot X^{\mathrm{T}}Y\right)c^{(k)} = \left(\left(t^{(k+1)\mathrm{T}}t^{(k+1)}\right)\cdot\left\|w^{(k+1)}\right\|\cdot\left(u^{(k)\mathrm{T}}u^{(k)}\right)\cdot\left\|c^{(k)}\right\|\right)\cdot c^{(k+1)}$$

3. 比较最小二乘方法、主成分分析法及偏最小二乘法

最小二乘方法、主成分分析法与偏最小二乘方法，均是从误差最小化方面去寻找合适的参数。这些方法在处理误差时，都默认了误差分布是理想的[1]，即模型误差是独立同分布的且服从其均值为 0、方差为 σ^2 的正太分布。下面以最小二乘方法进行说明。

给定需要拟合的数据 $\left\{\left(x_{i1},x_{i2},\cdots,x_{ik},y_i\right),i=1,2,\cdots,n,n>k\right\}$，利用下面的多元回归模型(即因变量关于自变量的数学期望)进行数据拟合：$\mu_{y|x_1,x_2,\cdots,x_k}=\beta_0+\beta_1r_1\left(x\right)+\beta_2r_2\left(x\right)+\cdots+\beta_kr_k\left(x_k\right)$，利用最小二乘估计得到参数 $\beta_0,\beta_1,\beta_2,\cdots,\beta_k$，每组拟合数据 $\left(x_{i1},x_{i2},\cdots,x_{ik},y_i\right)$ 均满足方程

$$y_i=\beta_0+\beta_1r_1\left(x_{i1}\right)+\beta_2r_2\left(x_{i2}\right)+\cdots+\beta_kr_k\left(x_{ik}\right)+\varepsilon_i$$

其中 ε_i 是模型的随机误差，它独立同分布，服从均值为 0，方差为 σ^2 的正太分布。

这里考虑因变量是自变量的多元线性回归模型，且只有一个因变量。从上述描写可知，在利用最小二乘法时，要求 $X^{\mathrm{T}}X$ 可逆。如果 X 的列之间存在多重共线，纵然 $X^{\mathrm{T}}X$ 可逆，由(1)计算得 $\beta=\left(\beta_0,\beta_1,\cdots,\beta_i,\cdots,\beta_m\right)^{\mathrm{T}}$ 建立的数学模型，很可能不具有通用性。

X 的列之间存在多重共线时，利用主成分分析法可消除这种共线性的不良影响。主成分分析法从原有的自变量向量中寻找它们的某些线性组合(即综合的指标或称之为综合自变量)，在信息损失最小的原则下，得到少数几个综合自变量。主成分分析法可以在一定程度上消除了原自变量之间的多重共线性，而且降低了待处理数据的维数。处理有周期性变化规律的数据时，如经济运行中的得到的各类经济数据，主成分分析法与谱分析方法相结合[9]，分析数据的变化趋势与规律。

主成分分析法对 X 进行主成分分析，没有考虑因变量 Y，也没有考虑 X 对 Y 的解释作用。PLS 方法

解决了这种缺憾。PLS 回归(Partial Least-Square Regression，简称 PLSR)方法充分利用了 PCR 方法对信息(自变量数据)提取的思想。

PLSR 方法是 S. Wold 和 C. Albano [4]在 1983 年提出的。许多研究工作者[5] [6] [7] [8]对 PLSR 进行了研究，他们还研究了利用迭代的方式计算主成分(参考上文中“PLS 回归迭代过程”关内容)。PLSR 是主成分分析法与最小二乘法的综合体，它对自变量和因变量均提取主成分，同时还考虑了因变量的作用以及自变量 X 对因变量的解释作用，一定程度上消除了基于主成分的回归模型的不可靠性。

如果响应变量类似于二值情况，回归模型的随机误差 ε_i 明显不是理想状态，即 ε_i 不服从均值为 0、方差为 σ^2 的正太分布。此时均值与方差均可能是随着自变量而发生改变，不可以利用多元回归方法中的最小二乘法来解决问题[1]。例如，Logistic 回归模型是一种比较理想的对二值响应的拟合，可以利用梯度下降法得到 Logistic 回归函数中的待估参数。

4. 拟合回归模型优良性检验说明

对于已经给的数据分析研究得到相应的回归模型时，通常只需要作参数检验[1]，检查 R^2 或 R_{adj}^2 [6]。如果得到的多个自变量数据与正交性差距太大，即它们明显不具有正交性，那么得到的模型可能过度拟合，此时需要进行交互验证[1]。我们下一步要做的工作是针对具体问题，利用前述方法建立模型并进行参数检验与参数在实际情况的意义的解释。

参考文献

[1] Walpole, R.E., Myers, R.H., Myers, S.L. and Ye, K.Y. 理工科概率统计[M]. 周勇, 马昀蓓, 谢尚宇, 王晓婧, 译. 北京: 机械工业出版社, 2010.

[2] Hotelling, H. (1933) Analysis of a Complex of Statistical Variables into Principal Components. *Education Psychology*, **24**, 417-444. https://doi.org/10.1037/h0071325

[3] Massy, W.F. (1965) Principal Components Regression in Exploratory Statistical Research. *Journal of the American Statistical Association*, **60**, 234-256. https://doi.org/10.1080/01621459.1965.10480787

[4] Wold, S., Albano, C. and Dun, M. (1983) Pattern Regression Finding and Using Regularities in Multivariate Data. Analysis Applied Science Publication, London.

[5] Rosipal, R. and Krämer, N. (2006) Overview and Recent Advances in Partial Least Squares. *Subspace, Latent Structure and Feature Selection*, Bohinj, Slovenia, 23-25 February 2005, 34-51. https://doi.org/10.1007/11752790_2

[6] Wold, H. (1982) Soft Modeling: The Basic Design and Some Extensions. In: Jöreskog, J.-K. and Wold, H., Eds., *Systems under Indirect Observation*, *Volume* 2, North Holland, Amsterdam, 1-53.

[7] Wold, H. (1985) Partial Least Squares. In: Kotz, S. and Johnson, N.L., Eds., *Encyclopedia of the Statistical Sciences*, *Vol.* 6, John Wiley, New York, 581-591.

[8] Wold, S., Ruhe, H., Wold, H. and Dunn III, W.J. (1984) The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverse. *SIAM Journal of Scientific and Statistical Computations*, **5**, 735-743. https://doi.org/10.1137/0905052

[9] 张红, 谢娜. 基于主成分分析与谱分析的房地产市场周期研究[J]. 清华大学学报(自然科学版), 2008, 48(9): 24-27.

[6] R^2 或 R_{adj}^2 含义见附录。

附　录

定理[1]：对于线性回归方程

$$y = X\beta + \varepsilon ,$$

方差 σ^2 的无偏估计可由下面的均方误差或均方残差 s^2 给出，即

$$s^2 = \frac{SSE}{n-k-1}, \text{ 其中 } SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 ,$$

其中 ε 是一个服从均值为 0、方差接近于 0 的正太分布的随机变量，k 为自变量的个数，n 为数据组的个数，y_i 与 $\hat{y}_i$ 分别是因变量 y 的第 i 个的观测值与模型估计值。

平方和等式

$$\underset{\uparrow}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} = \underset{\uparrow}{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2} + \underset{\uparrow}{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$$

$$SST \qquad SSR \qquad SSE$$

中，$\bar{y}$ 为因变量 y 的 n 个观测数据的平均值；SSR 的自由度为 k，而 SST 的自由度为 $n-1$，因此 SSE 的自由度为 $n-k-1$。

下述系数 R^2, R_{adj}^2 是用来衡量拟合回归模型优良性的一个常用指标

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} = 1 - \frac{SSE}{SST}$$ .....线性回归方程所解释的**变差占总变差**的**比例**。

$$R_{adj}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$ .....依据自由度调整后的 R^2 值。