统计与机器学习

第三章：多重共线性

倪 葎

DaSE@ECNU
(lni@dase.ecnu.edu.cn)

2021 年 3 月 24 日

目录

多重共线性的定义与原因

动机

- 在线性模型中，最小二乘估计

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

  - 如果自变量 $x_1, x_2, \cdots, x_p$ 是**完全线性相关**，即

$$|\boldsymbol{X}'\boldsymbol{X}| = 0,$$

    那么我们无法得到最小二乘估计 $\hat{\boldsymbol{\beta}}$.

多重共线性的定义与原因

动机

- 在线性模型中，最小二乘估计

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

  - 如果自变量 $x_1, x_2, \cdots, x_p$ 之间**相关性很高，但不完全线性相关**，即

$$|\boldsymbol{X}'\boldsymbol{X}| \approx 0,$$

    那么我们虽然可以得到最小二乘估计 $\hat{\boldsymbol{\beta}}$.

  - 但是，由于

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

    因此，虽然可以得到最小二乘估计 $\hat{\boldsymbol{\beta}}$，但是精度很低.

多重共线性的定义与原因

例子

- 考虑对因变量 y 和两个自变量 x_1 和 x_2 建立线性回归.
- 假定 y 与 x_1, x_2 都已经中心化.
- 回归方程为

$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

- 将 x_1 与 x_2 之间的相关系数记为

$$r_{12} = \frac{l_{12}}{\sqrt{l_{11} l_{22}}}.$$

其中

$$l_{11} = \sum_{i=1}^{n} x_{i1}^2, \quad l_{22} = \sum_{i=1}^{n} x_{i2}^2, \quad l_{12} = \sum_{i=1}^{n} x_{i1} x_{i2}.$$

多重共线性的定义与原因

例子（续）

- 最小二乘估计 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)'$ 的方差-协方差矩阵为

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2 \begin{pmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} \\ \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 \end{pmatrix}^{-1} \\
&= \sigma^2 \begin{pmatrix} l_{11} & l_{12} \\ l_{12} & l_{22} \end{pmatrix}^{-1}
\end{aligned}$$

多重共线性的定义与原因

线性代数知识（补充）

- 如何 n 阶方阵 A 的逆矩阵？采用伴随矩阵的方法.
- 第一步，计算 A 的行列式 $|A|$.
- 第二步，计算行列式 $|A|$ 的各个元素的代数余子式 A_{ij}.
- 第三步，构造方阵 A 的伴随矩阵，即

$$A^* = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{12} & \cdots & A_{n2} \\ \vdots & \vdots & & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{pmatrix}$$

- 第四步，A 的逆矩阵为 $A^{-1} = \frac{A^*}{|A|}$.

多重共线性的定义与原因

例子（续）

- 最小二乘估计 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)'$ 的方差-协方差矩阵为

$$
\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2 \begin{pmatrix} l_{11} & l_{12} \\ l_{12} & l_{22} \end{pmatrix}^{-1} \\
&= \sigma^2 \frac{1}{|\boldsymbol{X}'\boldsymbol{X}|} \begin{pmatrix} l_{22} & -l_{12} \\ -l_{12} & l_{11} \end{pmatrix} \\
&= \sigma^2 \frac{1}{l_{11}l_{22} - l_{12}^2} \begin{pmatrix} l_{22} & -l_{12} \\ -l_{12} & l_{11} \end{pmatrix} \\
&= \sigma^2 \frac{1}{l_{11}l_{22}(1 - r_{12}^2)} \begin{pmatrix} l_{22} & -l_{12} \\ -l_{12} & l_{11} \end{pmatrix}
\end{aligned}
$$

多重共线性的定义与原因

例子（续）

- 因此，我们得到

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{(1-r_{12}^2)l_{11}} \quad (1)$$

$$\mathrm{Var}(\hat{\beta}_2) = \frac{\sigma^2}{(1-r_{12}^2)l_{22}} \quad (2)$$

- 结论：随着自变量 x_1 和 x_2 的相关性**增加**，$\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差均会**增大**.
- 特别地，当 x_1 与 x_2 接近于完全相关时，$r \approx 1$，那么回归参数的估计值的方差将**趋于无穷**.

多重共线性的定义与原因

说明

- 自变量之间完全不想干的情形非常少见.
- 尤其，当自变量个数较多时，我们很难找到一组自变量，它们不但互相不相关，而且对因变量有显著影响.
- 考虑多个自变量，
  - 当自变量之间的相关性**较弱**时，我们一般认为符合多元线性回归模型设计矩阵的要求；
  - 当自变量之间的相关性**较强**时，我们一般认为违背多元线性回归模型的基本假设.

多重共线性的定义与原因

说明

- 自变量之间完全不想干的情形非常少见.
- 尤其，当自变量个数较多时，我们很难找到一组自变量，它们不但互相不相关，而且对因变量有显著影响.
- 考虑多个自变量，
  - 当自变量之间的相关性**较弱**时，我们一般认为符合多元线性回归模型设计矩阵的要求；
  - 当自变量之间的相关性**较强**时，我们一般认为违背多元线性回归模型的基本假设.

多重共线性的定义与原因

常见场景

- 对于分类变量，设置过多的虚拟变量.
  - 以**性别**为例.
  - 通常采用的虚拟变量

$$x_m = I(\text{性别为男性}) \quad \text{和} \quad x_f = I(\text{性别为女性}).$$

  - 问题：在建模时，不会将这两个虚拟变量同时纳入模型. 这是为什么?
  - 原因：$x_f = 1 - x_m$，即 x_f 可由 x_m 完全线性表示.
  - 解决方案: 通常有 J 个分类的变量, 至多可以设置 $J-1$ 个虚拟变量.

多重共线性的定义与原因

常见场景

- 对于分类变量，设置过多的虚拟变量.
  - 以**性别**为例.
  - 通常采用的虚拟变量

$$x_m = I(\text{性别为男性}) \quad \text{和} \quad x_f = I(\text{性别为女性}).$$

  - 问题：在建模时，不会将这两个虚拟变量同时纳入模型. 这是为什么?
  - 原因：$x_f = 1 - x_m$，即 x_f 可由 x_m 完全线性表示.
  - 解决方案: 通常有 J 个分类的变量, 至多可以设置 $J-1$ 个虚拟变量.

多重共线性的定义与原因

常见场景

- 某一个变量是由其他变量计算而成的.
  - 例如，在研究体型越大的鸟类更容易找到配偶的问题中，这类鸟有一种特别形态的尾部，研究者想探索鸟的整体大小和尾部大小是否有助于其找到配偶.
  - 研究者考虑了三个自变量：鸟的体长、尾长以及整体的长度.
  - 注意到，整体长度是体长和尾长之和.
  - 问题：是否可以将这三个自变量同时纳入模型?
  - 不能!
  - 解决方案：选择合适的自变量.

多重共线性的定义与原因

常见场景

- 模型中选用同样的或相似的自变量.
  - 同一概念但采用不同的测量方法，由此构造自变量纳入模型.
  - 例如，在研究收入与压力水平的关系时，度量收入的自变量有很多，如：个人收入、家庭收入等. 由于这些自变量都可以度量“收入”这一概念，往往具有很高的相关性.
  - 解决方案：通过**主成分分析**合并为一个表示“收入”量，作为自变量放入模型中.

方差扩大因子法

定义

- 自变量 $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p)$ 做标准化，记为 $\boldsymbol{X}_s$.
- 而标准化后自变量的相关系数矩阵为 $\boldsymbol{X}_s{}'\boldsymbol{X}_s = (r_{ij})$.
- 其逆矩阵为

$$\boldsymbol{C} = (c_{ij}) = (\boldsymbol{X}_s{}'\boldsymbol{X}_s)^{-1}$$

- 称方阵 $\boldsymbol{C}$ 主对角线元素

$$\mathrm{VIF}_j = c_{jj}$$

为自变量 x_j 的**方差扩大因子** (variance inflation factor, VIF) .

方差扩大因子法

命名的由来

- 在对因变量 y 和自变量 $x_1, \cdots, x_p$ 均进行标准化后，我们已经证明过使用标准化的数据和未使用标准化的数据所得到的最小二乘估计的关系为

$$\hat{\beta}_{s,j} = \frac{\sqrt{l_{jj}}}{\sqrt{l_{yy}}}\hat{\beta}_j.$$

- 这里我们仅仅考虑对自变量进行标准化，那么所得到的最小二乘估计的关系为

$$\hat{\beta}_{s,j} = \sqrt{l_{jj}}\hat{\beta}_j$$

其中，$l_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$.

方差扩大因子法

命名的由来

- 因此，易证

$$\mathrm{Var}(\hat{\beta}_j) = \frac{c_{jj}}{l_{jj}}\sigma^2, \quad j = 1, 2, \cdots, p$$

- 说明：由于 c_{jj} 越大，自变量 x_j 所对应回归系数 β_j 的最小二乘估计 $\hat{\beta}_j$ 的方差也越大。

问题：**为什么方差扩张因子能够用于诊断自变量间存在多重共线性呢?**

方差扩大因子法

另一个角度来看 VIF

- 我们将 x_j 作为因变量，与其余的 $p-1$ 个自变量建立多元线性回归模型。
- 记 R_j^2 表示其复决定系数。
- 可以证明，

$$c_{jj} = \frac{1}{1 - R_j^2}$$

- 说明：
  - R_j^2 度量了自变量 $\boldsymbol{x}_j$ 与其余自变量的线性相关程度。
  - 如果 R_j^2 越大，$\mathrm{VIF}_j(c_{jj})$ 也越大。

方差扩大因子法

如何利用 VIF 判断?

- 基本思想：VIF_j 越大，自变量 x_j 与其他自变量之间的多重共线性程度更严重。

- 判断标准：
  - 当 $\mathrm{VIF}_j < c_{\mathrm{VIF}}$ 时，自变量 $\boldsymbol{x}_j$ 与其余自变量之间不存在多重共线性；
  - 当 $\mathrm{VIF}_j \geq c_{\mathrm{VIF}}$ 时，自变量 $\boldsymbol{x}_j$ 与其余自变量之间存在多重共线性；

- 临界值 c_{VIF} 常见的取值有 $5, 10, 100$。

方差扩大因子法

如何利用 VIF 判断?

- 如何度量整个设计矩阵的多重共线性?
- 用 p 个自变量所对应的方差扩大因子的平均数来度量多重共线性，即

$$\overline{\mathrm{VIF}} = \frac{1}{p}\sum_{j=1}^{p}\mathrm{VIF}_j.$$

- 判断准则：当 $\overline{\mathrm{VIF}}$ 特别大时，表示存在严重的多重共线性问题。
- 值得注意的是，当样本量比较小时，R^2 较容易接近 1。因此 $\overline{\mathrm{VIF}}$ 的讨论需要基于样本量而讨论。

特征值判定法

线性代数知识（补充）

- 假定一个 n 阶方阵 $\boldsymbol{A}$ 是一个实对称矩阵。根据特征值分解，

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'$$

  注意到，

  - $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$，其中特征值分别为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$.
  - $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n)$，其中 $\boldsymbol{v}_i$ 是特征值 λ_i 所对应的特征向量，$i = 1, 2, \cdots, n$。

- A 的行列式等于其特征值的乘积，即

$$|\boldsymbol{A}| = \prod_{i=1}^{n} \lambda_i$$

特征值判定法

概述

- 这里仅仅考虑 $\boldsymbol{X}$ 是经过标准化后的。
- 根据线性代数的知识可知，行列式 $|\boldsymbol{X}'\boldsymbol{X}| \approx 0$ 时，矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 至少存在一个特征值近似为零。
- 反之，当矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 至少存在一个特征值近似为零时，$\boldsymbol{X}$ 的列向量间必然存在多重共线性。

特征值判定法

具体来说

- 假定 λ 是矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 的一个近似为零的特征值，即

$$\lambda \approx 0$$

- 而 $\boldsymbol{v} = (v_1, \cdots, v_p)'$ 是特征值 λ 所对应的单位特征向量，则

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{v} = \lambda\boldsymbol{v} \approx 0$$

- 于是，在上式中等式两端都左乘 $\boldsymbol{v}'$，可得

$$\boldsymbol{v}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{v} \approx 0 \Rightarrow (\boldsymbol{X}\boldsymbol{v})'\boldsymbol{X}\boldsymbol{v} \approx 0 \Rightarrow \boldsymbol{v}'\boldsymbol{X} \approx 0$$

- 这与多重共线性的定义是一致的。

特征值判定法

判定方法

- 假设 $\boldsymbol{X}'\boldsymbol{X}$ 的特征值分别为 $\lambda_1 \geq \cdots \geq \lambda_p$。
- 称

$$\kappa_j = \sqrt{\frac{\lambda_1}{\lambda_j}}, j = 1, 2, \cdots, p$$

为特征值 λ_j 的条件数（condition index）

- 基本想法：
  - 如果设计矩阵 $\boldsymbol{X}$ 没有多重共线性，即最小特征值 λ_p 不会接近零，那么条件数 κ_p 不会特别大；
  - 设计矩阵 $\boldsymbol{X}$ 存在多重共线性，即最小特征值 λ_p 接近零，那么条件数 κ_p 会特别大。

特征值判定法

判定方法

- 常用的判断标准
  - $0 < \kappa_p < c_\kappa$ 时，设计矩阵 $\boldsymbol{X}$ 没有多重共性性；
  - $\kappa_p \geq c_\kappa$ 时，设计矩阵 $\boldsymbol{X}$ 存在多重共线性；
- 临界值 c_κ 的常见取值有 $10, 100, 1000$。

直观判定法

总结

- 量化标准
  - 方差扩大因子
  - 条件数

- 这种量化标准并不是识别多重共线性的绝对标准，还应该结合一些直观方法综合识别多重共线性。

直观判定法

判定方法

- 当**增加或剔除**一个自变量，其他自变量的回归系数的估计值或显著性发生**较大**变化时，我们就认为回归方程存在严重的多重共线性；

- 当定性分析认为一些**重要**的自变量在回归方程中从**没有通过显著性检验**时，可初步判断存在严重的多重共线性；

- 当与因变量之间的简单相关系绝对值数**很大**的自变量在回归方程中数**没有通过显著性检验**时，可初步判断存在严重的多重共线性；

直观判定法

判定方法

- 当有些自变量的回归系数的数值大小与预期相差数**很大**，甚至正负号与定性分析结果数**相反**时，存在严重的多重共线性问题；
- 在自变量的相关矩阵中，当自变量间的相关系数数**较大**时会出现多重共线性的问题；
- 当一些重要的自变量的回归系数的标准误差数**较大**时，我们认为可能存在多重共线性。

消除多重共线性的方法

三种常见方法

- 删除一些不重要的自变量;
- 增加样本量;
- 改进经典的最小二乘估计（岭回归、主成分回归）。

假定

- 在讨论岭回归和主成分回归时，假定设计矩阵 $\boldsymbol{X}$ 是经过标准化，而因变量 y 未经过标准化。

目录

岭回归

原因

- 线性回归模型的最小二乘估计为

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

  其方差为

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

- 当设计矩阵 $\boldsymbol{X}$ 出现多重共线性时，回归系数的最小二乘估计的效果明显变差.

- 原因是

$$|\boldsymbol{X}'\boldsymbol{X}| \approx 0$$

  这导致了 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 计算不稳定.

岭回归

定义

- 为了求解逆矩阵更方便，我们采用

$$\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}, \quad k > 0$$

  代替 $\boldsymbol{X}'\boldsymbol{X}$.

- 优点：
  - k 很小时，$\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I} \approx \boldsymbol{X}'\boldsymbol{X}$
  - $\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}$ 可以避免是奇异矩阵.

岭回归

定义

- 称

$$\hat{\boldsymbol{\beta}}(k) = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

为回归系数 $\boldsymbol{\beta}$ 的岭回归估计，其中，称 k 为岭参数.

- 注意:
  - 由于 $\boldsymbol{X}$ 已经标准化，所以 $\boldsymbol{X}'\boldsymbol{X}$ 就是自变量样本相关系数矩阵.
  - 因为岭参数 k 不唯一确定，所以岭回归估计

$$\hat{\boldsymbol{\beta}}(k) = (\hat{\beta}_1(k), \hat{\beta}_2(k), \cdots, \hat{\beta}_p(k))'$$

是关于回归参数 $\boldsymbol{\beta}$ 的一个估计族.

岭回归

性质 1

- 为什么说岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 是有偏估计吗?

- 我们计算 $\hat{\boldsymbol{\beta}}(k)$ 的期望，即

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}(k)) &= E((\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\ &= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'E(\boldsymbol{y}) \\ &= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \neq \boldsymbol{\beta} \end{aligned}$$

定理 3-1

$\hat{\boldsymbol{\beta}}(k)$ 是回归参数 $\boldsymbol{\beta}$ 的有偏估计.

岭回归

性质 1

- 为什么说岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 是有偏估计吗?
- 我们计算 $\hat{\boldsymbol{\beta}}(k)$ 的期望，即

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}(k)) &= E((\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\ &= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'E(\boldsymbol{y}) \\ &= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \neq \boldsymbol{\beta} \end{aligned}$$

定理 3-1

$\hat{\boldsymbol{\beta}}(k)$ 是回归参数 $\boldsymbol{\beta}$ 的有偏估计.

岭回归

性质 1

- 为什么说岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 是有偏估计吗?
- 我们计算 $\hat{\boldsymbol{\beta}}(k)$ 的期望，即

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}(k)) &= E((\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\ &= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'E(\boldsymbol{y}) \\ &= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \neq \boldsymbol{\beta} \end{aligned}$$

定理 3-1

$\hat{\boldsymbol{\beta}}(k)$ 是回归参数 $\boldsymbol{\beta}$ 的有偏估计.

岭回归

性质 2

- 岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 与最小二乘估计 $\hat{\boldsymbol{\beta}}$ 有什么关系?

- 根据岭回归估计的定义可知，我们可以得到

$$\begin{aligned}\hat{\boldsymbol{\beta}}(k) &= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}\\ &= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\\ &= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}}\end{aligned}$$

- 如果岭参数 k 是与因变量 y 无关，$\hat{\boldsymbol{\beta}}(k)=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 是 $\hat{\boldsymbol{\beta}}$ 的一种线性变换;

- 岭回归估计 $\hat{\boldsymbol{\beta}}(k)=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 是 $\boldsymbol{y}$ 的线性函数.

岭回归

性质 2

- 岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 与最小二乘估计 $\hat{\boldsymbol{\beta}}$ 有什么关系?
- 根据岭回归估计的定义可知，我们可以得到

$$\begin{aligned}\hat{\boldsymbol{\beta}}(k) &= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}\\ &= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\\ &= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}}\end{aligned}$$

- 如果岭参数 k 是与因变量 y 无关,$\hat{\boldsymbol{\beta}}(k)=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 是 $\hat{\boldsymbol{\beta}}$ 的一种线性变换;
- 岭回归估计 $\hat{\boldsymbol{\beta}}(k)=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 是 $\boldsymbol{y}$ 的线性函数.

岭回归

性质 2

$$\begin{aligned}\hat{\boldsymbol{\beta}}(k) &= (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}}\end{aligned}$$

- 但是，在实际中，岭参数 k 也是由数据确定的，因而 k 也是依赖于因变量 $\boldsymbol{y}$.
- 因此，从本质上来说，$\hat{\boldsymbol{\beta}}(k)$ 并非 $\hat{\boldsymbol{\beta}}$ 的线性变换，也不是 $\boldsymbol{y}$ 的线性函数.

岭回归

性质 3

- 除了解决设计矩阵中的多重共线性之外，岭回归估计还有什么优势吗?
- 或者说，岭回归估计能否比最小二乘估计更优?
- 由于岭回归估计是有偏的，一般我们以**均方误差**作为标准来比较岭回归估计和最小二乘估计.
- 注意到，最小二乘估计可以看作一种特殊的岭回归估计，即

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + 0 \cdot \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} = \hat{\boldsymbol{\beta}}(0)$$

岭回归

性质 3

- 除了解决设计矩阵中的多重共线性之外，岭回归估计还有什么优势吗?
- 或者说，岭回归估计能否比最小二乘估计更优?
- 由于岭回归估计是有偏的，一般我们以**均方误差**作为标准来比较岭回归估计和最小二乘估计.
- 注意到，最小二乘估计可以看作一种特殊的岭回归估计，即

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + 0 \cdot \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} = \hat{\boldsymbol{\beta}}(0)$$

性质 3

定理 3-2

存在 $k > 0$，使得岭估计的均方误差小于最小二乘估计的均方误差，即

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}(k)) \leq \mathrm{MSE}(\hat{\boldsymbol{\beta}}(0))$$

证明： 为了简化符号，我们令 $H(k) = \mathrm{MSE}(\hat{\boldsymbol{\beta}}(k))$，即

$$\begin{aligned} H(k) &= \mathrm{MSE}(\hat{\boldsymbol{\beta}}(k)) = E\left(\hat{\boldsymbol{\beta}}(k) - \boldsymbol{\beta}\right)'\left(\hat{\boldsymbol{\beta}}(k) - \boldsymbol{\beta}\right) \\ &= E\left(\hat{\boldsymbol{\beta}}(k) - E(\hat{\boldsymbol{\beta}}(k))\right)'\left(\hat{\boldsymbol{\beta}}(k) - E(\hat{\boldsymbol{\beta}}(k))\right) \\ &\quad + (E(\hat{\boldsymbol{\beta}}(k)) - \boldsymbol{\beta})'(E(\hat{\boldsymbol{\beta}}(k)) - \boldsymbol{\beta}) \\ &=: I_1(k) + I_2(k) \end{aligned}$$

性质 3

为了证明存在 $k > 0$ 使得

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}(k)) \leq \mathrm{MSE}(\hat{\boldsymbol{\beta}}(0)),$$

只需要证明 $H(k)$ 在 $k = 0$ 处的导数 $\frac{\partial H(k)}{\partial k}|_{k=0} < 0$ 即可. 根据 $H(k) = I_1(k) + I_2(k)$ 可知，

$$\frac{\partial H(k)}{\partial k} = \frac{\partial I_1(k)}{\partial k} + \frac{\partial I_2(k)}{\partial k}.$$

于是，我们进一步分析这两个导数.

岭回归

性质 3

我们先讨论一些岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 不同的表达形式

$$\begin{aligned}\hat{\boldsymbol{\beta}}(k) &= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} \overset{\text{def}}{=} \boldsymbol{W}_k\boldsymbol{X}'\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}} \overset{\text{def}}{=} \boldsymbol{W}_k^*\hat{\boldsymbol{\beta}}\end{aligned}$$

于是，$\boldsymbol{W}_k^*$ 与 $\boldsymbol{W}_k$ 之间存在关系，即

$$\begin{aligned}\boldsymbol{W}_k^* &= \boldsymbol{W}_k(\boldsymbol{X}'\boldsymbol{X}) \\ &= (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}-k\boldsymbol{I}) \\ &= \boldsymbol{I}-k(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1} \\ &= \boldsymbol{I}-k\boldsymbol{W}_k\end{aligned}$$

性质 3

假定 $\boldsymbol{X}'\boldsymbol{X}$ 的特征值为

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0,$$

而相应正交化后的特征向量记为 $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_p$，则有

$$(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j, \quad j = 1, 2, \cdots, p$$

在上式的等式两端同时加上 $k\boldsymbol{I} \cdot \boldsymbol{v}_j$，

$$(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})\,\boldsymbol{v}_j = (\boldsymbol{X}'\boldsymbol{X})\boldsymbol{v}_j + k\boldsymbol{I}\cdot\boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j + k\boldsymbol{I}\cdot\boldsymbol{v}_j = (\lambda_j + k)\boldsymbol{v}_j$$

那么，

$$(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{v}_j = \frac{1}{\lambda_j + k}\boldsymbol{v}_j \Rightarrow \boldsymbol{W}_k\boldsymbol{v}_j = \frac{1}{\lambda_j + k}\boldsymbol{v}_j$$

性质 3

另一方面，根据

$$(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j, \quad j = 1, 2, \cdots, p$$

可知，

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{v}_j = \frac{1}{\lambda_j}\boldsymbol{v}_j.$$

于是，

$$(\boldsymbol{I} + k(\boldsymbol{X}'\boldsymbol{X})^{-1})\boldsymbol{v}_j = \left(1 + \frac{k}{\lambda_j}\right)\boldsymbol{v}_j = \frac{\lambda_j + k}{\lambda_j}\boldsymbol{v}_j.$$

从而

$$\left(\boldsymbol{I} + k(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)^{-1}\boldsymbol{v}_j = \frac{\lambda_j}{\lambda_j + k}\boldsymbol{v}_j \Rightarrow \boldsymbol{W}_k^*\boldsymbol{v}_j = \frac{\lambda_j}{\lambda_j + k}\boldsymbol{v}_j$$

性质 3

这里我们总结一下:

- $\boldsymbol{W}_k$ 的特征值分别为 $\frac{1}{\lambda_1+k}, \cdots, \frac{1}{\lambda_p+k}$;
- $\boldsymbol{W}_k^*$ 的特征值分别为 $\frac{\lambda_1}{\lambda_1+k}, \cdots, \frac{\lambda_p}{\lambda_p+k}$;
- $\boldsymbol{W}_k$ 和 $\boldsymbol{W}_k^*$ 的特征向量与 $\boldsymbol{X}'\boldsymbol{X}$ 的特征向量相同，与岭参数 k 无关.

性质 3

首先，我们考虑 $I_1(k)$.

$$\begin{aligned} I_1(k) &= E\left(\hat{\boldsymbol{\beta}}(k) - E(\hat{\boldsymbol{\beta}}(k))\right)' \left(\hat{\boldsymbol{\beta}}(k) - E(\hat{\boldsymbol{\beta}}(k))\right) \\ &= E(\boldsymbol{W}_k^*\hat{\boldsymbol{\beta}} - \boldsymbol{W}_k^*\boldsymbol{\beta})'(\boldsymbol{W}_k^*\hat{\boldsymbol{\beta}} - \boldsymbol{W}_k^*\boldsymbol{\beta}) \\ &= E((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\boldsymbol{W}_k^*)'(\boldsymbol{W}_k^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) \\ &= E(\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{W}_k^*)'(\boldsymbol{W}_k^*)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}) \end{aligned}$$

最后一个等号成立，是因为

$$\begin{aligned} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} \\ &= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} \end{aligned}$$

统计知识（补充）

- 假定 $\boldsymbol{A}$ 是对称矩阵. $\boldsymbol{x}$ 是一个 p 维随机变量，并假定 $\boldsymbol{\mu} = E(\boldsymbol{x})$ 和 $\Sigma = \mathrm{Var}(\boldsymbol{x})$. 那么，

$$E(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = \mathrm{tr}(\boldsymbol{A}\Sigma) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}.$$

性质 3

$$
\begin{aligned}
I_1(k) &= E(\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{W}_k^*)'(\boldsymbol{W}_k^*)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}) \\
&= \sigma^2\mathrm{tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{W}_k^*)'(\boldsymbol{W}_k^*)) \\
&= \sigma^2\mathrm{tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{W}_k(\boldsymbol{I}-k\boldsymbol{W}_k)) \\
&= \sigma^2\left(\mathrm{tr}(\boldsymbol{W}_k)-k\mathrm{tr}(\boldsymbol{W}_k^2)\right) \\
&= \sigma^2\left(\sum_{j=1}^{p}\frac{1}{\lambda_j+k}-k\sum_{j=1}^{p}\frac{1}{(\lambda_j+k)^2}\right) \\
&= \sigma^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j+k)^2} \\
\Rightarrow \frac{\partial I_1(k)}{\partial k} &= -2\sigma^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j+k)^3} < 0 \Rightarrow k \text{ 越大 } I_1(k) \text{ 越小}
\end{aligned}
$$

性质 3

接下来，我们考虑 $I_2(k)$.

$$
\begin{aligned}
I_2(k) &= \left(E(\hat{\boldsymbol{\beta}}(k)) - \boldsymbol{\beta}\right)' \left(E(\hat{\boldsymbol{\beta}}(k)) - \boldsymbol{\beta}\right) \\
&= (\boldsymbol{W}_k^* \boldsymbol{\beta} - \boldsymbol{\beta})' (\boldsymbol{W}_k^* \boldsymbol{\beta} - \boldsymbol{\beta}) \\
&= \boldsymbol{\beta}'(\boldsymbol{W}_k^* - \boldsymbol{I})'(\boldsymbol{W}_k^* - \boldsymbol{I})\boldsymbol{\beta} \\
&= k^2 \boldsymbol{\beta}' \boldsymbol{W}_k^2 \boldsymbol{\beta} \\
&= k^2 \boldsymbol{\beta}' \boldsymbol{V}' \boldsymbol{L} \boldsymbol{V} \boldsymbol{\beta} \qquad \text{令} (\boldsymbol{W}_k^2 = \boldsymbol{V}' \boldsymbol{L} \boldsymbol{V}) \\
&=: k^2 \boldsymbol{\alpha}' \boldsymbol{L} \boldsymbol{\alpha} \qquad \text{令} (\boldsymbol{\alpha} = \boldsymbol{V} \boldsymbol{\beta}) \\
&= k^2 \sum_{j=1}^{p} \frac{\alpha_j^2}{(\lambda_j + k)^2}
\end{aligned}
$$

其中，$\boldsymbol{\alpha} = \boldsymbol{V}\boldsymbol{\beta} = (\alpha_1, \alpha_2, \cdots, \alpha_p)'$ 与岭参数 k 无关.

性质 3

由于

$$I_2(k) = k^2 \sum_{j=1}^{p} \frac{\alpha_j^2}{(\lambda_j + k)^2}$$

因此，

$$\begin{aligned} \frac{\partial I_2(k)}{\partial k} &= 2\sum_{j=1}^{p} \frac{k\alpha_j^2}{(\lambda_j + k)^2} - 2\sum_{j=1}^{p} \frac{k^2\alpha_j^2}{(\lambda_j + k)^3} \\ &= 2k\sum_{j=1}^{p} \frac{\lambda_j\alpha_j^2}{(\lambda_j + k)^3} \geq 0 \end{aligned}$$

即当 k 越大时，$I_2(k)$ 越大.

性质 3

$$\begin{aligned}\frac{\partial H(k)}{\partial k} &= \frac{\partial I_1(k)}{\partial k} + \frac{\partial I_2(k)}{\partial k} \\ &= -2\sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^3} + 2k \sum_{j=1}^{p} \frac{\lambda_j \alpha_j^2}{(\lambda_j + k)^3}\end{aligned}$$

考虑 $k = 0$ 时，

$$\left.\frac{\partial H(k)}{\partial k}\right|_{k=0} = \left.\frac{\partial I_1(k)}{\partial k}\right|_{k=0} + \left.\frac{\partial I_2(k)}{\partial k}\right|_{k=0} = -2\sigma^2 \sum_{j=1}^{p} \lambda_j^{-2} < 0$$

由连续性可知，在以 0 为中心的一个领域内，存在 $k > 0$，使得 $H(k) < H(0)$. 此定理证毕.

说明

- 注意到

$$\begin{aligned}\frac{\partial H(k)}{\partial k} &= \frac{\partial I_1(k)}{\partial k} + \frac{\partial I_2(k)}{\partial k} \\ &= -2\sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^3} + 2k \sum_{j=1}^{p} \frac{\lambda_j \alpha_j^2}{(\lambda_j + k)^3} \\ &= \sum_{j=1}^{m} \frac{2\lambda_j}{(\lambda_j + k)^3} (k\alpha_j^2 - \sigma^2)\end{aligned}$$

- 根据上式，易知使得 $\frac{\partial H(k)}{\partial k} = 0$ 的 k 与 $\sigma^2, \boldsymbol{\beta}$ 有关；
- 但是，σ^2 与 β 均为未知参数，因此无法找到一个对一切 σ^2 及 $\boldsymbol{\beta}$ 都成立的 k 使得 $\boldsymbol{H}(k)$ 达到最小.

岭回归

另一个角度看岭回归估计

- 由于最小二乘估计 $\hat{\boldsymbol{\beta}}$ 是最小化离差平方和的解，即

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

- 可以证明，岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 是最小化带有 L_2 正则项的离差平方和的解，即

$$\hat{\boldsymbol{\beta}}(k) = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}'\boldsymbol{\beta}$$

- 等价于最小化

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \quad \text{s.t.} \quad \boldsymbol{\beta}'\boldsymbol{\beta} \leq s$$

岭回归

另一个角度看岭回归估计

- 考虑带约束的最优化问题:

$$\min(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \quad \text{s.t.} \quad \boldsymbol{\beta}'\boldsymbol{\beta} \leq s$$

这个约束会对解 $\hat{\boldsymbol{\beta}}(k)$ 带来什么影响?

- 如果最小二乘估计 $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} \leq s$，那么 $\hat{\boldsymbol{\beta}}$ 就是我们想要的解，也就是说 $\hat{\boldsymbol{\beta}}(k) = \hat{\boldsymbol{\beta}}$;
- 如果最小二乘估计 $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} > s$，那么，我们所得到的解 $\hat{\boldsymbol{\beta}}(k)$ 应该满足

$$\hat{\boldsymbol{\beta}}(k)'\hat{\boldsymbol{\beta}}(k) \leq s < \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$$

岭回归

另一个角度看岭回归估计

- 考虑带约束的最优化问题:

$$\min(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \quad \text{s.t.} \quad \boldsymbol{\beta}'\boldsymbol{\beta} \leq s$$

  这个约束会对解 $\hat{\boldsymbol{\beta}}(k)$ 带来什么影响?

- 如果最小二乘估计 $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} \leq s$，那么 $\hat{\boldsymbol{\beta}}$ 就是我们想要的解，也就是说 $\hat{\boldsymbol{\beta}}(k) = \hat{\boldsymbol{\beta}}$;

- 如果最小二乘估计 $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} > s$，那么，我们所得到的解 $\hat{\boldsymbol{\beta}}(k)$ 应该满足

$$\hat{\boldsymbol{\beta}}(k)'\hat{\boldsymbol{\beta}}(k) \leq s < \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$$

岭回归

另一个角度看岭回归估计

- 考虑带约束的最优化问题:

$$\min(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \quad \text{s.t.} \quad \boldsymbol{\beta}'\boldsymbol{\beta} \leq s$$

这个约束会对解 $\hat{\boldsymbol{\beta}}(k)$ 带来什么影响?

- 如果最小二乘估计 $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} \leq s$，那么 $\hat{\boldsymbol{\beta}}$ 就是我们想要的解，也就是说 $\hat{\boldsymbol{\beta}}(k) = \hat{\boldsymbol{\beta}}$;

- 如果最小二乘估计 $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} > s$，那么，我们所得到的解 $\hat{\boldsymbol{\beta}}(k)$ 应该满足

$$\hat{\boldsymbol{\beta}}(k)'\hat{\boldsymbol{\beta}}(k) \leq s < \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$$

性质 4

定理 3-3

对任意 $k > 0$，$\|\hat{\boldsymbol{\beta}}\| \neq 0$，总有

$$\|\hat{\boldsymbol{\beta}}(k)\| < \|\hat{\boldsymbol{\beta}}\|$$

证明：假定 $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p$ 是 $\boldsymbol{X}'\boldsymbol{X}$ 的特征值，而 $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_p$ 为其相应的特征向量. 于是，我们有

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{V}'\boldsymbol{\Lambda}\boldsymbol{V}$$

其中，$\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_p\}$，$\boldsymbol{V}'$ 是以 $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_p$ 为列向量的矩阵.

岭回归

性质 4

回归模型可写为

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \boldsymbol{X}\boldsymbol{V}'\boldsymbol{V}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &=: \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \end{aligned}$$

注意到，$\boldsymbol{\alpha} = \boldsymbol{V}\boldsymbol{\beta} \Rightarrow \boldsymbol{\beta} = \boldsymbol{V}'\boldsymbol{\alpha}$.

由此，$\boldsymbol{\alpha}$ 的最小二乘估计为

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y} = (\boldsymbol{V}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{V}')^{-1}\boldsymbol{Z}'\boldsymbol{y} \\ &= (\boldsymbol{V}\boldsymbol{V}'\boldsymbol{\Lambda}\boldsymbol{V}\boldsymbol{V}')^{-1}\boldsymbol{Z}'\boldsymbol{y} = \boldsymbol{\Lambda}^{-1}\boldsymbol{Z}'\boldsymbol{y} \end{aligned}$$

性质 4

而 $\boldsymbol{\beta}$ 的最小二乘估计 $\hat{\boldsymbol{\beta}}$ 与 $\hat{\boldsymbol{\alpha}}$ 存在如下关系

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{V}'\boldsymbol{\Lambda}^{-1}\boldsymbol{V}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{V}'\boldsymbol{\alpha}$$

类似地，关于 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$ 的岭估计分别为

$$\begin{aligned} \hat{\boldsymbol{\alpha}}(k) &= (\boldsymbol{\Lambda} + k\boldsymbol{I})^{-1}\boldsymbol{Z}'\boldsymbol{y} \\ \hat{\boldsymbol{\beta}}(k) &= \boldsymbol{V}'\hat{\boldsymbol{\alpha}}(k) \end{aligned}$$

所以，

$$\|\hat{\boldsymbol{\beta}}(k)\| = \|\hat{\boldsymbol{\alpha}}(k)\| = \|(\boldsymbol{\Lambda} + k\boldsymbol{I})^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\alpha}}\| < \|\hat{\boldsymbol{\alpha}}\| = \|\hat{\boldsymbol{\beta}}\|$$

由此，定理得证.

说明

- $\hat{\boldsymbol{\beta}}(k)$ 是对 $\hat{\boldsymbol{\beta}}$ 向原点的压缩.
- 这是因为

$$\begin{aligned} \mathrm{MSE}(\hat{\boldsymbol{\beta}}) &= E\left((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right) \\ &= E(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}'\boldsymbol{\beta} = E\|\hat{\boldsymbol{\beta}}\|^2 - \|\boldsymbol{\beta}\|^2 \end{aligned}$$

- 因此，

$$E\|\hat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{\beta}\|^2 + \mathrm{MSE}(\hat{\boldsymbol{\beta}}) = \|\boldsymbol{\beta}\|^2 + \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}$$

- 当设计矩阵 $\boldsymbol{X}$ 出现多重共线性时，上式中的第二项比较大，因此，对其做压缩是应该的.

岭回归

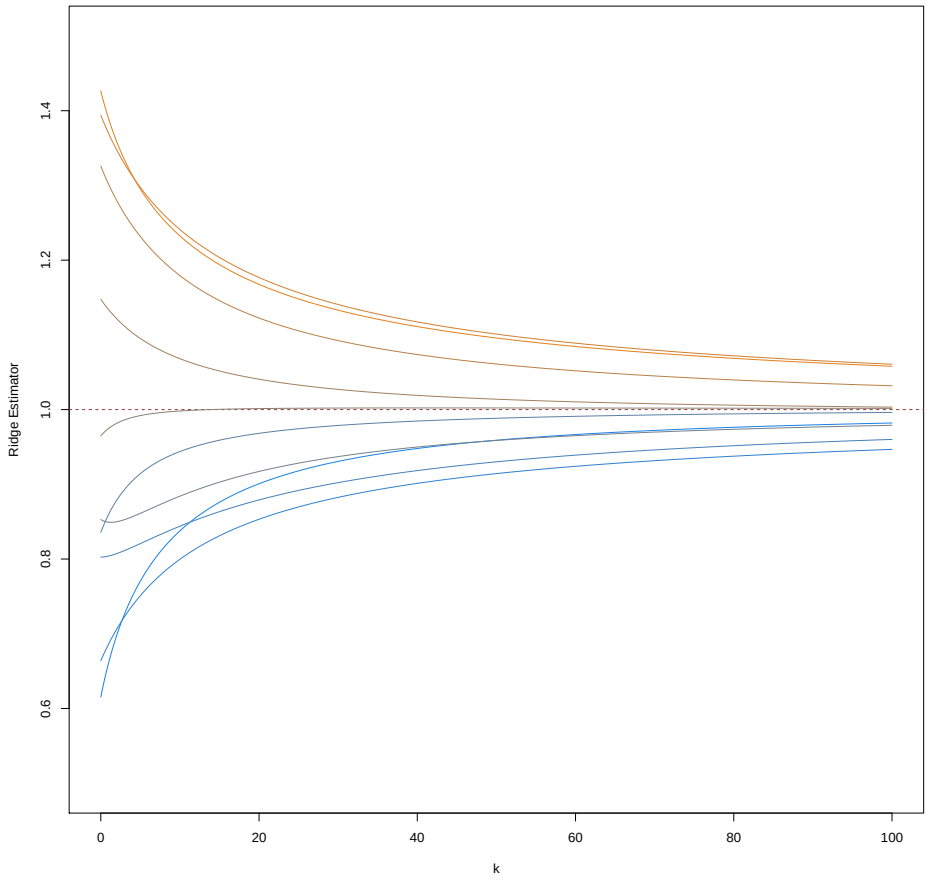
岭参数的参数选择（岭迹法）

- 岭估计 $\hat{\boldsymbol{\beta}}(k) = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 的分量 $\hat{\beta}_j(k)$ 作为岭参数 k 的函数.

- 当 k 在 $[0, +\infty)$ 变化时，在平面直角坐标系中，我们称 $k - \hat{\beta}_j(k)$ 的图像为**岭迹**.

岭回归

岭参数的参数选择（岭迹法）



岭回归

岭参数的参数选择（岭迹法）

- 岭迹法的一般原则
  - 各回归系数的岭估计基本稳定;
  - 用最小二乘估计时，符号不合理的回归系数的岭估计的符号变得合理;
  - 回归系数没有不合理的符号;
  - 残差平方和增大不多.
- 优点：容易计算;
- 缺点：具有主观性;

岭回归

岭参数的参数选择（方差扩大因子法）

- 根据方差扩大因子判定多重共线性，即 $c_{jj} > c_{\text{VIF}}$.
- 岭回归估计 $\hat{\boldsymbol{\beta}}(k)$ 的方差为

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}(k)) &= \sigma^2(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1} \\ &\stackrel{\text{def}}{=} \sigma^2\boldsymbol{C}(k) \end{aligned}$$

- 我们可以类似地定义矩阵 $\boldsymbol{C}(k)$ 中对角线的元素 $c_{jj}(k)$ 为岭估计的方差扩大因子.
- $c_{jj}(k)$ 随着 k 的增大而减少.
- 通过选择 k 使得所有方差扩大因子 $c_{jj}(k) \leq c_{\text{VIF}}$，从而确定岭参数 k.

岭参数的参数选择（Hoerl-Kennad 公式）

- 回顾

$$\begin{aligned}\frac{\partial H(k)}{\partial k} &= \frac{\partial I_1(k)}{\partial k} + \frac{\partial I_2(k)}{\partial k} \\ &= \sum_{j=1}^{m} \frac{2\lambda_j}{(\lambda_j + k)^3}(k\alpha_j^2 - \sigma^2)\end{aligned}$$

- 在 1970 年，霍尔和肯纳德提出了

$$k_{\mathrm{HK}} = \frac{\hat{\sigma}^2}{\max_j \hat{\alpha}_j^2}$$

- 易证 $\frac{\partial H(k)}{\partial k}\big|_{k=k_{\mathrm{HK}}} < 0.$

岭回归

岭参数的参数选择（Mcdorard-Garaneau 公式）

- 回顾

$$E\|\hat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{\beta}\|^2 + \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}$$

- 令

$$Q = \|\hat{\boldsymbol{\beta}}\|^2 - \hat{\sigma}^2 \sum_{j=1}^{p} \lambda_j^{-1}$$

- 如果 $Q > 0$，那么认为 $\hat{\boldsymbol{\beta}}$ 中某一分量过大，需要对其进行压缩. 压缩量由 $\sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}$ 决定.
- 如果 $Q \leq 0$，那么认为 $\hat{\boldsymbol{\beta}}$ 的各个分量都差不多，此时，对 $\hat{\boldsymbol{\beta}}$ 不进行压缩，选择 $k = 0$.

岭回归

岭参数的参数选择（Mcdorard-Garaneau 公式）

- Mcdorard 和 Garaneau 建议选择岭参数 k，使得

$$\|\hat{\boldsymbol{\beta}}\|^2 - \|\hat{\boldsymbol{\beta}}(k)\|^2 \approx \hat{\sigma}^2 \sum_{j=1}^{p} \lambda_j^{-1}$$

即选择 k 使得

$$\|\hat{\boldsymbol{\beta}}(k)\|^2 \approx \|\hat{\boldsymbol{\beta}}\|^2 - \hat{\sigma}^2 \sum_{j=1}^{p} \lambda_j^{-1}$$

统计与机器学习

第三章：多重共线性 - Part II

倪 葎

DaSE@ECNU
(lni@dase.ecnu.edu.cn)

2021 年 3 月 24 日

目录

主成分回归

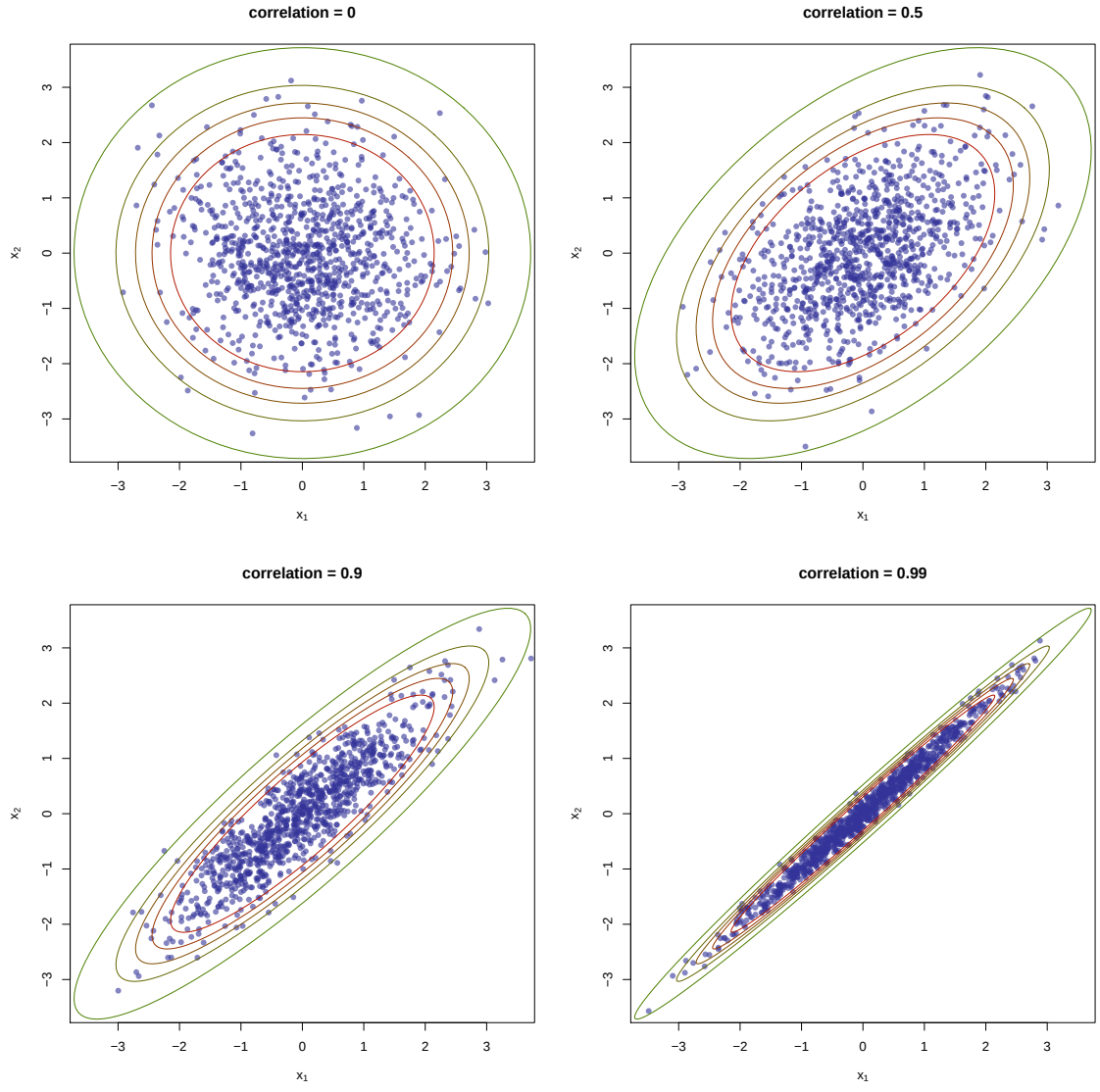
基本思想

主成分回归 $=$ **主成分**分析 $+$ **回归**分析

- 回归分析: 研究因变量 y 与自变量 $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)'$ 之间的关系，即
$$y = \boldsymbol{x}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$
- 主成分分析: 用 k 个自变量的线性变换（主成分）代替原本的 p 个自变量（$k < p$）;
- 关键问题:
  - 如何求主成分?
  - 如何利用主成分来估计回归参数 $\boldsymbol{\beta}$?
  - 主成分回归估计 $\hat{\boldsymbol{\beta}}_{\mathrm{PC}}$ 与最小二乘估计 $\hat{\boldsymbol{\beta}}$ 有什么关系?

主成分分析

动机


correlation = 0
correlation = 0.5
correlation = 0.9
correlation = 0.99
x1
x2

主成分分析

动机

- 由于自变量个数太多，往往自变量之间存在着一定的相关性，因而使得所观测到的数据在一定程度上反映的信息有所重叠。

- 当自变量较多时，我们自然想到能用**较少**的综合变量代替原本多个自变量，而这几个综合变量有能够**尽可能多**地反映原本变量的信息，并且彼此之间**互不相关**。

主成分分析

基本想法

- 设 $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)'$ 是 p 维随机向量
  - 均值 $E(\boldsymbol{x}) = \boldsymbol{\mu}$;
  - 方差-协方差矩阵 $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma}$.
- 考虑 $\boldsymbol{x}$ 的线性变换，即

$$\begin{cases} z_1 = \boldsymbol{a}_1'\boldsymbol{x} = a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p \\ z_2 = \boldsymbol{a}_2'\boldsymbol{x} = a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p \\ \vdots \\ z_p = \boldsymbol{a}_p'\boldsymbol{x} = a_{1p}x_1 + a_{2p}x_2 + \cdots + a_{pp}x_p \end{cases}$$

- 易知

$$\mathrm{Var}(z_i) = \boldsymbol{a}_i\boldsymbol{\Sigma}\boldsymbol{a}_i, \quad \mathrm{Cov}(z_i, z_j) = \boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_j$$

主成分分析

基本想法

- 如果我们希望用 z_1 来代替原来的 p 个变量 $x_1, x_2, \cdots, x_p$，那么这个“新”变量 z_1 需要满足
  - z_1 能够尽可能多地反映原来 p 个变量的信息。
- 问题：如何度量这个“信息”？
- 在统计学中，常常用“方差”度量“信息”。
- 自然想法：选取新变量 z_1，如果 $\mathrm{Var}(z_1)$ 越大，那么 z_1 包含的信息越多。

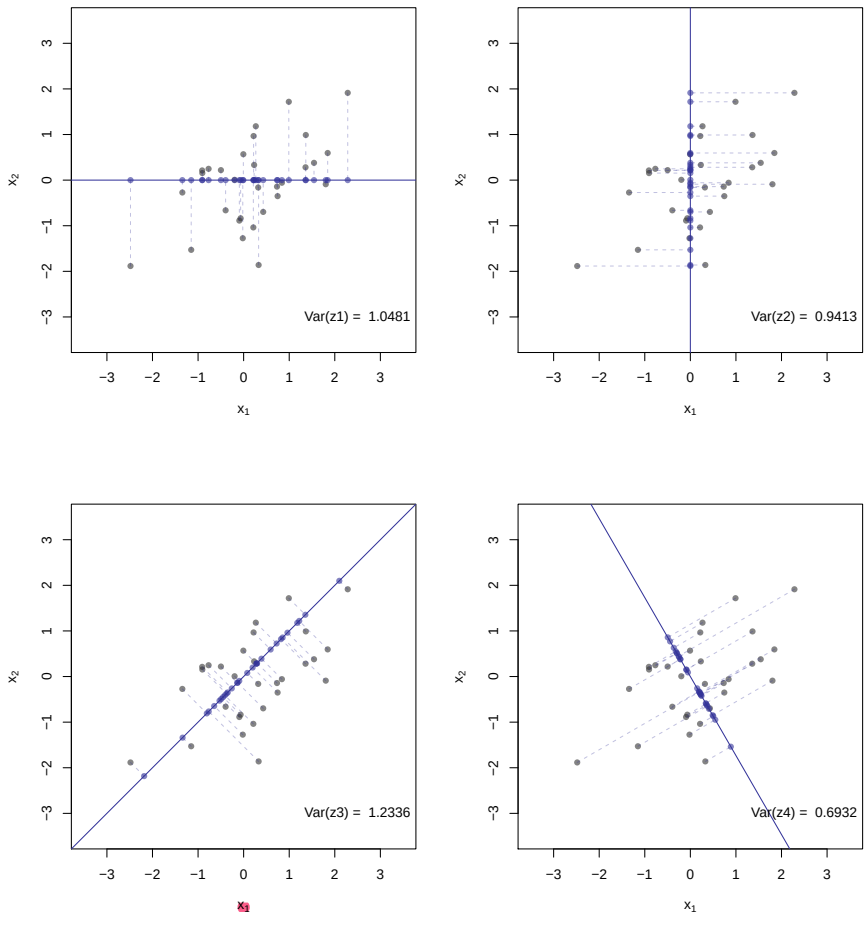
主成分分析

基本想法

- 如果我们希望用 z_1 来代替原来的 p 个变量 $x_1, x_2, \cdots, x_p$，那么这个“新”变量 z_1 需要满足
  - z_1 能够尽可能多地反映原来 p 个变量的信息。
- 问题：如何度量这个“信息”？
- 在统计学中，常常用“方差”度量“信息”。
- 自然想法：选取新变量 z_1，如果 $\mathrm{Var}(z_1)$ 越大，那么 z_1 包含的信息越多。

主成分分析

动机



主成分分析

基本想法

- 如果我们希望用 z_1 来代替原来的 p 个变量 $x_1, x_2, \cdots, x_p$，那么这个“新”变量 z_1 需要满足
  - z_1 能够尽可能多地反映原来 p 个变量的信息。
- 问题：如何度量这个“信息”?
- 在统计学中，常常用“方差”度量“信息”。
- 自然想法：选取新变量 z_1，如果 $\mathrm{Var}(z_1)$ 越大，那么 z_1 包含的信息越多。
- 由于 $\mathrm{Var}(z_1) = \boldsymbol{a}_1' \boldsymbol{\Sigma} \boldsymbol{a}_1$, 因此，我们需要对 $\boldsymbol{a}_1$ 做出一些限制。最常用的限制是：$\boldsymbol{a}_1' \boldsymbol{a}_1 = 1$.
- 若存在满足以上约束 $\boldsymbol{a}_1$，使得 $\mathrm{Var}(z_1)$ 达到最大，称 z_1 为**第一主成分**。

主成分分析

基本想法

- 如果第一主成分不足以代表原来的 p 个变量的绝大部分信息，我们需要进一步考虑 $\boldsymbol{x}$ 的第二个线性组合 z_2.
- 为了有效地代表原始变量的信息，z_1 已体现的信息不希望在 z_2 中出现.
- 用统计语言表示，就是要求

$$\mathrm{Cov}(z_2, z_1) = \boldsymbol{a}_2' \boldsymbol{\Sigma} \boldsymbol{a}_1 = 0$$

- 同样，在两个约束 $\boldsymbol{a}_2' \boldsymbol{a}_2 = 1$ 和 $\boldsymbol{a}_2' \boldsymbol{\Sigma} \boldsymbol{a}_1 = 0$ 下，求 a_2 使得 $\mathrm{Var}(z_2)$ 达到最大。
- 类似地，可以求第三主成分、第四主成分、$\cdots$

主成分分析

基本定义

定义（主成分）

设 $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)'$ 为 p 维随机向量且 $\boldsymbol{a}_i = (a_{1i}, a_{2i}, \cdots, a_{pi})'$ 是个 p 维常数向量。如果 $z_i = \boldsymbol{a}_i'\boldsymbol{x}$ 是 $\boldsymbol{x}$ 的线性组合，且满足

- $\boldsymbol{a}_i'\boldsymbol{a}_i = 1$;
- 当 $i > 1$ 时，$\boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_j = 0$;
- $\mathrm{Var}(z_i) = \max\limits_{\boldsymbol{a}'\boldsymbol{a}=1, \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}_j=0, j=1,\cdots,i-1} \mathrm{Var}(\boldsymbol{a}'\boldsymbol{x})$.

那么，称 z_i 为 $\boldsymbol{x}$ 的第 i 个主成分。

主成分分析

主成分的求法

- 设 p 维随机向量 $\boldsymbol{x}$ 的均值 $E(\boldsymbol{x}) = 0$，而方差-协方差矩阵 $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma}$.
- 问题：如何求第一主成分 $z_1 = \boldsymbol{a}_1'\boldsymbol{x}$?
- 求 $\boldsymbol{a}_1 = (a_{11}, a_{21}, \cdots, a_{p1})'$ 满足

$$\boldsymbol{a}_1 = \arg\max_{\boldsymbol{a}} \boldsymbol{a}'\boldsymbol{x} \quad \text{s.t.} \quad \boldsymbol{a}_1'\boldsymbol{a}_1 = 1$$

主成分分析

主成分的求法

- 采用拉格朗日乘子法，令

$$l(\boldsymbol{a}_1) = \mathrm{Var}(\boldsymbol{a}_1'\boldsymbol{x}) - \lambda(\boldsymbol{a}_1'\boldsymbol{a}_1 - 1) = \boldsymbol{a}_1'\boldsymbol{\Sigma}\boldsymbol{a}_1 - \lambda(\boldsymbol{a}_1'\boldsymbol{a}_1 - 1)$$

- 于是有

$$\begin{cases} \dfrac{\partial l}{\partial \boldsymbol{a}_1} = 2(\boldsymbol{\Sigma} - \lambda \boldsymbol{I})\boldsymbol{a}_1 = 0 \\ \dfrac{\partial l}{\partial \lambda} = \boldsymbol{a}_1'\boldsymbol{a}_1 - 1 = 0. \end{cases}$$

- 因为 $\boldsymbol{a}_1 \neq \boldsymbol{0}$，所以，$|\boldsymbol{\Sigma} - \lambda \boldsymbol{I}| = 0$.
- 求第一主成分的问题等价于求 $\boldsymbol{\Sigma}$ 的特征值和特征向量问题。

主成分分析

基本定义

定理 3-4

设 $\boldsymbol{x} = (x_1, \cdots, x_p)'$ 是 p 维随机向量，且 $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma}$ 且满足

- $\boldsymbol{\Sigma}$ 的特征值为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$;
- $\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_p$ 为相应的单位正交特征向量.

则 $\boldsymbol{x}$ 的第 i 主成分为

$$z_i = \boldsymbol{a}_i' \boldsymbol{x}, i = 1, 2, \cdots, p$$

主成分分析

基本定义

- 在实际问题中，不同变量往往有不同的量纲，而通过协方差矩阵 $\boldsymbol{\Sigma}$ 来求主成分总是优先考虑方差大的变量，有时会造成很不合理的结果。
- 为了消除由于量纲的不同可能带来的一些不合理的理想，常采用将变量标准化的方法。
- 记 $E(x_i) = \mu_i$ 和 $\mathrm{Var}(x_i) = \sigma_i^2$，那么对变量 x_i 进行标准化，即

$$x_i^* = \frac{x_i - E(x_i)}{\sqrt{\mathrm{Var}(x_i)}} = \frac{x_i - \mu_i}{\sigma_i}$$

- 随机向量 $\boldsymbol{x}^*$ 的协方差矩阵 $\boldsymbol{\Sigma}^*$ 就是原随机向量 $\boldsymbol{x}$ 的相关阵 $\mathrm{Corr}(\boldsymbol{x})$，因此，由 $\mathrm{Corr}(\boldsymbol{x})$ 来求主成分。

主成分分析

基本定义

假定由 $\mathrm{Corr}(\boldsymbol{x})$ 所确定的主成分为 $\boldsymbol{z}^* = (z_1^*, \cdots, z_p^*)'$, 则 $\boldsymbol{z}^*$ 的性质如下:

- 主成分的协方差阵为

$$\mathrm{Var}(\boldsymbol{z}^*) = \Lambda^* = \mathrm{diag}\{\lambda_1^*, \lambda_2^*, \cdots, \lambda_p^*\},$$

其中，$\lambda_1^* \geq \cdots \geq \lambda_p^*$;

- 特征值满足

$$\sum_{i=1}^{p} \lambda_i^* = p$$

主成分分析

基本定义

- 之前我们讨论了总体的主成分，在实际问题中，一般方差-协方差矩阵 $\boldsymbol{\Sigma}$ 未知，需要通过样本来估计。
- 样本数据（设计矩阵）为

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix}.$$

主成分分析

基本定义

- 协方差矩阵 Σ 的估计为样本协方差 S，即

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})' \stackrel{\text{def}}{=} (s_{kl})_{p\times p}$$

其中

$$\begin{aligned}\bar{\boldsymbol{x}} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \stackrel{\text{def}}{=} (\bar{x}_1, \cdots, \bar{x}_p)' \\ s_{kl} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l)\end{aligned}$$

主成分分析

基本定义

- 样本相关阵 R 为

$$R = (r_{kl})_{p\times p}$$

其中

$$r_{kl} = \frac{s_{kl}}{\sqrt{s_{kk}s_{ll}}}$$

主成分分析

基本定义

- 假定标准化后的矩阵为

$$\boldsymbol{X}^* = \begin{pmatrix} x_{11}^* & x_{12}^* & \cdots & x_{1p}^* \\ x_{21}^* & x_{22}^* & \cdots & x_{2p}^* \\ \vdots & \vdots & & \vdots \\ x_{n1}^* & x_{n2}^* & \cdots & x_{np}^* \end{pmatrix}$$

- x_{ij} 与 x_{ij}^* 之间的关系

$$x_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{l_{jj}}}$$

其中，$l_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 = (n-1)s_{jj}$

主成分分析

基本定义

- 考虑 $(\boldsymbol{X}^*)'\boldsymbol{X}^*$ 中的每一个元素

$$\begin{aligned}\sum_{i=1}^{n}(x_{ik}^*)(x_{il}^*) &= \sum_{i=1}^{n}\frac{x_{ik}-\bar{x}_k}{\sqrt{l_{kk}}}\frac{x_{il}-\bar{x}_l}{\sqrt{l_{ll}}} \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_{ik}-\bar{x}_k)(x_{il}-\bar{x}_l)}{\sqrt{s_{kk}s_{ll}}} \\ &= r_{kl}\end{aligned}$$

- $(\boldsymbol{X}^*)'\boldsymbol{X}^*$ 可以等价于原始设计矩阵 $\boldsymbol{X}$ 的样本相关阵 R。

主成分回归

动机

- 假定设计矩阵 $\boldsymbol{X}$ 已标准化。假设 $\boldsymbol{X}'\boldsymbol{X}$ 的
  - 特征值为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$;
  - 其相应单位正交化后的特征向量为 $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_p$.
- 我们令
  - $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_p\}$;
  - $\boldsymbol{V}' = (\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_p)$.
- 易知，
  - $\boldsymbol{V}'$ 是由相互正交的列向量组成，且 $\boldsymbol{V}\boldsymbol{V}' = \boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}$;
  - 特征值分解为 $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{V}'\boldsymbol{\Lambda}\boldsymbol{V}$ 即

$$\boldsymbol{V}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{V}' = (\boldsymbol{X}\boldsymbol{V}')'(\boldsymbol{X}\boldsymbol{V}') = \boldsymbol{\Lambda}.$$

主成分回归

动机

- 令 $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{V}'$，那么

$$\boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{\Lambda}.$$

- 在矩阵 $\boldsymbol{Z}_{n\times p} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_p)$ 中第 j 个列向量 $\boldsymbol{z}_j$ 为 $n \times 1$ 向量，且满足

$$\boldsymbol{z}_j \quad = \quad \boldsymbol{X}\boldsymbol{v}_j = (z_{1j}, z_{2j}, \cdots, z_{nj})'$$

- 第 j 个主成分为

$$\boldsymbol{z}_j = v_{1j}\boldsymbol{x}_1 + v_{2j}\boldsymbol{x}_2 + \cdots + v_{pj}\boldsymbol{x}_p$$

主成分回归

动机

- 由于 $\boldsymbol{X}$ 是已经标准化，即

$$\sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = 1, \quad j = 1, 2, \cdots, p.$$

- 而且 $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{V}'$ 和 $\boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{\Lambda}$，我们有

$$\sum_{i=1}^{n} z_{ij} = 0, \quad \sum_{i=1}^{n} z_{ij}^2 = \lambda_j, \quad \sum_{i=1}^{n} z_{ij} z_{ik} = 0 \ (j \neq k).$$

- 这表明了
  - 矩阵 $\boldsymbol{Z}$ 的各列之间正交；
  - 当 $\lambda_j \approx 0$ 时，$z_{1j}, z_{2j}, \cdots, z_{nj}$ 均近似为 0；

主成分回归

主成分估计的定义

- 当 $|\boldsymbol{X}'\boldsymbol{X}| \approx 0$ 时，存在一个 k，使得 $\lambda_{k+1}, \cdots, \lambda_p$ 均近似为 0。因此，$\boldsymbol{z}_{k+1}, \cdots, \boldsymbol{z}_p$ 近似为 0。
- 线性回归模型简化为

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{X}\boldsymbol{V}'\boldsymbol{V}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{Z}_{n\times p}\boldsymbol{\alpha}_{p\times 1} + \boldsymbol{\varepsilon}$$

- 我们将矩阵 $\boldsymbol{Z}$ 和向量 $\boldsymbol{\alpha}$ 按以下方式拆分

$$\begin{aligned}\boldsymbol{\alpha} &= (\alpha_1, \cdots, \alpha_k, \alpha_{k+1}, \cdots, \alpha_p)' = (\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2')',\\ \boldsymbol{Z} &= (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_k, \boldsymbol{z}_{k+1}, \cdots, \boldsymbol{z}_p) = (\boldsymbol{Z}_1, \boldsymbol{Z}_2).\end{aligned}$$

- 回归模型也可以写为

$$\boldsymbol{y} = \boldsymbol{Z}_1\boldsymbol{\alpha}_1 + \boldsymbol{Z}_2\boldsymbol{\alpha}_2 + \boldsymbol{\varepsilon} = \boldsymbol{Z}_1\boldsymbol{\alpha}_1 + \boldsymbol{\varepsilon}$$

主成分回归

主成分估计的定义

- 线性回归模型

$$\boldsymbol{y} = \boldsymbol{Z}_1\boldsymbol{\alpha}_1 + \boldsymbol{\varepsilon},$$

- $\boldsymbol{\alpha}_1$ 的最小二乘估计为

$$\hat{\boldsymbol{\alpha}}_1 = \left(\boldsymbol{Z}_1'\boldsymbol{Z}_1\right)^{-1}\boldsymbol{Z}_1'\boldsymbol{y} = \boldsymbol{\Lambda}_1^{-1}\boldsymbol{Z}_1'\boldsymbol{y}$$

其中，$\boldsymbol{\Lambda}_1 = \mathrm{diag}\{\lambda_1, \cdots, \lambda_k\}$，$\boldsymbol{\Lambda}_2 = \mathrm{diag}\{\lambda_{k+1}, \cdots, \lambda_p\}$.

- 那么，$\boldsymbol{\Lambda}$ 可以拆分为 $\boldsymbol{\Lambda}_1$ 和 $\boldsymbol{\Lambda}_2$，即

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda}_2 \end{pmatrix}$$

- 相应地，我们也可以将 $\boldsymbol{V}'$ 做拆分，即 $\boldsymbol{V}' = (\boldsymbol{V}_1', \boldsymbol{V}_2')$.

主成分回归

主成分估计的定义

- 根据

$$\boldsymbol{\beta} = \boldsymbol{V}'\boldsymbol{\alpha},$$

- 回归参数 $\boldsymbol{\beta}$ 的估计为

$$\hat{\boldsymbol{\beta}}_{\mathrm{PC}} = \boldsymbol{V}' \begin{pmatrix} \hat{\boldsymbol{\alpha}}_1 \\ \mathbf{0} \end{pmatrix} = \boldsymbol{V}_1' \hat{\boldsymbol{\alpha}}_1 = \boldsymbol{V}_1' \boldsymbol{\Lambda}_1^{-1} \boldsymbol{Z}_1' \boldsymbol{y}$$

- 称 $\hat{\boldsymbol{\beta}}_{\mathrm{PC}}$ 为 $\boldsymbol{\beta}$ 的**主成分估计**。

主成分回归

性质 1

- 主成分估计与最小二乘估计有什么关系吗?
- 主成分估计可写为最小二乘估计的线性变换，即

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\mathrm{PC}} &= \boldsymbol{V}_1'\boldsymbol{\Lambda}_1^{-1}\boldsymbol{Z}_1'\boldsymbol{y} \\
&= \boldsymbol{V}_1'\boldsymbol{\Lambda}_1^{-1}\boldsymbol{V}_1\boldsymbol{X}'\boldsymbol{y} \\
&= \boldsymbol{V}_1'\boldsymbol{\Lambda}_1^{-1}\boldsymbol{V}_1\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= \boldsymbol{V}_1'\boldsymbol{\Lambda}_1^{-1}\boldsymbol{V}_1\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{V}_1'\boldsymbol{\Lambda}_1^{-1}\boldsymbol{V}_1\boldsymbol{V}'\boldsymbol{\Lambda}\boldsymbol{V}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{V}_1'\boldsymbol{\Lambda}_1^{-1}\boldsymbol{V}_1(\boldsymbol{V}_1', \boldsymbol{V}_2')\begin{pmatrix}\boldsymbol{\Lambda}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{V}_1 \\ \boldsymbol{V}_2\end{pmatrix}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{V}_1'\boldsymbol{V}_1\hat{\boldsymbol{\beta}}.
\end{aligned}
$$

主成分回归

性质 2

- 主成分估计是无偏估计吗?
- 主成分估计的期望为

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_{\mathrm{PC}}) &= E(\boldsymbol{V}_1'\boldsymbol{V}_1\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{V}_1'\boldsymbol{V}_1 E(\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{V}_1'\boldsymbol{V}_1\boldsymbol{\beta} \end{aligned}$$

- 由于 $\boldsymbol{I}_p = \boldsymbol{V}'\boldsymbol{V} = \boldsymbol{V}_1'\boldsymbol{V}_1 + \boldsymbol{V}_2'\boldsymbol{V}_2$, 那么 $\boldsymbol{V}_1'\boldsymbol{V}_1 = \boldsymbol{I}_p - \boldsymbol{V}_2'\boldsymbol{V}_2$.
- 当 $k < p$ 时，$\boldsymbol{V}_1'\boldsymbol{V}_1\boldsymbol{\beta} = (\boldsymbol{I} - \boldsymbol{V}_2'\boldsymbol{V}_2)\boldsymbol{\beta} \neq \boldsymbol{\beta}$。
- 此时，主成分估计是有偏估计。

主成分回归

性质 3

- 虽然主成分估计是有偏估计，但是在均方误差的意义下，主成分估计是否优于最小二乘估计?

定理 3-5

当设计矩阵 $\boldsymbol{X}$ 存在多重共线性，选择合适的 k，可使得

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{PC}}) < \mathrm{MSE}(\hat{\boldsymbol{\beta}})$$

主成分回归

性质 3

证明：由于

$$\hat{\boldsymbol{\beta}}_{\mathrm{PC}} = \boldsymbol{V}' \begin{pmatrix} \hat{\boldsymbol{\alpha}}_1 \\ \mathbf{0} \end{pmatrix}$$

我们有

$$\begin{aligned}
\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{PC}}) &= E(\hat{\boldsymbol{\beta}}_{\mathrm{PC}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{\mathrm{PC}} - \boldsymbol{\beta}) \\
&= E\left(\boldsymbol{V}' \begin{pmatrix} \hat{\boldsymbol{\alpha}}_1 \\ \mathbf{0} \end{pmatrix} - \boldsymbol{V}'\boldsymbol{\alpha}\right)' \left(\boldsymbol{V}' \begin{pmatrix} \hat{\boldsymbol{\alpha}}_1 \\ \mathbf{0} \end{pmatrix} - \boldsymbol{V}'\boldsymbol{\alpha}\right) \\
&= E\left(\begin{pmatrix} \hat{\boldsymbol{\alpha}}_1 \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}\right)' \boldsymbol{V}\boldsymbol{V}' \left(\begin{pmatrix} \hat{\boldsymbol{\alpha}}_1 \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}\right) \\
&= E\left(\begin{pmatrix} \hat{\boldsymbol{\alpha}}_1 \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}\right)' \left(\begin{pmatrix} \hat{\boldsymbol{\alpha}}_1 \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}\right)
\end{aligned}$$

主成分回归

性质 3

证明：

$$
\begin{aligned}
\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{PC}}) &= E(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1)'(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1) + \|\boldsymbol{\alpha}_2\|^2 \\
&= E(\boldsymbol{\varepsilon}' \boldsymbol{Z}_1 \boldsymbol{\Lambda}_1^{-2} \boldsymbol{Z}_1' \boldsymbol{\epsilon}) + \|\boldsymbol{\alpha}_2\|^2 \\
&= \sigma^2 \mathrm{tr}\left(\boldsymbol{\Lambda}_1^{-1}\right) + + \|\boldsymbol{\alpha}_2\|^2 \\
&= \sigma^2 \sum_{j=1}^{k} \lambda_j^{-1} + \sum_{j=k+1}^{p} \alpha_j^2 \\
&= \mathrm{MSE}(\hat{\boldsymbol{\beta}}) + \left(\sum_{j=k+1}^{p} \alpha_j^2 - \sigma^2 \sum_{j=k+1}^{p} \lambda_j^{-1} \right)
\end{aligned}
$$

主成分回归

性质 3

证明： 我们已证明了

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{PC}}) = \mathrm{MSE}(\hat{\boldsymbol{\beta}}) + \left(\sum_{j=k+1}^{p} \alpha_j^2 - \sigma^2 \sum_{j=k+1}^{p} \lambda_j^{-1} \right)$$

由于设计矩阵存在多重共线性，因此有一部分特征值 λ_j 非常接近于零，不妨设后 $p-k$ 个特征值接近于零，则 $\sum_{j=k+1}^{p} \lambda_j^{-1}$ 将会很大，这导致了第二项为负。所以，

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{PC}}) < \mathrm{MSE}(\hat{\boldsymbol{\beta}})$$

主成分回归

性质 4

- 主成分估计的长度 vs 最小二乘估计的长度
- 主成分估计的模的平方为

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}}_{\mathrm{PC}}\|^2 &= (\hat{\boldsymbol{\beta}}_{\mathrm{PC}})'\hat{\boldsymbol{\beta}}_{\mathrm{PC}} \\
&= (\boldsymbol{V}_1'\boldsymbol{V}_1\hat{\boldsymbol{\beta}})'(\boldsymbol{V}_1'\boldsymbol{V}_1\hat{\boldsymbol{\beta}}) \\
&= \hat{\boldsymbol{\beta}}'\boldsymbol{V}_1'\boldsymbol{V}_1\boldsymbol{V}_1'\boldsymbol{V}_1\hat{\boldsymbol{\beta}} \\
&= \hat{\boldsymbol{\beta}}'\boldsymbol{V}_1'\boldsymbol{V}_1\hat{\boldsymbol{\beta}} \\
&\leq \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} \\
&= \|\hat{\boldsymbol{\beta}}\|^2
\end{aligned}$$

- 所以，$\|\hat{\boldsymbol{\beta}}_{\mathrm{PC}}\| \leq \|\hat{\boldsymbol{\beta}}\|$ 主成分估计是压缩估计。

主成分回归

如何选择主成分的个数?

- **保留特征值比重大的主成分**
  - 由于 $\sum_{j=1}^{p} \lambda_j = p$, 通常称 $\frac{\lambda_j}{p}$ 为第 j 个主成分 z_j 的贡献率。而 $\sum_{j=1}^{k} \frac{\lambda_j}{p}$ 为前 k 个主成分的累积贡献率。
  - 具体方案：给定一个定值 $c_{pc}(0 < c_{pc} < 1)$，如果存在 k，使得

$$\sum_{j=1}^{k-1} \frac{\lambda_j}{p} < c_{\mathrm{pc}}, \quad \sum_{j=1}^{k} \frac{\lambda_j}{p} \geq c_{\mathrm{pc}}.$$

    由此选取 k。
  - 通常，$c_{\mathrm{pc}} = 70\%, 75\%$ 或 80%。

主成分回归

如何选择主成分的个数?

- **删除特征值接近于令的主成分**
  - 具体方案：给定一个定值 c_0，如果

$$\lambda_k \geq c_0, \lambda_{k+1} < c_0$$

    由此选取 k。

- **均方误差确定 k**
  - 由于 $\sum_{j=1}^{p} \lambda_j^{-1}$ 与估计的均方误差有关，我们并不希望这个值太大，我们可以选取 k 满足

$$\sum_{j=1}^{k} \lambda_j^{-1} \leq 5k.$$