可视化方法在统计分析课程应用–以主元分析为例

张成　徐涛
沈阳化工大学　理学院　辽宁沈阳　110142

【摘　要】统计分析是研究生教育中的一门基础课程，统计分析方法在研究生的研究工作中起着至关重要的作用。为了使学生能够更加深入理解和掌握基本的数据分析方法，以主元分析为例，依托Tennessee Eastman过程数据开展可视化教学方法实践研究。将被广泛应用的实验数据引入教学过程当中，既可以拓宽学生视野，又可以提高学生学习兴趣。而将复杂繁琐的统计分析公式与理论进行可视化既可以激发学生的创造力与思维能力，又可以使学生在学习过程中不断探索、实践，更好地掌握相关知识。

Visualization method applied in statistical analysis course-take principle element analysis as an example

Zhang Cheng　Xu Tao
School of Science，Shenyang University of Chemical Engineering Shenyang，Liaoning　110142

【Abstract】Statistical analysis is a basic course in graduate education，and statistical analysis methods play a vital role in graduate research work. In order to enable students to more deeply understand and master the basic data analysis methods，meta-analysis as an example is carried out based on Tennessee Eastman process data. Integrating widely used experimental data into the teaching process can not only broaden students 'horizons，but also improve students' interest in learning. The visualization of complex and complicated statistical analysis formulas and theories can not only stimulate students' creativity and thinking ability，but also enable students to continuously explore and practice in the learning process，so as to better master relevant knowledge.

1. 引言

统计分析方法是一种通过数学模型对收集到的数据进行整理、分析和解释的方法。实践表明，统计分析课程的教学内容包含大量复杂数学公式、定理和定义，这一课程特点严重影响教与学的质量与效果。在当前大数据时代，统计分析方法在机器学习、深度学习等领域起到了至关重要的作用。因此，为提高学生统计分析能力而适当改进统计分析课程教学方式是必要的。

近些年，统计分析方法在基于数据驱动的生产过程监控领域已经取得了显著的成果。为了提高生产过程的安全性，一系列基于主元分析的方法被引入过程状态监控领域[1-2]。该类方法既保障了生产安全，又提高了生产效率。针对过程动态性分析问题，一种基于数据增广矩阵的主元分析方法被提出，同时被应用到线性动态过程的性能监控中[3]。该方法通过主元分析方法对高维数据进行维数约减以实现降维目的，同时能够有效捕获过程变量的线性相关关系。考虑过程变量间存在非线性关系，一种基于核技巧的主元分析方法被提出[4]。该方法是主元分析方法的一种变异，其核心思想是通过一个非线性映射将原始数据映射到一个线性可分的高维空间，然后在高维空间中应用主元分析对数据进行降维并完成过程状态监控任务。

综上，可以看出主元分析方法在生产过程中已经发挥了重要作用。为了学生能够更加准确理解和掌握主元分析方法，在教学过程中引入 TE 过程（Tennessee Eastman Process）数据[5]，同时结合可视化教学方法以提高学生在学习过程中的主动探索能力和实践能力。在这种教学模式下，学生需要主动探索现实世界的问题和挑战，通过可视化方法的实施来获取更深刻的知识和技能。

2. 主元分析简介

主元分析是一种将高维数据变换为低维数据的降维方法。其目标通常可以理解为找到新的投影方向，使数据在新

的方向上的投影分散程度最大化，具体过程如下。

假设 $\boldsymbol{X}_{m\times n}$ 是一个包含 m 个样本、n 个变量的数据集。在进行主元分析之前，对变量标准化是一个必要的步骤，如式（1）所示。

$$\tilde{x}_{ij}=(x_{ij}-\frac{1}{m}\sum_{k=1}^{m}x_{kj})/[\sqrt{\frac{1}{m-1}\sum_{l=1}^{m}(x_{lj}-\frac{1}{m}\sum_{k=1}^{m}x_{kj})^2}] \quad (1)$$

接下来，计算标准化后数据集 $\tilde{\boldsymbol{X}}$ 的协方差矩阵（相关系数矩阵），如式（2）所示。

$$\boldsymbol{C}=\frac{1}{m-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}} \quad (2)$$

在式（2）中，当 $i=j$ 时，$\boldsymbol{C}_{ij}$ 为变量 $\tilde{x}_i$ 方差，而当 $i\neq j$ 时，$\boldsymbol{C}_{ij}$ 为变量 $\tilde{x}_i$ 与 $\tilde{x}_j$ 的协方差。易知，矩阵 $\boldsymbol{C}$ 是一个对称矩阵，即 $\boldsymbol{C}^{\mathrm{T}}=\boldsymbol{C}$。

然后，将协方差矩阵 $\boldsymbol{C}$ 进行特征值分解，如式（3）所示。

$$\boldsymbol{C}\xi=\lambda\xi \quad (3)$$

其中，λ 为 $\boldsymbol{C}$ 的特征值，而 ξ 为特征值 λ 所对应的特征向量。将特征值由大到小进行排序，记为 $\lambda_1,\lambda_2,\cdots,\lambda_r,\cdots,\lambda_n$，相应的特征向量为 $\xi_1,\xi_2,\cdots,\xi_r,\cdots,\xi_n$，满足 $\boldsymbol{C}\xi_i=\lambda_i\xi_i$。

主成分的个数通常通过累计方差贡献率的方法进行确定，如式（4）所示。通过设置阈值 α，如 $\alpha=0.95$，当 cpv=α 时，相应的主成分个数 r 即被确定下来。

$$cpv=\sum_{j=1}^{r}\lambda_j/\sum_{i=1}^{n}\lambda_i \quad (4)$$

一旦主成分数量 r 被确定下来，数据集 $\tilde{\boldsymbol{X}}$ 在主成分方向的投影可以通过式（5）计算获得。

$$\boldsymbol{t}_i=\tilde{\boldsymbol{x}}_i\boldsymbol{P}_r \quad (5)$$

其中，$\tilde{\boldsymbol{x}}_i$ 为数据 $\tilde{\boldsymbol{X}}$ 中的第 i 个样本，而 $\boldsymbol{P}_r=[\xi_1,\xi_2,\cdots,\xi_r]$ 是负载矩阵。$\tilde{\boldsymbol{X}}$ 的投影可表示为 $\boldsymbol{T}_{m\times r}=[\boldsymbol{t}_1^{\mathrm{T}}\ \ \boldsymbol{t}_2^{\mathrm{T}}\ \ \cdots\ \ \boldsymbol{t}_m^{\mathrm{T}}]^{\mathrm{T}}=\tilde{\boldsymbol{X}}\boldsymbol{P}_r$。可以证明，数据集 $\tilde{\boldsymbol{X}}$ 的第 j 个主成分 t_j 的方差为 $\tilde{\boldsymbol{X}}$ 的协方差矩阵 $\boldsymbol{C}$ 的第 j 个特征值 λ_j，同时，舍弃的特征值 $\lambda_{r+1},\cdots,\lambda_n$ 通常较小。当数据集 $\tilde{\boldsymbol{X}}$ 存在显著线性相关关系时，$\lambda_l\approx 0\ \ (l=r+1,r+2,\cdots,n)$。于是，可以得到 t_l 的方差近似为 0，即 $Var(\tilde{\boldsymbol{x}}\xi_l)\approx 0$，其中向量 $\tilde{\boldsymbol{x}}=[\tilde{x}_1\ \ \tilde{x}_2\ \ \cdots\ \ \tilde{x}_n]$ 是由输入变量构成。由于数据集已被中心化处理，这意味着 t_l 的均值为 0。综上，$\tilde{\boldsymbol{x}}\xi_l\approx 0$，于是可以通过数据变量与舍弃的特征向量内积运算的方式获取输入变量的线性相关关系。

3. 依托 Tennessee Eastman 过程数据开展主元分析知识点可视化研究

TE 过程是由美国 Eastman 化学公司过程控制小组的 J.J.Downs 和 E.F.Vogel 提出的仿真模拟模型，主要用于过程控制技术的研究[5]。该模型描述了装置、物料和能量之间的非线性关系，广泛应用于装置控制方案设计和教学等领域。在 Downs 等提出的 TE 过程模型基础上，Bathelt 等提出了修订版的 TE 过程模型，扩大了仿真模型的可用性[6]。为了开展主元分析方法的可视化教学，选取了 TE 过程模态 1 数据进行分析和讨论。数据集中包含 1000 个样本，每个样本包含 13 个变量，名称分别为：物料 A 流量 x_1，物料 D 流量 x_2，物料 E 流量 x_3，总进料流量 x_4，压缩机返回物料流量 x_5，反应器给料流量 x_6，反应器压力 x_7，反应器液位 x_8，反应器温度 x_9，排空物料流量 x_{10}，气液分离器温度 x_{11}，气液分离器液位 x_{12}，气液分离器压力 x_{13}。

3.1 标准化结果可视化

数据标准化是数据分析过程中的一个重要环节。在标准化知识的介绍过程中，以 TE 过程数据为例，展现数据标准化前后的区别。图 1 展示了变量 x_7 和 x_{13} 分布散点图。可以发现数据标准化前后的位置发生了变化。标准化后的数据中心变成了坐标原点，即变量的均值变为 0，这为后面的协方差矩阵计算带来了便利。同时，新的变量具有了统一的单位方差。通过将标准化过程进行可视化，可以帮助学生深入理解对数据集进行标准化的两个目的：一是将数据中心由最初的非坐标原点位置移动到了坐标原点位置，二是消除不同数据量级的影响，使得不同数据之间的比较变得更加公平和有效。

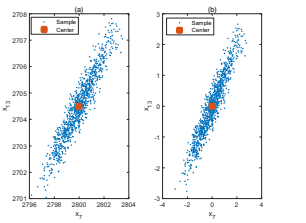

图 1 标准化：（a）原始数据；（b）标准化后数据

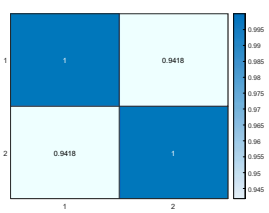

图 2 相关系数矩阵

3.2 协方差矩阵可视化

在主元分析中，协方差矩阵反应了输入变量之间的相关关系。依概率论相关知识，总体协方差可以表示为 $\mathbf{cov}(X,Y)=E[(X-E(X))(Y-E(Y))]$，而从抽样的角度，样本的协

方差 $S_{x_ix_j}=\frac{1}{m-1}\sum_{k=1}^{m}(x_{ki}-m_{x_i})(x_{kj}-m_{x_j})\overset{m_{x_i}=0}{\underset{m_{x_j}=0}{=}}\frac{1}{m-1}\sum_{k=1}^{m}x_{ki}x_{kj}$，其中 m_{x_i} 和 m_{x_j} 分别为变量

x_i 和 x_j 的样本均值。可以看出，样本的协方差矩阵是一个对称矩阵。在 TE 数据中，通过分析变量 x_7 和 x_{13} 的协方差矩阵的特点，能够使学生能够进一步理解协方差及协方差矩阵的含义。图 2 给出变量 x_7 和 x_{13} 的协方差矩阵，即相关系数矩阵热力图。可以看出，协方差矩阵呈现对称结构，同时相关系数矩阵暗示上述两个变量存在一定的线性相关性，其中图中横、纵轴标签 1 和 2 分别表示变量 x_7 和 x_{13}。

3.3 主成分可视化分析

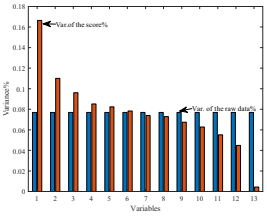

图 3 主元分析前后数据集变量方差占比变化

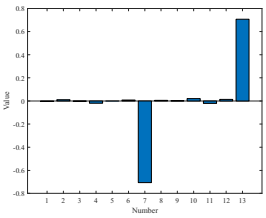

图 4 最小特征值所对应的特征向量

根据主元分析理论，需要计算协方差矩阵的由大到小排列的特征值，同时计算各个特征值占特征值总和的百分比，可视化结果如图（3）所示。为了对比分析，图中同时给出标准化后的各个原始变量方差占比结果。可以看出，通过主元分析方法获得的新的变量的方差是对原始数据变量方差和的一次分解。这种分解方式，使得原始数据的信息总量大部分分解到少数的几个主要变量上，如图中的前六个变量。值得注意的是变换后的最后一个变量 t_{13} 方差最小且接近于 0，这说明原始数据集中包含一种近似线性相关关系。这种线性相关关系需要通过对该特征值所对应的特征向量进行分析来确定。仿真实验得到协方差矩阵最小特征值 λ_{13} 的特征向量 $\xi_{13}=[-0.004,\ 0.011,\ -0.003,\ -0.019,\ -0.001,\ 0.008,\ -0.706,\ 0.005,\ 0.003,\ 0.020,\ -0.021,\ 0.014,\ 0.707]^{\mathrm{T}}$，如图（4）所示。由 $\tilde{x}\xi_{13}\approx 0$，经过进一步处理可以得到 $x_7\approx x_{13}$，这说明输入变量 x_7 和 x_{13} 之间的确存在一种较强的线性相关关系。

通过 TE 过程的实验数据分析主成分的方差变化，可以使学生能够了解主元分析是数据信息的重组方法，其目标是将总的信息即原始变量的方差和重新分解到少数几个正交方向上，这些新的方向可以包含原始数据的大部分信息，即在这些新的方向上数据呈现出较大的分散程度，从而避免了数据信息遮盖效应。同时，通过考察协方差矩阵的最小特征值，可以确定原始数据中隐藏的线性相关关系，这也为统计分析课程中回归分析内容的学习起到了铺垫作用。

参考文献

[1]周东华，李钢，李元. 数据驱动的工业过程故障诊断技术：基于主元分析与偏最小二乘的方法[M]. 科学出版社，2011.

[2]赵春晖，王福利，姚远. 基于时段的间歇过程统计建模、在线监测及质量预报[J]. 自动化学报，2010，36（3）：366–374.

[3]张成，戴絮年，李元. 基于 DPCA 残差互异度的故障检测与诊断方法[J]. 自动化学报，2022，48（1）：292–301.

[4]周卫庆，司风琪. 基于 KPCA 残差方向梯度的故障检测方法及应用[J]. 仪器仪表学报，2017，38（10）：2518–2524.

[5]Downs J J，Vogel E F. A plant–wide industrial process control problem. Computers & Chemical Engineering，1993（17）：245–255.

[6]Bathelt A，Ricker N L，et al. Revision of the Tennessee Eastman Process Model. IFAC–PapersOnLine，2015（48）：309–314.

基金项目：辽宁省研究生教育教学改革项目（LNYJG2022177）；辽宁省教育厅基本科研项目（LJKMZ20220792）.

作者简介：张成，男，博士，副教授，研究方向：基于数据驱动的工业过程监控。