基于 PCA 的信息压缩: 从一阶到高阶

夏志明^{1,2}, 徐宗本^{1*}

1. 西安交通大学数学与统计学院, 西安 710049

2. 西北大学数学学院, 西安 710127

* 通信作者. E-mail: statxzm@nwu.edu.cn

收稿日期: 2017–11–13; 接受日期: 2018–01–10; 网络出版日期: 2018–09–19

国家自然科学基金 (批准号: 11771353, 11201372, 91330204, 11690011, 11626252) 资助项目

摘要 本文概述信息压缩背景下从一阶到高阶主成分分析的统计原理, 并从 3 个不同角度揭示各阶主成分分析的特点和局限, 同时指出可能的研究方向. 首先以相似的结构综述一阶到高阶主成分分析的原理及已有的发展, 并进一步分析其内在相似统计意义, 揭示其共性结构——“勾股定理”, 并以此为基础展示主成分分析的两种等价表达——“变异性最大”与“平方损失最小”. 其次深入分析了一阶到高阶 PCA (principle component analysis) 的可能发展: 从“勾股定理”出发, PCA 可以推广到更一般损失函数形式的“稳健”或“稀疏”类 PCA; 从张量分解与主成分分析之间的关联出发, PCA 可以为构造不同的张量分解提供新思路; 从分析一阶到高阶主成分分析在揭示“各向异性”结构上的固有局限出发, PCA 能够推广到更有价值的“深度”PCA

关键词 主成分分析, 信息压缩, 高阶张量, Tucker 分解, 各向异性

1 引言

信息量通常指对数据中的不确定性的解释程度或消除程度. 在统计学和信息技术文献中经常用方差来度量数据所呈现出来的不确定性 (uncertainty) 或变异性 (variation), 在某种模型复杂度的条件下, 对方差的解释多少可度量其结构的信息多少. 在大数据的背景下, 数据量越来越大, 从而为数据的存储、传输等造成极大困难, 于是“信息压缩” (information compression) 或“特征提取” (feature extraction) 变得必不可少, 已成为进一步数据处理的基础.

信息压缩的方法与工具很多, 其中有一个结构简单、应用广泛的方法是主成分分析 (principle component analysis, PCA). 经典的 PCA 方法发端于压缩向量形式的数据或者称为“一阶” (one-order) 形式的数据, 这对大多数的数据形式都是适应的; 然而亦存在很多矩阵形式的数据, 譬如: 图像、推荐系统等, 这种数据表现为两个方向 (mode) 的变化, 称为“二阶” (two-order) 数据. 虽然理论上任何二阶

引用格式: 夏志明, 徐宗本. 基于 PCA 的信息压缩: 从一阶到高阶. 中国科学: 信息科学, 2018, 48: 1622–1633, doi: 10.1360/N112017-00238
Xia Z M, Xu Z B. Information compression based on principal component analysis: from one-order to higher-order (in Chinese). Sci Sin Inform, 2018, 48: 1622–1633, doi: 10.1360/N112017-00238

数据都能够直接“拉直”(reshaping)成向量即“一阶”数据来处理,但这种操作将或多或少破坏其原有数据结构,以至于对这些数据的后期处理造成不利影响.当然,现实中也面对 3 个或 3 个以上方向变化的数据,譬如:视频数据是一个“数据方体”(data cuboid),包含 3 个方向的变化,保持这种结构对视频数据处理显得非常必要,这类数据可称为“三阶”数据,推荐系统面对的可能是多个属性的同时观测,其数据结构也常常表现为 3 个或以上的方向.含 3 个方向以上变化的数据统称为“高阶”(high-order)数据.

2 从一阶 PCA 到高阶 PCA

2.1 一阶情形: 向量 PCA

多样本向量数据的主成分分析方法(PCA)起源最早^[1,2],其基本思想是寻找数据投影的“变异性”(用方差度量)最大的方向作为主成分.具体来说,考虑 N 个独立向量数据 $\mathbf{y}^{(n)} \in \mathbb{R}^I, n = 1, \dots, N$,依次构造 r 个标准正交的方向如下:

$$\begin{aligned}\hat{\mathbf{u}}_1 &= \arg \max_{\|\mathbf{u}_1\|=1} \widehat{\text{Var}}[\mathbf{u}_1^T \mathbf{y}] = \arg \max_{\|\mathbf{u}_1\|=1} \mathbf{u}_1^T \mathbf{S}_v \mathbf{u}_1, \\ \hat{\mathbf{u}}_2 &= \arg \max_{\|\mathbf{u}_2\|=1, \mathbf{u}_2 \perp \mathbf{u}_1=0} \widehat{\text{Var}}[\mathbf{u}_2^T \mathbf{y}] = \arg \max_{\|\mathbf{u}_2\|=1, \mathbf{u}_2 \perp \mathbf{u}_1=0} \mathbf{u}_2^T \mathbf{S}_v \mathbf{u}_2, \\ &\vdots \\ \hat{\mathbf{u}}_r &= \arg \max_{\|\mathbf{u}_r\|=1, \mathbf{u}_r \perp \mathbf{u}_i=0, i \in [r-1]} \widehat{\text{Var}}[\mathbf{u}_r^T \mathbf{y}] = \arg \max_{\|\mathbf{u}_r\|=1, \mathbf{u}_r \perp \mathbf{u}_i=0, i \in [r-1]} \mathbf{u}_r^T \mathbf{S}_v \mathbf{u}_r,\end{aligned}\quad (1)$$

其中 $\widehat{\text{Var}}[\mathbf{u}_i^T \mathbf{y}]$ 是随机向量 $\mathbf{y}$ 在方向 $\mathbf{u}_i$ 上的投影变异性, $\mathbf{S}_v = \frac{1}{N} \sum_{n=1}^N (\mathbf{y}^{(n)} - \bar{\mathbf{y}})(\mathbf{y}^{(n)} - \bar{\mathbf{y}})^T$ 表示向量数据的“协方差矩阵”.上述统计推断过程亦可用矩阵符号等价的表达为

$$\hat{\mathbf{U}} = \arg \max_{\mathbf{U}^T \mathbf{U} = \mathbf{I}_r} \widehat{\text{Var}}[\mathbf{U}^T \mathbf{y}] = \arg \max_{\mathbf{U}^T \mathbf{U} = \mathbf{I}_r} \text{Tr}[\mathbf{U}^T \mathbf{S}_v \mathbf{U}], \quad (2)$$

其中 $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_r) \in \mathbb{R}^{I \times r}$, $\mathbf{I}_r$ 为 r 阶单位阵, $\text{Tr}(\cdot)$ 表示矩阵的“迹”(trace).

式(1)中构造欧氏空间的标准正交基的方法不同于经典的“施密特正交化”(Schmidt orthogonalization),因为它本质上是由数据的“协方差矩阵”(covariance matrix)决定的或者说反应了数据各分量的方差及相互间的线性关系,被称为“Karhunen-Loève 变换”或“Hotelling 变换”^[1,2].这种 K-L 变换的一个突出特点是各方向 $\hat{\mathbf{u}}_i, i = 1, \dots, r$ 上的数据投影波动性 $\sigma_i^2 = \hat{\mathbf{u}}_i^T \mathbf{S}_v \hat{\mathbf{u}}_i$ 常常从最大迅速下降,即 $\sigma_1^2 \geq \sigma_2^2 \geq \dots \geq \sigma_r^2$,且前几个方向主导着数据的波动性.严格来说,它们之间满足以下性质.

性质1 给定 $\hat{\mathbf{U}}$ 满足式(2),则

$$\begin{aligned}\sum_{n=1}^N \|\mathbf{y}_o^{(n)}\|_F^2 &= \sum_{n=1}^N \|\mathbf{y}_o^{(n)} - \hat{\mathbf{U}} \hat{\mathbf{v}}^{(n)}\|_F^2 + \sum_{n=1}^N \|\hat{\mathbf{v}}^{(n)}\|_F^2 \\ &= \sum_{n=1}^N \|\mathbf{y}_o^{(n)} - \hat{\mathbf{U}} \hat{\mathbf{v}}^{(n)}\|_F^2 + N \sum_{i=1}^r \sigma_i^2,\end{aligned}\quad (3)$$

其中 $\mathbf{y}_o^{(n)} = \mathbf{y}^{(n)} - \bar{\mathbf{y}}, \hat{\mathbf{v}}^{(n)} = \hat{\mathbf{U}}^T \mathbf{y}_o^{(n)}$.

等式 (3) 揭示了多样本向量 PCA 的一个重要的等价描述 — 中心化数据矩阵 $\mathbf{Y} - \bar{\mathbf{y}}\mathbf{1}^T$ 的矩阵近似问题:

$$(\hat{\mathbf{V}}, \hat{\mathbf{U}}) = \arg \min_{\mathbf{V}, \mathbf{U}^T \mathbf{U} = \mathbf{I}_r} \sum_{n=1}^N \left\| \mathbf{y}_o^{(n)} - \hat{\mathbf{U}} \hat{\mathbf{v}}^{(n)} \right\|_F^2. \quad (4)$$

向量数据的主成分分析方法在图像、视频等领域应用广泛, 其相应的统计理论也慢慢发展起来, 其中主要关心经验特征值与特征向量的大样本性质, 譬如相合性与渐近正态性等. 当样本量 N 趋于无穷且向量维数 p 固定时, 比较早的代表性论文是由 Anderson^[3] 在数据服从多元正态分布的假设下证明了样本协方差的特征值和特征向量的渐近正态性. 最近十几年的研究主要针对当样本量 $N, p \rightarrow \infty$ 条件下的 PCA 的大样本性质. 这个领域的统计性质的代表论文包括 Johnstone 的 [4]、Baik 等的 [5]、Paul 的 [6]、Johnstone 和 Lu 的 [7]、Jung 和 Marron 的 [8]、Onatski 的 [9]、Shen 等的 [10]、Wang 和 Fan 的 [11] 等. 这些优秀的工作揭示出经验特征值和特征向量的渐近行为 (包括相合性与渐近正态性等) 取决于 N, d 的相对大小以及协方差矩阵的特点. 特别地, 当 $N \ll d$ 且均趋于无穷时, 除非最大特征值发散充分得快, 否则最大经验特征向量是不相合的, 这种研究结果刺激了该条件下关于 PCA 的稀疏版本的研究, 下一节的方法拓展中将对此有重点的简单介绍.

2.2 二阶情形: 矩阵 PCA

在实际数据处理和信息压缩应用中, 图像等常常是以矩阵形式存在的. 这类数据的传统处理手段是通过“拉直”将其化归为向量数据处理, 显然该方法一方面破坏了原有矩阵数据的原始结构, 另一方面也增加了统计模型复杂度, 进而加剧了计算上的“维数诅咒” (curse of dimension). 事实上, 在不“拉直”矩阵的条件下, 向量 PCA 的“投影变异性最大”的统计思想是可以优雅地被推广到矩阵 (二阶张量) 情形的. Yang 等^[12] 于 2004 年首次提出了一种“二维主成分分析” (two-dimensional PCA, 2DPCA) 的方法, 直接对矩阵数据进行压缩, 其基本思想是将矩阵沿某一个方向投影以使得“投影特征向量”的变异性最大. 具体来说, 假定给定 N 个独立的矩阵样本或二阶张量样本 (two-order tensor) $\mathbf{Y}^{(n)} \in \mathbb{R}^{I \times J}, n = 1, 2, \dots, N$, 可以类似地依次构造 r 个标准正交方向如下:

$$\begin{aligned} \hat{\mathbf{u}}_1 &= \arg \max_{\|\mathbf{u}_1\|=1} \widehat{\text{Var}}[\mathbf{Y}\mathbf{u}_1] = \arg \max_{\|\mathbf{u}_1\|=1} \mathbf{u}_1^T \mathbf{S}_m \mathbf{u}_1, \\ \hat{\mathbf{u}}_2 &= \arg \max_{\|\mathbf{u}_2\|=1, \mathbf{u}_2 \perp \mathbf{u}_1=0} \widehat{\text{Var}}[\mathbf{Y}\mathbf{u}_2] = \arg \max_{\|\mathbf{u}_2\|=1, \mathbf{u}_2 \perp \mathbf{u}_1=0} \mathbf{u}_2^T \mathbf{S}_m \mathbf{u}_2, \\ &\vdots \\ \hat{\mathbf{u}}_r &= \arg \max_{\|\mathbf{u}_r\|=1, \mathbf{u}_r \perp \mathbf{u}_i=0, i \in [r-1]} \widehat{\text{Var}}[\mathbf{Y}\mathbf{u}_r] = \arg \max_{\|\mathbf{u}_r\|=1, \mathbf{u}_r \perp \mathbf{u}_i=0, i \in [r-1]} \mathbf{u}_r^T \mathbf{S}_m \mathbf{u}_r, \end{aligned} \quad (5)$$

其中 $\widehat{\text{Var}}[\mathbf{Y}\mathbf{u}_i]$ 是随机矩阵 $\mathbf{Y}$ 各行在方向 $\mathbf{u}_i$ 上的投影变异性, 此时投影为 I 维列向量, 其变异性为各分量样本方差之和, 相应地,

$$\mathbf{S}_m = \sum_{i=1}^I \left[\frac{1}{N} \sum_{n=1}^N \left(\mathbf{y}_{i:}^{(n)} - \bar{\mathbf{y}}_{i:} \right) \left(\mathbf{y}_{i:}^{(n)} - \bar{\mathbf{y}}_{i:} \right)^T \right] \quad (6)$$

表示各行向量的“样本协方差矩阵”之和, 其中 $\mathbf{y}_{i:}^{(n)}, \bar{\mathbf{y}}_{i:}$ 表示第 n 个矩阵数据 $\mathbf{Y}^{(n)}$ 的第 i 行向量及其样本均值. 类似于式 (2), 上述统计推断过程亦可用矩阵符号简洁地表达为

$$\hat{\mathbf{U}} = \arg \max_{\mathbf{U}^T \mathbf{U} = \mathbf{I}_r} \widehat{\text{Var}}[\mathbf{Y}\mathbf{U}] = \arg \max_{\mathbf{U}^T \mathbf{U} = \mathbf{I}_r} \text{Tr}[\mathbf{U}^T \mathbf{S}_m \mathbf{U}]. \quad (7)$$

式 (5) 描述的是数据矩阵 $\mathbf{Y}$ 的所有行向量同时向一个方向投影而获得“投影特征向量”并计算其变异性, 所以该特征向量就是数据矩阵的列向量的一个线性组合, 它本质上反映的是各列向量之间的“向量水平”的线性关系和变异性, 部分地反映了各行内部元素之间的关系, 但并不能反映列向量内部各元素之间的关系, 因而不能反映行向量之间的“向量水平之间”的关系. 类似地, 各方向 $\hat{\mathbf{u}}_i, i = 1, \dots, r$ 上的数据投影波动性 $\sigma_i^2 = \hat{\mathbf{u}}_i^T \mathbf{S}_m \hat{\mathbf{u}}_i$ 常从最大迅速下降, 即 $\sigma_1^2 \geq \sigma_2^2 \geq \dots \geq \sigma_r^2$, 且前几个方向占据主导. 此处一个很自然的问题是, 在矩阵情形是否仍然存在类似于 (3) 的性质呢? 答案是肯定的.

性质2 给定 $\hat{\mathbf{U}}$ 满足式 (10), 则

$$\begin{aligned} \sum_{n=1}^N \|\mathbf{Y}_o^{(n)}\|_F^2 &= \sum_{n=1}^N \|\mathbf{Y}_o^{(n)} - \hat{\mathbf{V}}^{(n)} \hat{\mathbf{U}}^T\|_F^2 + \sum_{n=1}^N \|\hat{\mathbf{V}}^{(n)}\|_F^2 \\ &= \sum_{n=1}^N \|\mathbf{Y}_o^{(n)} - \hat{\mathbf{V}}^{(n)} \hat{\mathbf{U}}^T\|_F^2 + N \sum_{i=1}^r \sigma_i^2, \end{aligned} \quad (8)$$

其中 $\mathbf{Y}_o^{(n)} = \mathbf{Y}^{(n)} - \bar{\mathbf{Y}}, \hat{\mathbf{V}}^{(n)} = \mathbf{Y}_o^{(n)} \hat{\mathbf{U}}$.

类似于式 (3), 等式 (8) 同样揭示了多样本矩阵 PCA 的另一个重要的等价视角 —— 一组中心化数据矩阵 $\mathbf{Y}^{(n)} - \bar{\mathbf{Y}}$ 的矩阵近似优化问题:

$$(\hat{\mathbf{V}}^{(n)}, \hat{\mathbf{U}}) = \arg \min_{\mathbf{V}^{(n)}, \mathbf{U}^T \mathbf{U} = \mathbf{I}_r} \sum_{n=1}^N \|\mathbf{Y}_o^{(n)} - \hat{\mathbf{V}}^{(n)} \hat{\mathbf{U}}^T\|_F^2. \quad (9)$$

注意到 Yang 的 2DPCA 只是沿着该数据矩阵 $\mathbf{Y}^{(n)}$ 的右侧 (也就是各行) 进行投影压缩而对左侧或者列保持不变, 从而使其压缩能力仍能进一步提升, 这一点 Yang 在后续的模拟实验中也提到. Ye 等^[13] 从一个矩阵的行列同时展开“双线性变换” (bilinear transformation) 建立了“GPCA”, 改进了 2DPCA 的不足之处. Cai 等^[14] 和 Ye^[15] 把上述理论最后统一到最小二乘框架下作了统一描述. 具体来说, 针对同样的 N 个矩阵数据, 如下构造两个方向上的标准正交基:

$$(\hat{\mathbf{U}}_1, \hat{\mathbf{U}}_2) = \arg \max_{\mathbf{U}_k^T \mathbf{U}_k = \mathbf{I}_{r_k}, k=1,2} \widehat{\text{Var}} [\mathbf{U}_1^T \mathbf{Y} \mathbf{U}_2] = \arg \max_{\mathbf{U}_k^T \mathbf{U}_k = \mathbf{I}_{r_k}, k=1,2} \sum_{n=1}^N \|\hat{\mathbf{U}}_1^T (\mathbf{Y}^{(n)} - \bar{\mathbf{Y}}) \hat{\mathbf{U}}_2\|_F^2, \quad (10)$$

称为主成分; 称 $\mathbf{V}^{(n)} = \hat{\mathbf{U}}_1^T (\mathbf{Y}^{(n)} - \bar{\mathbf{Y}}) \hat{\mathbf{U}}_2$ 为“核矩阵” (core matrix). 易验证以下性质.

性质3 给定 $\hat{\mathbf{U}}_1, \hat{\mathbf{U}}_2$ 满足式 (10), 则

$$\begin{aligned} \sum_{n=1}^N \|\mathbf{Y}_o^{(n)}\|_F^2 &= \sum_{n=1}^N \|\mathbf{Y}_o^{(n)} - \hat{\mathbf{U}}_1 \hat{\mathbf{V}}^{(n)} \hat{\mathbf{U}}_2^T\|_F^2 + \sum_{n=1}^N \|\hat{\mathbf{V}}^{(n)}\|_F^2 \\ &= \sum_{n=1}^N \|\mathbf{Y}_o^{(n)} - \hat{\mathbf{U}}_1 \hat{\mathbf{V}}^{(n)} \hat{\mathbf{U}}_2^T\|_F^2 + \sum_{n=1}^N \|\hat{\mathbf{U}}_1^T (\mathbf{Y}^{(n)} - \bar{\mathbf{Y}}) \hat{\mathbf{U}}_2\|_F^2, \end{aligned} \quad (11)$$

其中 $\mathbf{Y}_o^{(n)} = \mathbf{Y}^{(n)} - \bar{\mathbf{Y}}, \hat{\mathbf{V}}^{(n)} = \hat{\mathbf{U}}_1^T \mathbf{Y}_o^{(n)} \hat{\mathbf{U}}_2$.

因此, 式 (10) 的另一个等价的优化表达式为

$$(\hat{\mathbf{V}}^{(n)}, \hat{\mathbf{U}}_1, \hat{\mathbf{U}}_2) = \arg \min_{\mathbf{V}^{(n)}, \mathbf{U}_k^T \mathbf{U}_k = \mathbf{I}_{r_k}, k=1,2} \sum_{n=1}^N \|\mathbf{Y}_o^{(n)} - \mathbf{U}_1 \mathbf{V}^{(n)} \mathbf{U}_2^T\|_F^2.$$

2.3 高阶情形: 张量 PCA

类似于 Yang 等^[12]的“二维主成分分析”(two-dimensional PCA, 2DPCA), 可以要求高阶张量在某一个方向的“投影”或“特征向量”的变异性最大来实现信息压缩. 具体来说, 给定 N 个独立的高阶张量样本 (high-order tensor) $\mathcal{Y}^{(n)} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_p}$, $n = 1, 2, \dots, N$, 假定它沿第 k 方向 (mode) 投影, 则依次可构造 r 个标准正交方向如下:

$$\begin{aligned}\hat{\mathbf{u}}_1 &= \arg \max_{\|\mathbf{u}_1\|=1} \widehat{\text{Var}} [\mathcal{Y} \times_k \mathbf{u}_1^T] = \arg \max_{\|\mathbf{u}_1\|=1} \mathbf{u}_1^T \mathbf{S}_t \mathbf{u}_1, \\ \hat{\mathbf{u}}_2 &= \arg \max_{\|\mathbf{u}_2\|=1, \mathbf{u}_2 \perp \mathbf{u}_1=0} \widehat{\text{Var}} [\mathcal{Y} \times_k \mathbf{u}_2^T] = \arg \max_{\|\mathbf{u}_2\|=1, \mathbf{u}_2 \perp \mathbf{u}_1=0} \mathbf{u}_2^T \mathbf{S}_t \mathbf{u}_2, \\ &\vdots \\ \hat{\mathbf{u}}_r &= \arg \max_{\|\mathbf{u}_r\|=1, \mathbf{u}_r \perp \mathbf{u}_i=0, i \in [r-1]} \widehat{\text{Var}} [\mathcal{Y} \times_k \mathbf{u}_r^T] = \arg \max_{\|\mathbf{u}_r\|=1, \mathbf{u}_r \perp \mathbf{u}_i=0, i \in [r-1]} \mathbf{u}_r^T \mathbf{S}_t \mathbf{u}_r,\end{aligned}\quad (12)$$

其中 $\widehat{\text{Var}} [\mathcal{Y} \times_k \mathbf{u}_i^T]$ 是随机张量 $\mathcal{Y}$ 沿第 k 方向在方向 $\mathbf{u}_i$ 上的投影变异性, 此时投影为 $p-1$ 阶张量, 其变异性为各元素样本方差之和, 相应地,

$$\mathbf{S}_t = \sum_{j \neq k} \sum_{i_j=1}^{I_j} \left[\frac{1}{N} \sum_{n=1}^N \left(\mathbf{y}_{i_1 \dots i_{k-1} i_{k+1} \dots i_p}^{(n)} - \bar{\mathbf{y}}_{i_1 \dots i_{k-1} i_{k+1} \dots i_p} \right) \left(\mathbf{y}_{i_1 \dots i_{k-1} i_{k+1} \dots i_p}^{(n)} - \bar{\mathbf{y}}_{i_1 \dots i_{k-1} i_{k+1} \dots i_p} \right)^T \right] \quad (13)$$

表示第 k 方向的向量的“样本协方差矩阵”之和, 其中 $\mathbf{y}_{i_1 \dots i_{k-1} i_{k+1} \dots i_p}^{(n)}$, $\bar{\mathbf{y}}_{i_1 \dots i_{k-1} i_{k+1} \dots i_p}$ 表示第 n 个张量数据 $\mathcal{Y}^{(n)}$ 的沿第 k 方向的向量及其样本均值. 类似于 (2), 上述统计推断过程亦可用矩阵符号等价地表达为

$$\hat{\mathbf{U}} = \arg \max_{\mathbf{U}^T \mathbf{U} = \mathbf{I}_r} \widehat{\text{Var}} [\mathcal{Y} \times_k \mathbf{U}^T] = \arg \max_{\mathbf{U}^T \mathbf{U} = \mathbf{I}_r} \text{Tr} [\mathbf{U}^T \mathbf{S}_t \mathbf{U}]. \quad (14)$$

式 (17) 描述的是数据张量 $\mathcal{Y}$ 的所有 “ $\{\mathbf{y}_{i_1 \dots i_{k-1} i_{k+1} \dots i_p} : i_j \in [I_j]\}$ ” 同时沿一个方向投影而获得的“特征张量”并计算其变异性, 所以该投影本质上是数据张量的 $p-1$ 阶子张量 $\{\mathcal{Y}_{\dots i_k \dots}; k = 1, \dots, I_k\}$ 的一个线性组合, 它本质上反映的是各子张量之间的“张量水平”的线性关系和变异性, 部分地反映了各子张量内部元素之间的关系, 但没有反映张量 $\mathcal{Y}_{\dots i_k \dots}$ 内部各元素之间的关系. 类似地, 各方向 $\hat{\mathbf{u}}_i, i = 1, \dots, r$ 上的数据投影波动性 $\sigma_i^2 = \hat{\mathbf{u}}_i^T \mathbf{S}_t \hat{\mathbf{u}}_i$ 也常常从最大迅速下降, 即 $\sigma_1^2 \geq \sigma_2^2 \geq \dots \geq \sigma_r^2$, 且前几个方向占据主导. 在高阶张量情形, 前述性质依然是满足的.

性质4 给定 $\hat{\mathbf{U}}$ 满足式 (17), 则

$$\begin{aligned}\sum_{n=1}^N \|\mathcal{Y}_o^{(n)}\|_F^2 &= \sum_{n=1}^N \|\mathcal{Y}_o^{(n)} - \hat{\mathcal{V}}^{(n)} \times_k \hat{\mathbf{U}}^T\|_F^2 + \sum_{n=1}^N \|\hat{\mathcal{V}}^{(n)}\|_F^2 \\ &= \sum_{n=1}^N \|\mathcal{Y}_o^{(n)} - \hat{\mathcal{V}}^{(n)} \times_k \hat{\mathbf{U}}^T\|_F^2 + N \sum_{i=1}^r \sigma_i^2,\end{aligned}\quad (15)$$

其中 $\mathcal{Y}_o^{(n)} = \mathcal{Y}^{(n)} - \bar{\mathcal{Y}}$, $\hat{\mathcal{V}}^{(n)} = \mathcal{Y}_o^{(n)} \times_k \hat{\mathbf{U}}$.

等式 (25) 揭示了多样本张量 PCA 也可以等价地表达为一组中心化数据矩阵 $\mathcal{Y}_o^{(n)}$ 的张量近似优化问题:

$$(\hat{\mathcal{V}}^{(n)}, \hat{\mathbf{U}}) = \arg \min_{\mathcal{V}^{(n)}, \mathbf{U}^T \mathbf{U} = \mathbf{I}_r} \sum_{n=1}^N \|\mathcal{Y}_o^{(n)} - \hat{\mathcal{V}}^{(n)} \times_k \hat{\mathbf{U}}^T\|_F^2. \quad (16)$$

Lu 等于 2008 年在文献 [16] 中提出了“多线性主成分分析”(multilinear PCA, MPCA), 该方法试图沿张量的所有方向同时降维, 保证降维之后的“核心张量”在各个样本之间的差异性能达到最大. 按此思路, 可以获得沿每个方向相应的标准正交基矩阵 $\{U_k \in \mathbb{R}^{I_k \times r_k}, k = 1, 2, \dots, p\}$, 如下:

$$\begin{aligned} (\hat{U}_1, \hat{U}_2, \dots, \hat{U}_p) &= \arg \max_{U_k^T U_k = I_{r_k}, k \in [p]} \widehat{\text{Var}} [\mathcal{Y} \times_1 \hat{U}_1^T \times_2 \cdots \times_p \hat{U}_p^T] \\ &= \arg \max_{U_k^T U_k = I_{r_k}, k \in [p]} \sum_{n=1}^N \left\| (\mathcal{Y}^{(n)} - \bar{\mathcal{Y}}) \times_1 \hat{U}_1^T \times_2 \cdots \times_p \hat{U}_p^T \right\|_F^2. \end{aligned} \quad (17)$$

容易验证类似的性质依旧成立.

性质5 给定 $\hat{U}$ 满足式 (17), 则

$$\sum_{n=1}^N \left\| \mathcal{Y}_o^{(n)} \right\|_F^2 = \sum_{n=1}^N \left\| \mathcal{Y}_o^{(n)} - \hat{\mathcal{V}}^{(n)} \times_1 \hat{U}_1 \times_2 \cdots \times_p \hat{U}_p \right\|_F^2 + \sum_{n=1}^N \left\| \hat{\mathcal{V}}^{(n)} \right\|_F^2, \quad (18)$$

其中 $\mathcal{Y}_o^{(n)} = \mathcal{Y}^{(n)} - \bar{\mathcal{Y}}$, $\hat{\mathcal{V}}^{(n)} = \mathcal{Y}_o^{(n)} \times_1 \hat{U}_1^T \times_2 \cdots \times_p \hat{U}_p^T$.

式 (18) 揭示出式 (17) 的另一个优化表达式:

$$(\hat{U}_1, \hat{U}_2, \dots, \hat{U}_p) = \sum_{n=1}^N \left\| \mathcal{Y}_o^{(n)} - \hat{\mathcal{V}}^{(n)} \times_1 \hat{U}_1 \times_2 \cdots \times_p \hat{U}_p \right\|_F^2. \quad (19)$$

显然 MPCA 是 GPCA [13] 和 2DPCA [12] 的一个自然推广, 但其通过最大化投影方差这一基本思想发端在于 Yang 的 2DPCA 方法.

3 理解 PCA 的 3 个重要视角

3.1 从“勾股定理”看 PCA

对给定的向量空间、矩阵空间或者高阶张量空间, 相对于通常的内积定义即: 对应元素乘积的和, 将会很自然形成一个 Hilbert 空间. 从一阶到高阶 PCA, (2), (7), (10), (14) 和 (17) 分别满足各自 Hilbert 空间里的“勾股定理”(3), (8), (11), (15) 和 (18), 而直观地从信息分解的角度看都可以统一成以下形式:

$$\text{总变异} = \text{噪声造成的变异} + \text{信号造成的变异}. \quad (20)$$

上述等式左边表示数据中的“总变异”(total variation), 这些“变异性”既可能由某种感兴趣的信号或者感兴趣的结构造成, 也可能由其他不感兴趣的噪声或误差造成, 等式右边的分解可以很好地表达这两部分的“变异性”来源. 这种分解有助于深入解析变异性背后形成的原因并可能推动我们选择合适的 PCA 方法, 这将马上在下段以及后面两小节分别深入阐述.

首先从上述表达可以立即获得 PCA 两个表达的等价性. 如果从右侧第二项出发, 就是经典的“变异性最大化”准则, 这是原始的统计角度的定义, 具有很好的统计可解释性; 如果从右侧第一项出发, 就是“平方损失最小化”或者“误差平方和最小”准则. 两者的本质都是为了在给定的模型复杂度下, 达到最佳的“信号-噪声比”. 此处我们特别关注这种优化视角下的 PCA 表述, 虽遮蔽了原有的统计意义, 但其显著优势在于将各阶 PCA 均转化为最小二乘框架或者 L^2 损失函数下的近似问题. 下面以 Zou 等 [17] 提出的“稀疏 PCA”(sparse PCA) 为例展示该优化视角下的推广思路, 基于上述“勾股定

理”在向量数据 PCA 的表达式 (3), Zou 等将 L^2 损失项叠加所谓 “Elastic net” 惩罚项, 构造了如下优化目标函数:

$$\min_{\mathbf{A}, \mathbf{B}, \mathbf{A}^T \mathbf{A} = \mathbf{I}_{r \times r}} \left\{ \sum_{n=1}^N \left\| \mathbf{y}_o^{(n)} - \mathbf{A} \mathbf{B}^T \mathbf{y}_o^{(n)} \right\|_F^2 + \lambda_1 \sum_{j=1}^r \|\mathbf{b}_j\|^2 + \sum_{j=1}^r \lambda_{2,j} \|\mathbf{b}_j\|_1 \right\},$$

其中 $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_r] \in \mathbb{R}^{I \times r}$, $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_r] \in \mathbb{R}^{I \times r}$. 当然上述优化问题求解并不容易, 具体算法除了最早的论文 [17] 外, 还可参看 2015 年 Hastie 等的专著 [18] 第 207 页所提出的转化算法. 注意到当 $\mathbf{A} = \mathbf{B} = \mathbf{U}$ 且 $\lambda_1 = \lambda_{2,j} = 0$ 时就退化到 “勾股定理” 的 L^2 损失函数, 其最小化问题就是经典的向量 PCA 求解问题. 基于 L^2 损失, 我们还可以做类似的推广. 我们知道当数据中出现 “异常点” (outliers) 或者 “强噪声” (strong noise) 的条件下, L^2 损失常常都不是最佳选择, 从而可以通过改变损失函数来引入各种 PCA 的推广形式, 如稳健 PCA [19~24] 或者通过基于某种先验的 “结构信息”, 譬如添加正则项而转化成 “稀疏 PCA” [7, 17, 24~29] 等, 或者添加其他的结构约束如投影方向的非负性 [30] 而构建了 “非负 PCA” 方法等. 同时, 在高阶张量 PCA 场合也有类似的推广, 已有文献涉及这个领域比较少, 这将是本文作者今后的努力方向之一.

3.2 从张量的 Tucker 分解看 PCA

(1) 张量的 Tucker 分解

张量的 Tucker 分解最早由 Tucker [31] 于 1963 年首次提出, 此后被 Levin [32] 和 Tucker [33, 34] 做了重要改进, 其中 Tucker 于文献 [34] 中系统阐述了这类张量分解的基础理论与有效算法, 成为该领域早期引用最多的一篇文献. 具体而言, 给定的张量 $\mathcal{Y} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_p}$, 张量 $\mathcal{Y}$ 的 Tucker 分解指的是将其表达为以下形式:

$$\mathcal{Y} = \mathcal{V} \times_1 \mathbf{U}_1 \times_2 \dots \times_p \mathbf{U}_p, \quad (21)$$

其中 $\mathcal{V} \in \mathbb{R}^{J_1 \times \dots \times J_p}$, $\mathbf{U}_k \in \mathbb{R}^{I_k \times J_k}$, $k = 1, \dots, p$. 最近几十年, “Tucker 分解” 在不同条件下获得一些研究结论, 所以又被冠以其他名字, 譬如: 三方向因子分析 (three-mode factor analysis, 3MFA/Tucker3) [34]、三方向主成分分析 (three-mode PCA, 3MPCA) [35]、 N -方向主成分分析 (N-mode PCA) [36]、高阶奇异值分解 (higher-order SVD, HOSVD) [37]、 N -方向奇异值分解 (N-mode SVD) [38] 等等. 由于这是一个从最早开始的研究方向, 基于上述解的有关优化算法也得到比较系统的发展, 获得了比较成熟的结论 [35, 37, 39]. 当然, 实际中给定的张量数据 $\mathcal{Y}$ 带有误差, 常常需要寻找一个在最小二乘误差下的 “秩 $(r_1, \dots, r_p)$ 近似”, 即

$$\left(\hat{\mathcal{V}}, \hat{\mathbf{U}}_k, k \in [p] \right) = \arg \min_{\mathcal{V}, \mathbf{U}_k^T \mathbf{U}_k = \mathbf{I}_{r_k}, k \in [p]} \left\| \mathcal{Y} - \mathcal{V} \times_1 \mathbf{U}_1 \times_2 \dots \times_p \mathbf{U}_p \right\|_F^2, \quad (22)$$

其中 $\mathcal{V} \in \mathbb{R}^{r_1 \times \dots \times r_p}$, $\mathbf{U}_k \in \mathbb{R}^{r_k \times I_k}$, $k \in [p]$. 容易验证满足以下等式:

$$\|\mathcal{Y}\|_F^2 = \left\| \mathcal{Y} - \hat{\mathcal{V}} \times_1 \hat{\mathbf{U}}_1 \times_2 \dots \times_p \hat{\mathbf{U}}_p \right\|_F^2 + \left\| \hat{\mathcal{V}} \right\|_F^2. \quad (23)$$

当 $p = 1$ 张量 $\mathcal{Y}$ 退化为向量情形 $\mathbf{y} \in \mathbb{R}^I$, 上述近似分解是不存在的, 或者说向量数据不存在拟合的问题, 因为对任意给定秩 r , 总存在一个空间包含该数据或者说使得拟合误差为零, 这就是说单个向量数据的上述分解总是过拟合 (over-fitting) 的. 但是当 $p = 2$ 时, 上述分解就退化为矩阵的 “奇异值

分解”(singular value decomposition, SVD) 问题. 具体而言, 给定矩阵数据 $\mathbf{Y}$ 时, 通常采用以下的 “秩 r 矩阵近似” 或者 “奇异值分解近似” 问题:

$$(\hat{\mathbf{V}}, \hat{\mathbf{U}}_1, \hat{\mathbf{U}}_2) = \arg \min_{\mathbf{U}_i^T \mathbf{U}_i = \mathbf{I}_{r_i}, i=1,2} \|\mathbf{Y} - \mathbf{U}_1 \mathbf{V} \mathbf{U}_2^T\|_F^2, \quad (24)$$

其中 $\mathbf{V} \in \mathbb{R}^{r_1 \times r_2}$, $\mathbf{U}_k \in \mathbb{R}^{r_k \times I}$, $k=1,2$, 且 $r_1 = r_2 = r$. 自然地, 也会成立以下等式:

$$\|\mathbf{Y}\|_F^2 = \|\mathbf{Y} - \hat{\mathbf{U}}_1 \hat{\mathbf{V}} \hat{\mathbf{U}}_2^T\|_F^2 + \|\hat{\mathbf{V}}\|_F^2 = \|\mathbf{Y} - \hat{\mathbf{U}}_1 \hat{\mathbf{V}} \hat{\mathbf{U}}_2^T\|_F^2 + \|\hat{\mathbf{U}}_1^T \mathbf{Y} \hat{\mathbf{U}}_2\|_F^2,$$

其中 $\hat{\mathbf{V}} = \hat{\mathbf{U}}_1^T \mathbf{Y} \hat{\mathbf{U}}_2$. 当然还有看出上述的最优化问题等价于

$$\hat{\mathbf{A}} = \arg \min_{\text{rank}(\mathbf{A})=r} \|\mathbf{Y} - \mathbf{A}\|_F^2,$$

其中 $\hat{\mathbf{A}}$ 就是相应的 “奇异值分解近似”, 即满足 $\hat{\mathbf{A}} = \hat{\mathbf{U}}_1 \hat{\mathbf{V}} \hat{\mathbf{U}}_2^T$.

(2) PCA 的 “Tucker 分解近似” 表达

多个向量样本 PCA 等式 (3) 可以等价表达为矩阵形式:

$$\|\mathbf{Y}\|_F^2 = \|\mathbf{Y} - \hat{\mathbf{U}} \hat{\mathbf{V}}\|_F^2 + \|\hat{\mathbf{V}}\|_F^2,$$

其中 $\mathbf{Y} = (\mathbf{y}_o^{(1)}, \dots, \mathbf{y}_o^{(N)}) \in \mathbb{R}^{I \times N}$, $\hat{\mathbf{V}} = \hat{\mathbf{U}}^T \mathbf{Y}_o$, 从而可以将 PCA 表达成矩阵 $\mathbf{Y} \in \mathbb{R}^{I \times N}$ 的矩阵分解 (24) 的特例, 其中 $\mathbf{U}_1 = \mathbf{U}$ 且同时约束 $\mathbf{U}_2$ 为单位阵 $\mathbf{I}_N$. 类似地, 多个矩阵样本的式 (8) 可以表达为三阶张量的近似 Tucker 分解形式

$$\|\mathcal{Y}\|_F^2 = \|\mathcal{Y} - \hat{\mathcal{V}} \times_2 \hat{\mathbf{U}}\|_F^2 + \|\hat{\mathcal{V}}\|_F^2,$$

其中 $\mathcal{Y} \in \mathbb{R}^{I \times J \times N}$ 满足 $\mathbf{Y}_{::n} = \mathbf{Y}_o^{(n)}$, $\hat{\mathcal{V}} = \mathcal{Y} \times_2 \hat{\mathbf{U}}^T \in \mathbb{R}^{I \times r \times N}$ 满足 $\hat{\mathbf{V}}_{::n} = \hat{\mathbf{V}}^{(n)}$, 从而类似地可以将矩阵 PCA (10) 表达成三阶张量 $\mathcal{Y} \in \mathbb{R}^{I \times J \times N}$ 的近似 Tucker 分解 (23) 的特例, 其中 $\mathbf{U}_1 = \mathbf{U}$ 且同时约束 $\mathbf{U}_2 = \mathbf{I}_r, \mathbf{U}_3 = \mathbf{I}_N$; 而式 (11) 等价表达为

$$\|\mathcal{Y}\|_F^2 = \|\mathcal{Y} - \hat{\mathcal{V}} \times_1 \hat{\mathbf{U}}_1^T \times_2 \hat{\mathbf{U}}_2^T\|_F^2 + \|\hat{\mathcal{V}}\|_F^2,$$

其中 $\mathcal{Y} \in \mathbb{R}^{I \times J \times N}$ 满足 $\mathbf{Y}_{::n} = \mathbf{Y}^{(n)} - \bar{\mathbf{Y}}$, $\hat{\mathcal{V}} = \mathcal{Y} \times_1 \hat{\mathbf{U}}_1^T \times_2 \hat{\mathbf{U}}_2^T \in \mathbb{R}^{r_1 \times r_2 \times N}$ 满足 $\hat{\mathbf{V}}_{::n} = \hat{\mathbf{V}}^{(n)}$, 此时也可以把矩阵 PCA (10) 表达成三阶张量 $\mathcal{Y} \in \mathbb{R}^{I \times J \times N}$ 的近似 Tucker 分解形式 (23), 其中也需要约束 $\mathbf{U}_3 = \mathbf{I}_N$. 对更一般的高阶张量 PCA 而言上述关系也是成立的, 首先注意到式 (18) 可以表达为 $p+1$ 阶张量的近似 Tucker 分解形式

$$\|\mathcal{Y}\|_F^2 = \|\mathcal{Y} - \hat{\mathcal{V}} \times_k \hat{\mathbf{U}}\|_F^2 + \|\hat{\mathcal{V}}\|_F^2,$$

其中 $\mathcal{Y} \in \mathbb{R}^{I_1 \times \dots \times I_p \times N}$ 满足 $\mathcal{Y}_{:, \dots, n} = \mathcal{Y}_o^{(n)}$, $\hat{\mathcal{V}} = \mathcal{Y} \times_k \hat{\mathbf{U}}^T \in \mathbb{R}^{I_1 \times \dots \times I_{k-1} \times r \times I_{k+1} \times \dots \times I_p \times N}$ 满足 $\hat{\mathcal{V}}_{:, \dots, n} = \hat{\mathcal{V}}^{(n)}$, 此时张量 PCA 可以视为 $p+1$ 阶张量 $\mathcal{Y}$ 的近似 Tucker 分解 (23), 其中 $\mathbf{U}_k = \mathbf{U}$ 且设定 $\mathbf{U}_j = \mathbf{I}_{r_j}, j \neq k, j \in [p], \mathbf{U}_{p+1} = \mathbf{I}_N$; 对于张量 PCA 等式 (18) 而言, 其 $p+1$ 阶张量的等价表达也是类似的,

$$\|\mathcal{Y}\|_F^2 = \|\mathcal{Y} - \hat{\mathcal{V}} \times_1 \hat{\mathbf{U}}_1 \times_2 \dots \times_p \hat{\mathbf{U}}_p\|_F^2 + \|\hat{\mathcal{V}}\|_F^2, \quad (25)$$

其中 $\mathcal{Y} \in \mathbb{R}^{I_1 \times \cdots \times I_p \times N}$ 满足 $\mathcal{Y}_{:, \dots, n} = \mathcal{Y}_o^{(n)}$, $\hat{\mathcal{V}} = \mathcal{Y} \times_1 \hat{U}_1^T \times \cdots \times_p \hat{U}_p^T \in \mathbb{R}^{r_1 \times \cdots \times r_p \times N}$ 满足 $\hat{\mathcal{V}}_{:, \dots, n} = \hat{\mathcal{V}}^{(n)}$, 则此时的张量 PCA 也是式 (17) 的特例, 不过设定 $U_{p+1} = I_N$.

综上所述, 多样本 p 阶 PCA 可以表达成一个 $p+1$ 阶张量的近似 Tucker 分解问题, 需要首先沿“样本”方向中心化, 使得其本质就是“正交投影的中心二阶矩”尽可能大, 这与通常的 $p+1$ 阶张量数据分解是不同的, 后者常常是不会沿某一个特定方向中心化的.

(3) “Tucker 分解近似”的 PCA 方法

上文所述多样本 p 阶张量 PCA 可以视为一个 $p+1$ 阶张量的“Tucker 分解近似”, 这种视角为我们压缩一个高阶张量 $\mathbf{Y}$ 提供了一种新的可能途径. 譬如: 给定一个视频数据 $\mathbf{Y} \in \mathbb{R}^{H \times W \times T}$, 其中 $H \times W$ 表示每一帧的分辨率, T 表示视频的时间长度, 我们可以将其剖分成 T 帧, 从而视为 T 个矩阵数据, 按多样本 PCA 处理. 一般来说, 给定 p 阶张量数据, 首先选择一个特殊方向——具有某种类似的样本重复性, 以此方向将张量数据剖分成多个 $p-1$ 阶张量数据, 并视为多样本 $p-1$ 阶张量数据中心化, 展开多样本 $p-1$ 阶张量的 PCA 的分析方法.

同时, 对于一个 p 阶张量数据 $\mathcal{Y}$, 还可以考虑剖分成多个同阶张量, 展开 PCA 分析. 具体而言, 给定一个 p 阶张量, 如果在某一方向具有重复性, 但是相互之间又存在某种关联, 为了保持这种关联性, 最好不要进行降阶拆分, 则可以进行分段——保持“阶”不变, 只是沿某一方向等分, 从而可以转化成同阶的多样本 PCA. 譬如: 视频在时间方向各帧之间具有某种重复性——背景相似性, 但是沿着时间方向各帧又相互关联——特别是前景(运动物体)和阴影变化等是在时间尺度或者各帧之间连续变化的, 要保持或很好地抓住这种前景或阴影变化, 就不能简单地将视频降一阶来分割为很多帧. 一个折中的办法就是将视频分段但又不至于“降阶”从而保持三方向结构, 同时获得多样本数据.

3.3 从“深度”(depth)看 PCA

给定 p 阶张量 $\mathcal{Y} \in \mathbb{R}^{I_1 \times \cdots \times I_p}$, 我们可以将其视为一个 p 维回归模型的离散观测. 具体来说, 给定一个 p 维回归模型

$$y = f(x_1, x_2, \dots, x_p) + \varepsilon, \quad (26)$$

其中 y 为感兴趣的观测指标或响应变量(response variable), $x_k, k=1, \dots, p$ 为 p 个自变量或解释变量(explanatory variable), 它表示张量 $\mathcal{Y}$ 沿 p 个方向分别变化时的潜在因子(latent factor), 此处这些因子可能并不是真实存在的. 当对第 k 个解释变量观测 I_k 次, 即 $x_{ki_k}, i_k=1, 2, \dots, I_k, k=1, 2, \dots, p$, 则相应的响应变量 y 可观测到相应的观测值 $y_{i_1 i_2 \dots i_p}, i_k=1, 2, \dots, I_k, k=1, 2, \dots, p$, 形成一个数据张量 $\mathcal{Y} = (y_{i_1 i_2 \dots i_p})$.

二阶或二阶以上张量数据中, 感兴趣的结构常常具有“各向异性”(anisotropy), 即: 这种结构表现为感兴趣的指标 y 只依赖于一部分自变量的取值不同而发生显著改变, 不妨设为 $\{x_k, k \in S_0\}, S_0 \subseteq S = \{1, 2, \dots, p\}$, 而对剩余的变量 $\{x_k, k \in \overline{S_0}\}$ 的不同取值并不是太敏感, 从而在属于 S_0 的方向上总是呈现出大的数据变异性而在 $\overline{S_0}$ 方向上不会呈现出大的差异性. 譬如: 三维监控视频 $\mathcal{Y} \in \mathbb{R}^{H \times W \times T}$, 前景或运动物体常常随着时间 $t=1, 2, \dots, T$ 推移而在各帧之间剧烈变动, 表现为其灰度值 y 随时间具有较大的变化, 而在每一帧的空间位置具有较小的活动区域, 即不大依赖于 $(w, h) \in [W] \times [H]$ 的变化而改变, 此处 $S_0 = \{3\}$; 又如在数字图片中, 可能“灰度值”随着“行指标”的不同而改变较大, 而对于不同列“灰度值”相差不大, 此时 $S_0 = \{1\}$, 等等诸如此类的现象也可能在高阶推荐系统、高光谱等数据中存在, 我们称之为张量数据中的“各向异性”.

如果我们希望在“信息压缩”的过程中,尽可能保留这些“各向异性”的结构以达到好的恢复精度,就需要设计相应的压缩方法. 给定一组数字图片 $Y^{(n)} \in \mathbb{R}^{I \times J}, n = 1, 2, \dots, N$, 按照上一段记号如果 $S_0 = \{1\}$, 可以采用 Yang 等^[12]的 2DPCA 方法对矩阵右乘一个“基矩阵”, 按行投影或者说对列做线性组合以获得列之间的“相似性”或“相依性”, 但尽可能保持列的内部结构, 这正是我们想要的; 类似地, 如果 $S_0 = \{2\}$, 我们需左乘“基矩阵”进行压缩; 不管怎么样, GPCA 方法同时对矩阵左右两侧同时乘以“基矩阵”总是可以解决无论何种方向的“各向异性”的矩阵压缩问题. 然而, 当数据是高阶张量时, 情况却没有这么简单了, 事实上相对于矩阵数据而言高阶张量存在一些本质的不同. 譬如: 给定一组三阶视频张量数据 $\mathcal{Y}^{(n)} \in \mathbb{R}^{H \times W \times T}, n = 1, 2, \dots, N$, 且 $S_0 = \{3\}$, 沿某个方向, 譬如: 3 方向按高阶 PCA 方法 (17) 进行压缩, 其本质就是对各帧做线性组合, 将会利用各帧之间的“相似性”与“相依性”进行压缩, 但是尽可能保持各帧内部结构, 由于此时前景只占据空间上在较小的区域内, 从而留下大量与前景无关的区域冗余而无法压缩, 事实上各帧之间由于前景不停变化, 使得各帧之间具有显著差异也并不容易压缩 (除非对于恰好无前景的帧之间相似度较大而可以压缩), 类似的问题同样存在沿其他两个方向的高阶 PCA 压缩. 显然, 上述困难是根植于一阶到高阶 PCA 的构造基因中, 当然如果沿各方向同时进行压缩似乎也是很难解决该问题的, 这种猜测已经在我们的模拟试验中获得不断验证, 在此不再将具体的试验结果呈现.

上述高阶张量的压缩受限根源于沿 k 方向高阶 PCA (17) 的“先天不足”, 从张量展开角度可以描述如下: 首先对张量 $\mathcal{Y}$ 沿 k 方向进行“矩阵化” (matricization) 得到 $Y_{(k)} \in \mathbb{R}^{I_k \times (\prod_{j \neq k} I_j)}$, 然后对转化后的矩阵 $Y_{(k)}$ 采用 Yang 的 2DPCA 的做法, 即沿矩阵 $Y_{(k)}$ 左乘“基矩阵”进行列投影或者行线性组合. 针对“各向异性”问题, 譬如: 感兴趣的结构是依赖于方向 S_0 的, 我们考虑一个新的可能压缩方法: 首先沿 $\overline{S_0} : S_0$ 对 $\mathcal{Y}$ 做矩阵展开得到 $Y_{\overline{S_0}, S_0} \in \mathbb{R}^{\prod_{j \in \overline{S_0}} I_j \times (\prod_{j \in S_0} I_j)}$, 然后自然的也转化成了 Yang 的 2DPCA 的类似做法. 譬如: 上述的三阶视频张量数据满足 $S_0 = \{3\}$, 相应的 $\overline{S_0} = \{1, 2\}$, 那么按 $\{1, 2\} : \{3\}$ 展开后的矩阵的每一行就是沿时间方向的“条” (fibre), 按 Yang 的 2DPCA 左乘“基矩阵”意味着对各行作线性组合, 从而保持每行内部结构不变, 行之间利用其“相似性”或“相依性”进行压缩. 注意到高阶 PCA 从形式上沿袭 2DPCA 的思想, 相当于总是做 $\{k\} : \{j | j \neq k\}$ 展开, 其 $\overline{S_0} = \{k\}$ 的方向数为 1, 完全保持了 2DPCA 每次沿一个方向压缩的思想, 我们称它为深度为 1 的 PCA, 这就是说, 已有的从一阶到高阶的 PCA 都是深度为 1 的, 正如前面所分析, 它并不利于挖掘感兴趣的“各向异性”结构; 更好的做法是, 沿 $\overline{S_0} : S_0$ 展开, 按最佳的深度 $d = |\overline{S_0}|$ (depth) 展开. 这种思想我们已经在模拟试验进行多次验证, 取得了不错的压缩效果, 据此思想我们将进一步系统发展所谓“深度 PCA” (deep-PCA) 的方法, 有关算法、理论本文不再详细阐述.

4 结论

本文在信息压缩背景下概述了从一阶到高阶主成分的发展状况, 同时指出理解 PCA 的 3 个视角及其可能的推广研究方向. 第 1 个视角揭示了从一阶到高阶 PCA 的共性结构——“勾股定理”, 从而说明经典的“变异性最大”的统计思想等价于“最小二乘误差最小”方法, 在出现“异常点”或“强影响点”的条件下, 最小二乘损失可以推广到一般损失函数情形, 这可以视为一种“稳健 PCA”或“稀疏 PCA”的可能研究方向; 第 2 个视角揭示了 PCA 与张量分解之间的关系, 在合适的张量特征条件下, 张量可以按多样本 PCA 的方式进行分解; 第 3 个视角揭示了已有的从一阶到高阶 PCA 都可以认为是“深度”为 1 的, 在张量中存在“各向异性”结构的条件下, 经典 PCA 的表现并不是最佳的, 从而可以将其推广到一般的“深度”情形, 这已经作为另一项专门的工作中获得进一步研究.

参考文献

- 1 Pearson K. On lines and planes of closest fit to systems of points in space. *London Edinburgh Dublin Philos Mag J Sci*, 1901, 2: 559–572
- 2 Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol*, 1933, 24: 417–441
- 3 Anderson T W. Asymptotic theory for principal component analysis. *Ann Math Stat*, 1963, 34: 122–148
- 4 Johnstone I M. On the distribution of the largest eigenvalue in principal components analysis. *Ann Stat*, 2001, 29: 295–327
- 5 Baik J, Arous G B, Pécché S. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann Probab*, 2005, 33: 1643–1697
- 6 Paul D. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Stat Sin*, 2007, 17: 1617–1642
- 7 Johnstone I M, Lu A Y. On consistency and sparsity for principal components analysis in high dimensions. *J Am Stat Assoc*, 2009, 104: 682–693
- 8 Jung S, Marron J S. PCA consistency in high dimension, low sample size context. *Ann Stat*, 2009, 37: 4104–4130
- 9 Onatski A. Asymptotics of the principal components estimator of large factor models with weakly influential factors. *J Econom*, 2012, 168: 244–258
- 10 Shen D, Shen H, Zhu H, et al. The statistics and mathematics of high dimension low sample size asymptotics. *Stat Sin*, 2016, 26: 1747–1770
- 11 Wang W, Fan J. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Ann Stat*, 2017, 45: 1342–1374
- 12 Yang J, Zhang D, Frangi A F, et al. Two-dimensional pca: a new approach to appearance-based face representation and recognition. *IEEE Trans Pattern Anal Machine Intell*, 2004, 26: 131–137
- 13 Ye J P, Janardan R, Li Q. GPCA: an efficient dimension reduction scheme for image compression and retrieval. In: *Proceedings of the 10th ACM Sigkdd International Conference on Knowledge Discovery & Data Mining*, Seattle, 2004. 354–363
- 14 Cai D, He X, Han J. Subspace Learning Based on Tensor Analysis. UIUCDCS-R-2005-2572. 2005
- 15 Ye J. Generalized low rank approximations of matrices. *Mach Learn*, 2005, 61: 167–191
- 16 Lu H P, Plataniotis K N, Venetsanopoulos A N. MPCA: multilinear principal component analysis of tensor objects. *IEEE Trans Neural Netw*, 2008, 19: 18–39
- 17 Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. *J Comput Graphical Stat*, 2006, 15: 265–286
- 18 Hastie T, Tibshirani R, Wainwright M. *Statistical Learning with Sparsity the Lasso and Generalizations*. London: Chapman and Hall/CRC, 2015
- 19 Huber P J. Robust estimation of a location parameter. *Ann Math Stat*, 1964, 35: 73–101
- 20 Huber P J. Robust regression: asymptotics, conjectures and Monte Carlo. *Ann Stat*, 1973, 1: 799–821
- 21 Chen X R, Zhao L C. *M-method in Linear Model*. Shanghai: Shanghai Scientific & Technical Publishers, 1996
- 22 de Torre F. A framework for robust subspace learning. *Int J Comput Vision*, 2003, 54: 117–142
- 23 Candes E J, Li X, Ma Y, et al. Robust principal component analysis? *J ACM*, 2011, 58: 1–37
- 24 Nie F, Yuan J, Huang H. Optimal mean robust principal component analysis. In: *Proceedings of the 31st International Conference on Machine Learning*, Beijing, 2014
- 25 Vu V Q, Lei J. Minimax rates of estimation for sparse PCA in high dimensions. In: *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, La Palma, 2012. 1278–1286
- 26 Shen D, Shen H, Marron J S. Consistency of sparse PCA in high dimension, low sample size contexts. *J Multivar Anal*, 2013, 115: 317–333
- 27 Cai T T, Ma Z, Wu Y. Sparse PCA: optimal rates and adaptive estimation. *Ann Stat*, 2013, 41: 3074–3110
- 28 Nadler B. Finite sample approximation results for principal component analysis: a matrix perturbation approach. *Ann Stat*, 2008, 36: 2791–2817
- 29 Reiss M, Wahl M. Non-asymptotic upper bounds for the reconstruction error of PCA. 2016. ArXiv: 1609.03779
- 30 Montanari A, Richard E. Non-negative principal component analysis: message passing algorithms and sharp asymptotics. *IEEE Trans Inform Theor*, 2016, 62: 1458–1484
- 31 Tucker L R. *Implications of Factor Analysis of Three-way Matrices for Measurement of Change*. Wisconsin: University

- of Wisconsin Press, 1963
- 32 Levin J. Three-mode factor analysis. Dissertation for Ph.D. Degree. Urbana: University of Illinois, 1963
 - 33 Tucker L R. The Extension of Factor Analysis to Three-dimensional Matrices. New York: Springer-Verlag, 1964
 - 34 Tucker L R. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 1966, 31: 279–311
 - 35 Kroonenberg P M, de Leeuw J. Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 1980, 45: 69–97
 - 36 Kapteyn A, Neudecker H, Wansbeek T. An approach ton-mode components analysis. *Psychometrika*, 1986, 51: 269–275
 - 37 Lathauwer L D, Moor B D, Vandewalle J. A multilinear singular value decomposition. *SIAM J Matrix Anal Appl*, 2000, 21: 1253–1278
 - 38 Vasilescu M A O, Terzopoulos D. Multilinear analysis of image ensembles: tensor faces. In: *Proceedings of the 7th European Conference on Computer Vision*. Berlin: Springer, 2002. 447–460
 - 39 ten Berge J M F, de Leeuw J, Kroonenberg P M. Some additional results on principal components analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 1987, 52: 183–191

Information compression based on principal component analysis: from one-order to higher-order

Zhiming XIA^{1,2} & Zongben XU^{1*}

1. *School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China;*

2. *School of Mathematics, Northwest University, Xi'an 710027, China*

* Corresponding author. E-mail: statxzm@nwu.edu.cn

Abstract In this paper, a statistical technique of principal component analysis (PCA) from one-order to higher-order data under the background of information compression is summarized, its characteristics and limitations of each order PCA are revealed from three different perspectives, and some possible research directions are pointed out. Firstly, the technique and some existing developments are summarized by a similar structure, their intrinsic similar statistical significance is further analyzed, and the common structure—the Pythagorean theorem that shows two equivalent expressions of PCA—“maximizing variability” and “minimizing square loss” is shown. Secondly, this paper analyzes three important angles of PCA: the first one starts from the perspective of the Pythagorean theorem and further points out that the PCA can be extended to more general loss function—“robust” or “sparse” PCA; the second view reveals the relationship between tensor decomposition and PCA, which leads to a new idea of constructing tensor decomposition from the PCA perspective; the last view shows that higher-order PCA has a limitation to reveal anisotropic structure and further points out that a new method, “depth PCA,” can be used to conquer this limitation.

Keywords principle component analysis, information compression, high-order tensor, Tucker decomposition, anisotropy

Zhiming XIA was born in 1978. He obtained his Ph.D. degree in statistics from Northwest University, Xi'an, in 2009. He currently serves as a professor of statistics at Northwest University. His research interests include tensor data processing, change-point analysis, distributed data processing, statistical process control, and nonparametric statistics.

Zongben XU was born in 1955. He obtained his Ph.D. degree in mathematics from Xi'an Jiaotong University, Xi'an, in 1987. He currently serves as a professor of mathematics & computer science and a Ph.D. supervisor, chief scientist of National Basic Research Program of China, and director of the Institute for Information and System Sciences at Xi'an Jiaotong University, Xi'an. He is an academican of the Chinese Academy of Sciences and a member of the New York Academy of Sciences. His current research interests include nonlinear functional analysis, mathematical foundations of information technology, and computational intelligence.