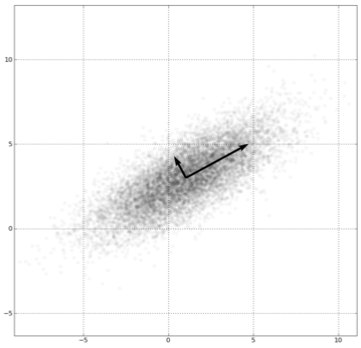
主成分分析

维基百科，自由的百科全书

在多元统计分析中，主成分分析（英语：**Principal components analysis**，**PCA**）是一种统计分析、简化数据集的方法。它利用正交变换来对一系列可能相关的变量的观测值进行线性变换，从而投影为一系列线性不相关变量的值，这些不相关变量称为主成分（Principal Components）。具体地，主成分可以看做一个线性方程，其包含一系列线性系数来指示投影方向。PCA对原始数据的正则化或预处理敏感（相对缩放）。



主成分分析实例：一个平均值为(1, 3)、标准差在(0.878, 0.478)方向上为3、在其正交方向为1的高斯分布。这里以黑色显示的两个向量是这个分布的协方差矩阵的特征向量，其长度按对应的特征值之平方根为比例，并且移动到以原分布的平均值为原点。

基本思想：

- 将坐标轴中心移到数据的中心，然后旋转坐标轴，使得数据在C1轴上的方差最大，即全部n个数据个体在该方向上的投影最为分散。意味着更多的信息被保留下来。C1成为**第一主成分**。
- C2**第二主成分**：找一个C2，使得C2与C1的协方差（相关系数）为0，以免与C1信息重叠，并且使数据在该方向的方差尽量最大。
- 以此类推，找到第三主成分，第四主成分。。。。第p个主成分。p个随机变量可以有p个主成分[1]。

主成分分析经常用于减少数据集的维数，同时保持数据集中的对方差贡献最大的特征。这是通过保留低阶主成分，忽略高阶主成分做到的。这样低阶成分往往能够保留住数据的最重要方面。但是，这也不是一定的，要视具体应用而定。由于主成分分析依赖所给数据，所以数据的准确性对分析结果影响很大。

主成分分析由卡尔·皮尔逊于1901年发明[2]，用于分析数据及建立数理模型，在原理上与主轴定理相似。之后在1930年左右由哈罗德·霍特林独立发展并命名。依据应用领域的不同，在信号处理中它也叫做离散K-L 转换（discrete Karhunen–Loève transform (KLT)）。其方法主要是通过对协方差矩阵进行特征分解[3]，以得出数据的主成分（即特征向量）与它们的权值（即特征值[4]）。PCA是最简单的以特征量分析多元统计分布的方法。其结果可以理解为对原数据中的方差做出解释：哪一个方向上的数据值对方差的影响最大？换而言之，PCA提供了一种降低数据维度的有效办法；如果分析者在原数据中除掉最小的特征值所对应的成分，那么所得的低维度数据必定是最优化的（也即，这样降低维度必定是失去讯息最少的方法）。主成分分析在分析复杂数据时尤为有用，比如人脸识别。

PCA是最简单的以特征量分析多元统计分布的方法。通常情况下，这种运算可以被看作是揭露数据的内部结构，从而更好的解释数据的变量的方法。如果一个多元数据集能够在一个高维数据空间坐标系中被显现出来，那么PCA就能够提供一幅比较低维度的图像，这幅图像即为在讯息最多的点上原对象的一个‘投影’。这样就可以利用少量的主成分使得数据的维度降低了。

PCA跟因子分析密切相关，并且已经有很多混合这两种分析的统计包。而真实要素分析则是假定底层结构，求得微小差异矩阵的特征向量。

目录

数学定义

PCA的数学定义是：一个正交化线性变换，把数据变换到一个新的坐标系统中，使得这一数据的任何投影的第一大方差在第一个坐标（称为第一主成分）上，第二大方差在第二个坐标（第二主成分）上，依次类推[5]。

定义一个$n \times m$的矩阵, $\mathbf{X}^{\mathrm{T}}$为去平均值（以平均值为中心移动至原点）的数据，其行为数据样本，列为数据类别（注意，这里定义的是$\mathbf{X}^{\mathrm{T}}$ 而不是$\mathbf{X}$）。则$\mathbf{X}$的奇异值分解为$\mathbf{X} = \mathbf{W}\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}}$，其中$m \times m$矩阵$\mathbf{W}$是$\mathbf{X}\mathbf{X}^{\mathrm{T}}$的特征向量矩阵， $\mathbf{\Sigma}$是$m \times n$的非负矩形对角矩阵，V是$n \times n$的$\mathbf{X}^{\mathbf{T}}\mathbf{X}$的特征向量矩阵。据此，

$$\begin{aligned}\boldsymbol{Y}^\top &= \boldsymbol{X}^\top \boldsymbol{W} \\ &= \boldsymbol{V}\boldsymbol{\Sigma}^\top \boldsymbol{W}^\top \boldsymbol{W} \\ &= \boldsymbol{V}\boldsymbol{\Sigma}^\top\end{aligned}$$

当 $m < n - 1$时，$\mathbf{V}$ 在通常情况下不是唯一定义的，而$\mathbf{Y}$ 则是唯一定义的。$\mathbf{W}$ 是一个正交矩阵，$\mathbf{Y}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}$=$\mathbf{X}^{\mathrm{T}}$，且$\mathbf{Y}^{\mathrm{T}}$的第一列由第一主成分组成，第二列由第二主成分组成，依此类推。

为了得到一种降低数据维度的有效办法，我们可以利用$\mathbf{W}_{\mathrm{L}}$把 $\mathbf{X}$ 映射到一个只应用前面L个向量的低维空间中去：

$$\mathbf{Y} = {\mathbf{W}_{\mathbf{L}}}^{\top}\mathbf{X} = \mathbf{\Sigma}_{\mathbf{L}}\mathbf{V}^{\top}$$

其中$\mathbf{\Sigma_L} = \mathbf{I}_{L \times m}\mathbf{\Sigma}$，且$\mathbf{I}_{L \times m}$为$L \times m$的单位矩阵。

$\mathbf{X}$ 的单向量矩阵$\mathbf{W}$相当于协方差矩阵的特征向量 $\mathbf{C} = \mathbf{X}\,\mathbf{X}^{\mathrm{T}}$,

$$\mathbf{X}\mathbf{X}^{\top} = \mathbf{W}\mathbf{\Sigma}\mathbf{\Sigma}^{\top}\mathbf{W}^{\top}$$

在欧几里得空间给定一组点数，第一主成分对应于通过多维空间平均点的一条线，同时保证各个点到这条直线距离的平方和最小。去除掉第一主成分后，用同样的方法得到第二主成分。依此类推。在$\mathbf{\Sigma}$中的奇异值均为矩阵$\mathbf{X}\mathbf{X}^{\mathrm{T}}$的特征值的平方根。每一个特征值都与跟它们相关的方差是成正比的，而且所有特征值的总和等于所有点到它们的多维空间平均点距离的平方和。PCA提供了一种降低维度的有效办法，本质上，它利用正交变换将围绕平均点的点集中尽可能多的变量投影到第一维中去，因此，降低维度必定是失去讯息最少的方法。PCA具有保持子空间拥有最大方差的最优正交变换的特性。然而，当与离散余弦变换相比时，它需要更大的计算需求代价。非线性降维技术相对于PCA来说则需要更高的计算要求。

PCA对变量的缩放很敏感。如果我们只有两个变量，而且它们具有相同的样本方差，并且成正相关，那么PCA将涉及两个变量的主成分的旋转。但是，如果把第一个变量的所有值都乘以100，那么第一主成分就几乎和这个变量一样，另一个变量只提供了很小的贡献，第二主成分也将和第二个原始变量几乎一致。这就意味着当不同的变量代表不同的单位（如温度和质量）时，PCA是一种比较武断的分析方法。但是在Pearson的题为 "On Lines and Planes of Closest Fit to Systems of Points in Space"的原始文件里，是假设在欧几里得空间里不考虑这些。一种使PCA不那么武断的方法是使用变量缩放以得到单位方差。

讨论

通常，为了确保第一主成分描述的是最大方差的方向，我们会使用平均减法进行主成分分析。如果不执行平均减法，第一主成分有可能或多或少的对应于数据的平均值。另外，为了找到近似数据的最小均方误差，我们必须选取一个零均值[6]。

假设零经验均值，数据集 $\mathbf{X}$ 的主成分w_1可以被定义为：

$$\mathbf{w}_1 = \underset{\|\mathbf{w}\|=1}{\arg\max}\ \mathrm{Var}\{\mathbf{w}^\top \mathbf{X}\} = \underset{\|\mathbf{w}\|=1}{\arg\max}\, E\left\{\left(\mathbf{w}^\top \mathbf{X}\right)^2\right\}$$

为了得到第 k个主成分，必须先从$\mathbf{X}$中减去前面的 $k-1$ 个主成分：

$$\hat{\mathbf{X}}_{k-1} = \mathbf{X} - \sum_{i=1}^{k-1} \mathbf{w}_i \mathbf{w}_i^\top \mathbf{X}$$

然后把求得的第k个主成分带入数据集，得到新的数据集，继续寻找主成分。

$$\mathbf{w}_k = \underset{\|\mathbf{w}\|=1}{\arg\max}\, E\left\{\left(\mathbf{w}^\top \hat{\mathbf{X}}_{k-1}\right)^2\right\}.$$

PCA相当于在气象学中使用的经验正交函数（EOF）,同时也类似于一个线性隐层神经网络。 隐含层 K 个神经元的权重向量收敛后，将形成一个由前 K 个主成分跨越空间的基础。但是与PCA不同的是，这种技术并不一定会产生正交向量。

PCA是一种很流行且主要的的模式识别技术。然而，它并不能最优化类别可分离性[7] 。另一种不考虑这一点的方法是线性判别分析。

符号和缩写表

Symbol符号	Meaning意义	Dimensions尺寸	Indices指数
$\mathbf{X} = \{X[m,n]\}$	由所有数据向量集组成的数据矩阵，一列代表一个向量	$M \times N$	$m = 1 \dots M$ $n = 1 \dots N$
N	数据集中列向量的个数	1×1	*标量*
M	每个列向量的元素个数	1×1	*标量*
L	子空间的维数, $1 \leq L \leq M$	1×1	*标量*
$\mathbf{u} = \{u[m]\}$	经验均值向量	$M \times 1$	$m = 1 \dots M$
$\mathbf{s} = \{s[m]\}$	经验标准方差向量	$M \times 1$	$m = 1 \dots M$
$\mathbf{h} = \{h[n]\}$	所有的单位向量	$1 \times N$	$n = 1 \dots N$
$\mathbf{B} = \{B[m,n]\}$	对均值的偏离向量	$M \times N$	$m = 1 \dots M$ $n = 1 \dots N$
$\mathbf{Z} = \{Z[m,n]\}$	Z–分数，利用均值和标准差计算得到	$M \times N$	$m = 1 \dots M$ $n = 1 \dots N$
$\mathbf{C} = \{C[p,q]\}$	协方差矩阵	$M \times M$	$p = 1 \dots M$ $q = 1 \dots M$
$\mathbf{R} = \{R[p,q]\}$	相关矩阵	$M \times M$	$p = 1 \dots M$ $q = 1 \dots M$
$\mathbf{V} = \{V[p,q]\}$	C的所有特征向量集	$M \times M$	$p = 1 \dots M$ $q = 1 \dots M$
$\mathbf{D} = \{D[p,q]\}$	主对角线为特征值的对角矩阵	$M \times M$	$p = 1 \dots M$ $q = 1 \dots M$
$\mathbf{W} = \{W[p,q]\}$	基向量矩阵	$M \times L$	$p = 1 \dots M$ $q = 1 \dots L$
$\mathbf{Y} = \{Y[m,n]\}$	X 和W矩阵的投影矩阵	$L \times N$	$m = 1 \dots L$ $n = 1 \dots N$

主成分分析的属性和限制

如上所述，主成分分析的结果依赖于变量的缩放。

主成分分析的适用性受到由它的派生物产生的某些假设[8] 的限制。

主成分分析和信息理论

通过使用降维来保存大部分数据信息的主成分分析的观点是不正确的。确实如此，当没有任何假设信息的信号模型时，主成分分析在降维的同时并不能保证信息的不丢失，其中信息是由香农熵[9]来衡量的。 基于假设得 $\mathbf{x} = \mathbf{s} + \mathbf{n}$ 也就是说，向量 x 是含有信息的目标信号 s 和噪声信号 n 之和，从信息论角度考虑主成分分析在降维上是最优的。

特别地，Linsker证明了如果 s 是高斯分布，且 n 是 与密度矩阵相应的协方差矩阵的高斯噪声，

使用统计方法计算PCA

以下是使用统计方法计算PCA的详细说明。但是请注意，如果利用奇异值分解（使用标准的软件）效果会更好。

我们的目标是把一个给定的具有 M 维的数据集$\mathbf{X}$ 变换成具有较小维度 L的数据集$\mathbf{Y}$。现在要求的就是矩阵$\mathbf{Y}$，$\mathbf{Y}$是矩阵$\mathbf{X}$ Karhunen–Loève变换。:$\mathbf{Y} = \mathbb{KLT}\{\mathbf{X}\}$

组织数据集

假设有一组 M 个变量的观察数据，我们的目的是减少数据，使得能够用L 个向量来描述每个观察值，$L < M$。进一步假设，该数据被整理成一组具有N个向量的数据集，其中每个向量都代表M 个变量的单一观察数据。

- $\mathbf{x}_1 \dots \mathbf{x}_N$为列向量，其中每个列向量有$M$ 行。
- 将列向量放入$M \times N$的单矩阵X 里。

计算经验均值

- 对每一维m = 1, ..., M计算经验均值
- 将计算得到的均值放入一个 $M \times 1$维的经验均值向量u中

$$u[m] = \frac{1}{N} \sum_{n=1}^{N} X[m,n]$$

计算平均偏差

对于在最大限度地减少近似数据的均方误差的基础上找到一个主成分来说，均值减去法是该解决方案的不可或缺的组成部分[10] 。因此，我们继续如下步骤：

- 从数据矩阵X的每一列中减去经验均值向量 u
- 将平均减去过的数据存储在$M \times N$矩阵B中

$$\mathbf{B} = \mathbf{X} - \mathbf{uh}$$
where h is a $1 \times N$ row vector of all 1s:

其中$\mathbf{h}$是一个全 1s:的$1 \times N$ 的行向量

$$h[n] = 1 \qquad \text{for } n = 1, \dots, N$$

求协方差矩阵

- 从矩阵B 中找到$M \times M$ 的经验协方差矩阵C

$$\mathbf{C} = \mathbb{E}\left[\mathbf{B} \otimes \mathbf{B}\right] = \mathbb{E}\left[\mathbf{B} \cdot \mathbf{B}^*\right] = \frac{1}{N-1} \sum \mathbf{B} \cdot \mathbf{B}^*$$

其中 $\mathbb{E}$为期望值

⊗是最外层运算符

$*$ 是共轭转置运算符。

请注意，如果B完全由实数组成，那么共轭转置与正常的转置一样。

- 为什么是N–1,而不是N，Bessel's correction[11]给出了解释

查找协方差矩阵的特征值和特征向量

- 计算矩阵C 的特征向量

$$\mathbf{V}^{-1}\mathbf{C}\mathbf{V} = \mathbf{D}$$

其中，D 是C的特征值对角矩阵，这一步通常会涉及到使用基于计算机的计算特征值和特征向量的算法。在很多矩阵代数系统中这些算法都是现成可用的，如R语言，MATLAB,[12][13] Mathematica,[14] SciPy, IDL(交互式数据语言), 或者GNU Octave以及OpenCV。

- 矩阵D为$M \times M$的对角矩阵
- 各个特征值和特征向量都是配对的，m个特征值对应m个特征向量。

相关源代码

- Cornell Spectrum Imager (https://code.google.com/p/cornell-spectrum-imager/wiki/Home) – An open-source toolset built on ImageJ. Enables quick easy PCA analysis for 3D datacubes.
- imDEV (https://web.archive.org/web/20120502214751/http://sourceforge.net/apps/mediawiki/imdev/index.php?title=Main_Page) Free Excel addin to calculate principal components using R package pcaMethods (http://www.bioconductor.org/packages/1.9/bioc/html/pcaMethods.html).
- "ViSta: The Visual Statistics System" (https://web.archive.org/web/20090323151854/http://www.mdp.edu.ar/psicologia/vista/vista.htm) a free software that provides principal components analysis, simple and multiple correspondence analysis.
- "Spectramap" (http://www.coloritto.com) is software to create a biplot using principal components analysis, correspondence analysis or spectral map analysis.
- XLSTAT is a statistical and multivariate analysis software including Principal Component Analysis among other multivariate tools.
- FinMath (https://rtmath.net/products/finmath/), a .NET numerical library containing an implementation of PCA.
- The Unscrambler is a multivariate analysis software enabling Principal Component Analysis (PCA) with PCA Projection.
- Computer Vision Library (http://sourceforge.net/projects/opencvlibrary/)
- In the MATLAB Statistics Toolbox, the functions `princomp` and `wmspca` give the principal components, while the function `pcares` gives the residuals and reconstructed matrix for a low-rank PCA approximation. Here is a link to a MATLAB implementation of PCA `PcaPress` (http://www.utdallas.edu/~herve/abdi-PCA4Wiley.zip) .
- In the NAG Library, principal components analysis is implemented via the `g03aa` routine (available in both the Fortran[15] and the C[16] versions of the Library).
- NMath, a proprietary numerical library containing PCA for the .NET Framework.
- in Octave, a free software computational environment mostly compatible with MATLAB, the function `princomp` (http://octave.sourceforge.net/statistics/function/princomp.html) gives the principal component.
- in the free statistical package R, the functions `princomp` (http://stat.ethz.ch/R-manual/R-patched/library/stats/html/princomp.html) and `prcomp` (http://stat.ethz.ch/R-manual/R-patched/library/stats/html/prcomp.html) can be used for principal component analysis; `prcomp` uses singular value decomposition which generally gives better numerical accuracy. Recently there has been an explosion in implementations of principal component analysis in various R packages, generally in packages for specific purposes. For a more complete list, see here: [1] (http://cran.r-project.org/web/views/Multivariate.html).
- In *XLMiner*, the Principal Components tab can be used for principal component analysis.
- In IDL, the principal components can be calculated using the function `pcomp`.

- Weka computes principal components (javadoc (https://web.archive.org/web/20120411235501/http://weka.sourceforge.net/doc/weka/attributeSelection/PrincipalComponents.html)).
- Software for analyzing multivariate data with instant response using PCA (http://www.qlucore.com)
- Orange (software) supports PCA through its Linear Projection widget.
- A version of PCA adapted for population genetics analysis can be found in the suite EIGENSOFT (https://web.archive.org/web/20120217121739/http://genepath.med.harvard.edu/~reich/Software.htm).
- PCA can also be performed by the statistical software Partek Genomics Suite (https://web.archive.org/web/20120614230058/http://www.partek.com/partekgs), developed by Partek (https://www.webcitation.org/5ElLwqoZN?url=http://www.partek.com/).

参见

- Correspondence analysis
- Canonical correlation
- CUR matrix approximation (can replace of low-rank SVD approximation)
- Detrended correspondence analysis
- Dynamic mode decomposition
- 特征脸(Eigenface)
- 多线性主成分分析(Multilinear PCA)
- Geometric data analysis
- Factorial code
- 独立成分分析
- 核主成分分析
- 图模式
- 马尔可夫链
- 马尔可夫逻辑网络
- 矩阵分解
- Nonlinear dimensionality reduction
- Oja's rule
- Point distribution model (PCA applied to morphometry and computer vision)
- Principal component regression
- Singular spectrum analysis
- 奇异值分解
- Sparse PCA
- 变换编码
- 最小二乘法
- Low-rank approximation

注释

1. 主成分分析（principal components analysis, PCA）——无监督学习.
2. Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space (PDF). Philosophical Magazine. 1901, **2** (6): 559–572. （原始内容 (PDF)存档于2013-10-20）.
3. Abdi. H., & Williams, L.J. Principal component analysis.. Wiley Interdisciplinary Reviews: Computational Statistics,. 2010, **2**: 433–459.
4. Shaw P.J.A. (2003) *Multivariate statistics for the Environmental Sciences*, Hodder-Arnold. ISBN 978-0-340-80763-7.
5. Jolliffe I.T. Principal Component Analysis (http://www.springer.com/west/home/new+%26+forthcoming+titles+%28default%29?SGWID=4-40356-22-2285433-0), Series: Springer Series in Statistics (http://www.springer.com/west/home/statistics/statistical+theory+and+methods?SGWID=4-10129-69-173621571-0), 2nd ed., Springer, NY, 2002, XXIX, 487 p. 28 illus. ISBN 978-0-387-95442-4
6. A. A. Miranda, Y. A. Le Borgne, and G. Bontempi. New Routes from Minimal Approximation Error to Principal Components (http://www.ulb.ac.be/di/map/yleborgn/pub/NPL_PCA_07.pdf), Volume 27, Number 3 / June, 2008, Neural Processing Letters, Springer
7. Fukunaga, Keinosuke. Introduction to Statistical Pattern Recognition. Elsevier. 1990. ISBN 0122698517.

8. Jonathon Shlens, A Tutorial on Principal Component Analysis. (http://www.snl.salk.edu/~shlens/pca.pdf) 互联网档案馆的存档 (https://web.archive.org/web/20120302185926/http://www.snl.salk.edu/~shlens/pca.pdf)，存档日期2012-03-02.
9. Geiger, Bernhard; Kubin, Gernot (Sep 2012), Relative Information Loss in the PCA (http://arxiv.org/abs/1204.0429)
10. A.A. Miranda, Y.-A. Le Borgne, and G. Bontempi. New Routes from Minimal Approximation Error to Principal Components (http://www.ulb.ac.be/di/map/yleborgn/pub/NPL_PCA_07.pdf), Volume 27, Number 3 / June, 2008, Neural Processing Letters, Springer
11. Bessel's correction (https://en.wikipedia.org/wiki/Bessel%27s_correction) Bessel's correction
12. eig function (http://www.mathworks.com/access/helpdesk/help/techdoc/ref/eig.html#998306) Matlab documentation
13. MATLAB PCA-based Face recognition software (http://www.mathworks.com/matlabcentral/fileexchange/24634)
14. Eigenvalues function (http://reference.wolfram.com/mathematica/ref/Eigenvalues.html) Mathematica documentation
15. The Numerical Algorithms Group. NAG Library Routine Document: nagf_mv_prin_comp (g03aaf) (PDF). NAG Library Manual, Mark 23. [2012-02-16].
16. The Numerical Algorithms Group. NAG Library Routine Document: nag_mv_prin_comp (g03aac) (PDF). NAG Library Manual, Mark 9. [2012-02-16]. （原始内容 (PDF)存档于2011-11-24）.

参考

- Jolliffe, I. T. Principal Component Analysis. Springer-Verlag. 1986: 487. ISBN 978-0-387-95442-4. doi:10.1007/b98835.