doi: 10.7690/bgzdh.2014.09.016

主成分分析法及其在数据降噪中的应用

周宪英[1]，高成文[2]，曹建华[2]
（1. 中国人民解放军 92941 部队 96 分队，辽宁 葫芦岛 125001；
2. 中国人民解放军 92853 部队 4 分队，辽宁 兴城 125106）

摘要：为从含有噪声的采集信号中提取有用信号，确保飞行器试验结果数据的准确性，提出采用主成分分析提取有用信号的方法。阐述主成分分析的基本原理，分析主成分分析与奇异值分解 SVD 分析的区别与联系，给出采用 Hankel 矩阵和采用不重复排列矩阵的主成分对单列信号进行降噪处理的方法，并对无趋势信号、有趋势项信号和含冲击成分示例信号进行降噪设计。结果表明：主成分分析对无趋势信号、有趋势项信号具有很好的去除白噪声的效果，但不适用于含冲击成分信号的降噪，该方法可为相关领域信号分析提供参考。

关键词：主成分分析；SVD；白噪声；降噪
中图分类号：TJ04　　**文献标志码：**A

Principal Component Analysis Method and Its Application in Data Noise Reducing

Zhou Xianying[1], Gao Chengwen[2], Cao Jianhua[2]
(1. *No*. 96 *Team*, *No*. 92941 *Unit of PLA*, *Huludao* 125001, *China*;
2. *No*. 4 *Team*, *No*. 92853 *Unit of PLA*, *Xingcheng* 125106, *China*)

Abstract: For acquiring useful signal from acquired signal with noise, ensuring the results data accuracy in the aircraft test, the principal component analysis is proposed to extract the useful signal. Firstly, the fundamental principle of principal component analysis is discussed, and then its relation with singular value decomposition (SVD) is illustrated. Put forward 2 methods using principal component, one uses Hankel matrix and the other uses none repeated matrix, in the noise reduction for single queue signal. 3 types of signals are taken as inputs for the verify simulation and discussion, which are the signal with no trend, the signal with trend and the signal with an impact component. Results show that this method works well in white noise reducing with signal with no trend and signal with trend. It works not so well for signal containing impact component. This method can be referenced for signal processing engineers.

Keywords: principal component analysis; SVD; white noise; de-noising

0　引言

作为基础的数学分析方法，主成分分析的应用十分广泛，比如人口统计学、经济学、数学建模和数理分析等，是一种常用多变量分析方法。已经有相关文献表明它可以对多维数据降噪，找出信号的关键成分[1-2]。对于含有白噪声的单维信号降噪存在很多实际应用场合，目前方法还极其有限[3]。主成分分析已经成为商业软件的一个信号处理函数[4]。笔者主要在研究主成分分析基本原理的基础上，通过对无趋势信号、有趋势项信号和含冲击成分信号的降噪设计，探寻其在数据降噪等方面的可能应用。该方法可为相关信号分析提供参考。

1　主成分分析基本原理[5-8]

1.1　方法的基本思路

n 个样品在二维空间中的分布大致为一个椭圆，每个样品有 2 个变量，如图 1。

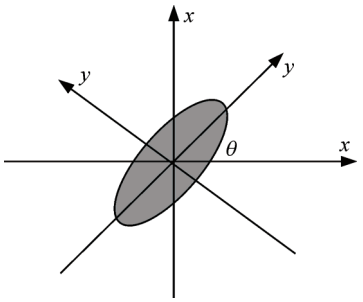


图 1　主成分分析几何解释

若将坐标系进行正交旋转一个角度 θ，使其椭圆长轴方向取坐标 y_1，在椭圆短轴方向取坐标 y_2，旋转公式为

$$\boldsymbol{Y}=\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \end{bmatrix}^{\mathrm{T}}= \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \boldsymbol{X}\boldsymbol{U}^{\mathrm{T}} \quad (1)$$

其中 $\boldsymbol{U}$ 为坐标旋转变换矩阵，它是正交矩阵，即有

收稿日期：2014-03-16；修回日期：2014-04-28
作者简介：周宪英(1966—)，女，天津人，工学学士，高级工程师，从事飞行器试验遥测数据处理方法研究。

$\boldsymbol{U}^{\mathrm{T}}=\boldsymbol{U}^{-1}$，$\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}=1$。经旋转变换后新坐标有如下性质：

1) n 个点的坐标与 y_1 和 y_2 的相关几乎为零。

2) 二维平面上的 n 个点的方差大部分都归结为 y_1 轴上，而 y_2 轴上的方差较小。

y_1 和 y_2 称为原始变量 x_1 和 x_2 的综合变量。由于 n 个点在 y_1 轴上的方差最大，因而将二维空间的点用在 y_1 轴上的一维综合变量来代替，所损失的信息量最小，由此称 y_1 轴为第一主成分。如果第一主成分不足以代表原来 m 个变量的信息，再考虑选取 y_2 即第二个线性组合，并且 y_1 已有的信息就不需要再出现在 y_2 中，用数学语言表达就是要求 $\mathrm{Cov}(y_1,y_2)=0$，y_2 轴上有较小的方差，称它为第二主成分。对于 n 维多坐标系，最多有 n 个主成分[9]。

1.2 主成分分析计算步骤[10]

主成分处理步骤如下：

1) 采集数据形成 $n\times m$ 的矩阵。m 为观测变量个数，n 为采样点个数。

2) 在每个观测变量(矩阵行向量)上减去该观测变量的平均值得到矩阵 $\boldsymbol{X}$。

3) 对 $\boldsymbol{X}$ 的协方差阵 $\boldsymbol{C}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 进行特征分解：

$$\underset{m\times m}{\boldsymbol{C}}=\underset{m\times n}{\boldsymbol{X}^{\mathrm{T}}}\ \underset{n\times m}{\boldsymbol{X}}=\underset{m\times m}{\boldsymbol{U}}\ \underset{m\times m}{\boldsymbol{\Lambda}}\ \underset{m\times m}{\boldsymbol{U}^{\mathrm{T}}} \tag{2}$$

式中：$\boldsymbol{\Lambda}$ 是对角阵，$\boldsymbol{\Lambda}=\mathrm{Diag}[\lambda_1,\lambda_2,\cdots,\lambda_m]$，各 λ_i 为 $\boldsymbol{C}$ 的特征根，$\boldsymbol{U}$ 为特征矩阵，它的各列 $u_1,u_2,\cdots,u_m$ 为特征矢量。求出特征根后和特征矩阵后，对特征根进行重新排列，使得 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m$，特征矢量 $u_1,u_2,\cdots,u_m$ 进行相应的交换。

4) 把 $\boldsymbol{U}^{\mathrm{T}}$ 后乘到数据阵 $\boldsymbol{X}$ 上，得 $\underset{n\times m}{\boldsymbol{Y}}=\underset{n\times m}{\boldsymbol{X}}\ \underset{m\times m}{\boldsymbol{U}^{\mathrm{T}}}$。

$\boldsymbol{Y}$ 的各列即为 $\boldsymbol{X}$ 的主分量，他们在 Y 中是依能量大小排列。

$\boldsymbol{Y}$ 的各列的方差是递减的，包含的信息量也是递减的，所以实际分析时，一般不是选取 m 个主成分，而是根据各个主成分累计贡献率的大小选取前 k 个主成分，这里贡献率是指某个主成分的方差占全部方差的比重，也就是某个特征值占全部特征值合计的比重。即：

$$\text{贡献率}=\lambda_i/\sum_{i=1}^{m}\lambda_i \tag{3}$$

贡献率越大，说明该主成分所包含的原始变量的信息越强。主成分个数 k 的选取，一般要求累计贡献率达到 85%以上，这样才能保证综合变量能包括原始变量的绝大多数信息。

1.3 PCA 与 SVD 区别联系

PCA 的解法是 SVD 的一种变形和弱化，对于 $n\times m$ 的矩阵 $\boldsymbol{X}$，通过奇异值分解可以直接得到如下形式：

$$\boldsymbol{X}=\boldsymbol{U}_S\boldsymbol{\Lambda}_S\boldsymbol{V}_S^{\mathrm{T}} \tag{4}$$

其中：$\boldsymbol{U}_S$ 是一个 $n\times n$ 的矩阵；$\boldsymbol{V}_S$ 是一个 $m\times m$ 的矩阵，而 $\boldsymbol{\Lambda}_S$ 是 $m\times n$ 的对角阵。$\boldsymbol{\Lambda}_S$ 形式如下：

$$\boldsymbol{\Lambda}_S=\begin{bmatrix}\sigma_1 & & & 0\\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_m\end{bmatrix} \tag{5}$$

其中 $\sigma_1\geqslant\sigma_2\geqslant\cdots\geqslant\sigma_m$，是原矩阵的奇异值。由简单推导可知，对奇异值分解加以约束，$\boldsymbol{U}_S$ 的向量必须正交，则矩阵 $\boldsymbol{V}_S$ 即为 PCA 的特征值分解中的 $\boldsymbol{U}$，而 $\boldsymbol{\Lambda}=\mathrm{Diag}[\lambda_1,\lambda_2,\cdots,\lambda_M]=\dfrac{1}{n-1}\mathrm{Diag}[\sigma_1^2,\sigma_2^2,\cdots,\sigma_m^2]$。说明 PCA 并不一定需要求取 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$，也可以直接对原数据矩阵 $\boldsymbol{X}$ 进行 SVD 奇异值分解即可得到特征向量矩阵，也就是主元向量。

2 采用主成分分析的降噪设计[11]

综上所述，主成分分析中越往后面的主成分包含信息越小，考虑若噪声没有固定指向性，则可以通过舍去最后几个比较小的主成分降噪。对于单维数据，特别是周期信号，这时似乎无法采用主成分分析法，但若考虑信号的周期性，可以扩展为多维数据。

2.1 采用 Hankel 矩阵

由原较长数据(长度为 N)采用下一行比上一行延迟一点的方法构造 $n\times m$ 的 Hankel 矩阵，

$$\boldsymbol{X}=\begin{bmatrix}x(1) & x(2) & \cdots & x(m)\\ x(2) & x(3) & \cdots & x(m+1)\\ \vdots & \vdots & & \vdots\\ x(n) & x(n+1) & \cdots & x(n+m-1)\end{bmatrix}_{n\times m} \tag{6}$$

矩阵的 n、m 参数与 N 存在关系：$n+m-1=N$。

1) 对矩阵进行 SVD 分解，得到 m 个由大到小排列的特征值 $\sigma_1\geqslant\sigma_2\geqslant\cdots\geqslant\sigma_m$ 以及其对应正交特征向量 $\boldsymbol{V}_S$ 以及 $\boldsymbol{U}_S$。

2) 特征值选择和矩阵剪裁：取前 k 个较大的特征值予以保留，后面的删除对应矩阵进行剪裁，$\tilde{\boldsymbol{U}}_S$ 取前 $\boldsymbol{U}_S$ 的 $n\times k$ 部分，$\tilde{\boldsymbol{\Lambda}}_S$ 取 $\boldsymbol{\Lambda}_S$ 的前 $k\times k$，$\tilde{\boldsymbol{V}}_S$ 取 $\boldsymbol{V}_S$ 的前 $m\times k$。

3) 信号重构：

$$\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{U}}_S \tilde{\boldsymbol{\Lambda}}_S \tilde{\boldsymbol{V}}_S^{\mathrm{T}} \tag{7}$$

2.2 采用不重复排列矩阵

采用不重复排列矩阵：

$$\boldsymbol{X} = \begin{bmatrix} x(1) & x(2) & \cdots & x(m) \\ x(m+1) & x(m+2) & \cdots & x(2m) \\ \vdots & \vdots & & \vdots \\ x((n-1)m+1) & x((n-1)m+2) & & x(nm) \end{bmatrix}_{n\times m} \tag{8}$$

3 计算例子与分析

笔者通过对几种信号降噪效果的分析，以讨论主成分分析降噪的应用场合和适用条件。下面将分别讨论均值为零的信号(也称为无趋势项信号)、有趋势项信号和含瞬态冲击成分的信号含有白噪声时的降噪效果。

3.1 无趋势信号

信号用 x_i 表示，无趋势信号的均值为 0，$\sum_{i=1}^{n} x_i = 0$。假设，原始信号函数：

$$x_i = \sin(\frac{200\pi}{1\,225} i) + \sin(\frac{40\pi}{1\,225} i) \quad i = 1,2,\cdots,1\,225$$

而由于白噪声影响，检测到的信号为：

$$x_i' = x_i + 10 \times (\mathrm{rand}(i) - 0.5) \quad i = 1,2,\cdots,1\,225$$

式中rand(i)为白噪声产生函数。

下面考察采用 2.1 节和 2.2 节所讨论的 2 种主成分分析方法能否从含噪声信号中还原出原始信号，并定义误差评价函数：

$$\mathrm{error}\% = \frac{\sum_{i=1}^{n} (x_i'' - x_i)^2}{\sum_{i=1}^{n} x_i^2} \tag{9}$$

其中 x_i'' 表示减噪后得到的信号。

采用 2.1 节的方法可以得到一个 613 行 613 列的矩阵，其主分量分布如图 2(采用 2.2 节方法的主分量类似，不再赘述)。

若都取前 6 个主分量，这时 2 种方法能量比都在 0.15 左右，采用 2.1 节方法和 2.2 节方法的信噪比如表 1。

从表中可以看出，采用 2.1 节方法降噪性能优于 2.2 节方法，所以后面只讨论 2.1 节方法的降噪效果。图 3 是采用 2.1 节方法降噪结果时域图。

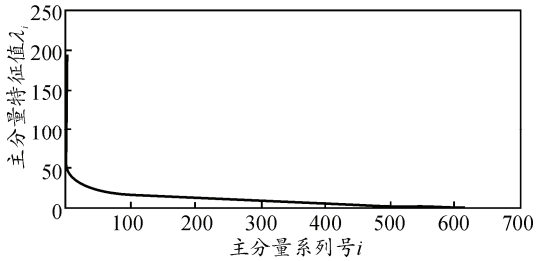


图 2　主分量特征值分布

表 1　不同方法降噪后信噪比

类型	信噪比	类型	信噪比
原始含噪信号	0.121 3	2.2 节方法处理信噪比	0.363 9
2.1 节方法处理信噪比	16.611 3		

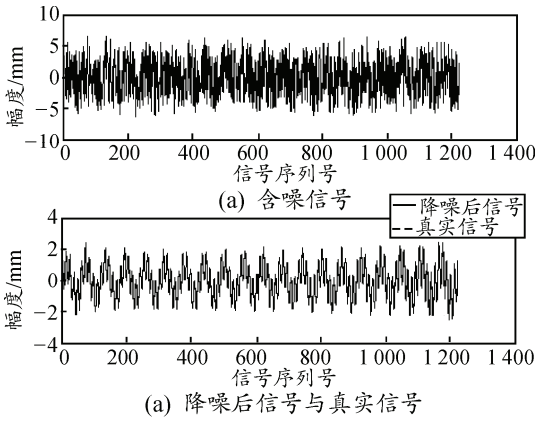


图 3　含噪声信号和降噪后信号

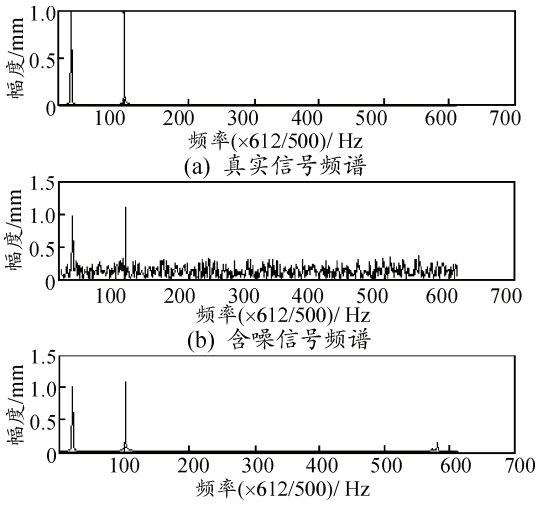


图 4　降噪后信号与原始信号频谱对比

可以看出，采用 2.1 节方法降噪效果很明显，图 3(a)是含噪声信号、图 3(b)是降噪后信号与原始信号对比，降噪后信号与原始信号几乎重合。

图 4 为降噪后频谱与原始信号频谱对比。其中图 4(a)是原始信号频谱，图 4(b)是含噪声信号频谱，图 4(c)是降噪后信号频谱，可以看出降噪后频谱与原始信号频谱几乎相同，降噪效果明显。

3.2　含趋势项信号

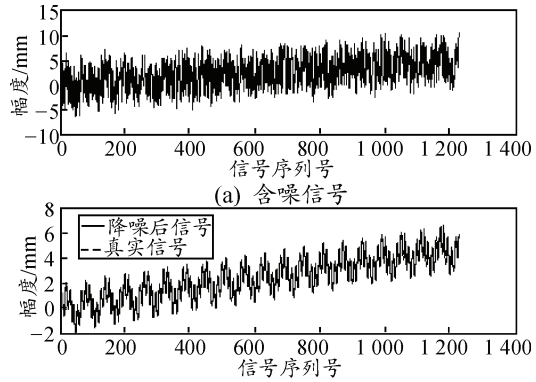
含趋势信号是指原始信号均值不为 0，含有趋势项，含趋势项的原始信号如下：

$$x_i = \sin(\frac{200\pi}{1\,225}i) + \sin(\frac{40\pi}{1\,225}i) + 5\times i/1\,225 \quad i = 1,2,\cdots,1\,225$$

在该信号基础上加上白噪声：

$$x_i' = x_i + 10\times(\mathrm{rand}(i) - 0.5) \quad i = 1,2,\cdots,1\,225$$

图 5 是采用 2.1 节方法降噪结果时域图。



(a) 含噪信号

(b) 降噪后信号与真实信号

图 5　含噪声信号和降噪后信号

采用 2.1 节方法降噪效果很明显，图 5(a)是含噪声信号，图 5(b)是降噪后信号与原始信号对比，降噪后信号与原始信号几乎重合，因此主成分分析对含趋势信号也适用。

3.3　含冲击信号

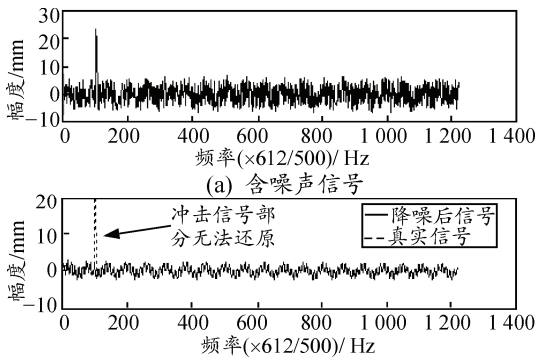
含冲击信号是在信号中加入瞬态冲击成分，信号如下：

$$x = \sin(\frac{200\pi}{1\,225}i) + \sin(\frac{40\pi}{1\,225}i) + 20\delta(i-100) \quad i = 1,2,\cdots,1\,225$$

其中 $\delta(n)$ 为狄拉克函数。

含噪声信号：

$$x_i' = x_i + 10\times(\mathrm{rand}(i) - 0.5) \quad i = 1,2,\cdots,1\,225$$



(a) 含噪声信号

(b) 降噪后信号与原始信号

图 6　含噪声信号和降噪后信号

图 6 是采用 2.1 节方法降噪结果时域图。

采用 2.1 节方法降噪效果很明显，图 6 中图(a)是含噪声信号，图(b)是降噪后信号与原始信号对比，降噪后信号没能还原出冲击信号，因此主成分分析对含冲击成分信号不适用。

4　结论

笔者讨论了主成分分析原理、计算步骤以及它与奇异值分解 SVD 的区别联系，提供了 2 种采用主成分分析降噪的设计，并对无趋势信号、有趋势信号和含白冲击信号的降噪效果进行分析，结果表明：

1) 采用 PCA 可以很好消除白噪声。这是因为一般周期信号会集中在前几个主分量上，而白噪声会均匀分布在所有分量上，所以通过去掉不重要分量就可以有效消除白噪声。

2) 采用 Hankel 矩阵可以最大限度重现信号的周期性，比分段方法降噪效果更明显。

3) 对于冲击信号，文中的 PCA 方法无法进行降噪，因此可以认为文中提出的 PCA 方法对于瞬态信号不适用。

参考文献：

[1] 胡云生, 郑继明. 基于主分量分析和遗传神经网络的电力负荷预测[J]. 自动化技术与应用, 2008, 27(8): 1-3.

[2] 夏鹏, 张浩然, 徐展敏. 一种增量 PCA 算法及其在人脸识别中的应用[J]. 计算机工程与应用, 2008, 44(6): 228-230.

[3] 顾绍红, 王永生, 王光霞. 主成分分析模型在数据处理中的应用[J]. 测绘科学技术学报, 2007(5): 387-390.

[4] 史峰. Matlab 函数速查手册[M]. 北京: 中国铁路出版社, 2011: 607-624.

[5] Lu Chen. 主元分析(PCA)理论分析及应用[OL]. http://www.cad.zju.edu.cn/home/chenlu/pca.htm.

[6] 余锦华, 杨维权. 多元统计分析与应用[M]. 广州: 中山大学出版社, 2005: 1-222.

[7] 王晓华. 过程监控与故障诊断的 ICA-MPCA 方法[D]. 大连: 大连理工大学, 2008: 1-50.

[8] Chen Q, Wynne R J, Goulding P, et al. The application of principal component analysis and kernel density estimation to enhance process monitoring[J]. Control Engineering Practice, 2000, 8(5): 531-543.

[9] 樊海锋, 徐凯, 江全元, 等. 考虑最大可观测通道数和关键故障线路约束的 PMU 优化配置研究[J]. 机电工程, 2013, 30(10): 1240-2045.

[10] 周东华, 叶银忠. 现代故障诊断与容错控制[M]. 北京: 清华大学出版社, 2000: 1-347.

[11] 余昌和, 李建黎. 低信噪比下相干信号的 DOA 估计的白噪声滤除方法[J]. 信号处理, 2012, 28(7): 957-962.