

相对主元分析及其在数据压缩和故障诊断中的应用研究

文成林^{1,2} 胡静² 王天真³ 陈志国²

摘要 传统主元分析 (Principal component analysis, PCA) 方法因忽视量纲对系统的影响, 从而使选取的主元难以具有代表性; 而在进行量纲标准化后, 又因得到的特征值常常是近似相等的而无法进行有效的主元提取. 针对这一主要问题, 本文通过引入相对化变换 (Relative transform, RT)、相对主元 (Relative principal components, RPCs) 和分布“均匀”等概念, 建立起一种相对主元分析 (Relative principal component analysis, RPCA) 的新方法. 该方法首先对系统各分量进行量纲标准化; 其次再根据系统的先验信息分析和确定各分量的重要程度; 然后在系统能量守恒的准则下, 赋以系统各分量相应的权值; 最后利用已建立起的相对主元模型, 对系统实施 RPCA. 同时运用数值例子, 开展了 RPCA 在数据压缩和系统故障诊断中的应用研究. 理论分析和仿真实验均表明, 采用 RPCA 方法选取出的主元更具代表性和显著几何意义, 加之选取主元的灵活性, 将使新方法具有更广泛的应用前景.

关键词 数据压缩, 故障诊断, 相对主元分析, 相对化变换, 量纲标准化, 分布“均匀”

中图分类号 TP391; TP277

Relative PCA with Applications of Data Compression and Fault Diagnosis

WEN Cheng-Lin^{1,2} HU Jing² WANG Tian-Zhen³ CHEN Zhi-Guo²

Abstract In traditional principal component analysis (PCA), because of the neglect of the influence of dimensions on the system, the selected principal components (PCs) often fail to be representative. For this problem, an improved algorithm, called relative principal component analysis (RPCA), is proposed by introducing some new concepts, such as relative transform (RT), relative principal components (RPCs), “rotundity” scatter and so on. Firstly, the algorithm standardizes every variable's dimension in this system. Secondly, according to priori information, it analyzes and determines the different important levels of different variables. And then it allocates weights for each variable under the criterion of conservation of system energy. Finally, the algorithm utilizes the relative-principal-component model established to analyze system. Meanwhile, its functions are illustrated by some numerical examples such as data compression and system fault diagnosis. Both theoretic analysis and computer simulation have shown that these selected RPCs are representative and their significance of geometry is notable. So we can say that the new method may have extensive applications, together with the flexibility of PCs selection.

Key words Data compression, fault diagnosis, relative principal component analysis (RPCA), relative transform (RT), dimensionless standardization, “rotundity” scatter

在工业生产过程中, 常存在着大量的高度相关的系统变量, 而这些变量在每一时刻的采样值中都蕴含着生产过程是否正常、产品质量是否合格等重要信息. 如何有效地利用这些信息并及时发现故障

或预测故障, 从而保证设备在工作期间始终能够安全、高效、可靠地运转, 是故障检测与诊断技术中的一个重要问题.

主元分析 (Principal component analysis, PCA) 是多元统计分析中常用的方法之一, 它以有限长度的多维变量时间序列构成的随机矩阵为基础, 通过建立数目较小的综合变量, 使其能更集中地反映原来变量中所包含的变化信息, 并有效分离系统信息与噪声. 目前, PCA 已在故障诊断、数据压缩、信号处理和模式识别等领域中均有广泛的应用^[1-2].

然而, 传统的 PCA 方法却常存在如下不足:

1) 传统的 PCA 方法是先计算出上述随机矩阵协方差阵的特征值, 随后再依据特征值的大小来确定各级主元变量的. 然而, 特征值与相应分变量协方差的大小密切相关, 而各个变量协方差的大小又与对应分量所使用的量纲 (单位) 密切相关. 正是由于各变量量纲的不同, 才使得在传统方法中出现了方

收稿日期 2007-06-29 收修改稿日期 2007-08-27
Received June 29, 2007; in revised form August 27, 2007
国家自然科学基金 (60434020, 60572051), 浙江省国际合作重点项目 (2006C24G2040012), 浙江省教育厅科技计划重点项目 (20050530) 资助

Supported by National Natural Science Foundation of China (60434020, 60572051), International Cooperation Item of Zhejiang Province (2006C24G2040012), and Key Project of Science and Technology Program of Education Department of Zhejiang Province (20050530)

1. 杭州电子科技大学信息与控制研究所 杭州 310018 2. 河南大学计算机与信息工程学院 开封 475001 3. 上海海事大学电气自动化系 上海 200135

1. Institute of Information and Control, Hangzhou Dianzi University, Hangzhou 310018 2. Department of Computer and Information Engineering, Henan University, Kaifeng 475001 3. Department of Electrical Automation, Shanghai Maritime University, Shanghai 200135
DOI: 10.3724/SP.J.1004.2008.01128

差大的分变量不一定就是实际系统中起主导作用的变量的反常现象^[3].

2) 从 PCA 建模的观点来看, 对原始随机矩阵进行去量纲标准化处理之后, 系统结构也发生了变化, 并常常导致协方差阵的特征值大小近似相等, 即得到的随机矩阵在几何上呈现分布“均匀”的现象. 因此, 很难提取出有代表性的主元, 也就不容易对信号实施有效地压缩和监测.

3) 从能量守恒的观点来看, 去量纲标准化前后的系统能量通常是不守恒的.

为了解决上述实际问题, 就需要在相应理论上有所突破. 因此, 本文提出相对主元 (Relative principal components, RPCs) 的概念, 在充分考虑系统中不同变量具有不同重要性的基础上, 建立一种相对主元分析 (Relative principal component analysis, RPCA) 的新方法, 不仅从理论上分析了量纲标准化过程可能引起的分布“均匀”现象以及相对化变换 (Relative transform, RT) 对系统特征结构的影响, 而且系统地探讨了 RPCA 在数据压缩和故障诊断中的应用. 基本思想是: 首先通过预处理, 消除因量纲差异所带来的虚假影响, 使得各个分变量处于“平等”的地位; 然后再利用系统各分量的先验知识, 并根据系统的实际要求, 在一定准则下赋予系统各分量不同的权值; 最终实现能更有效地获取、分析和利用系统所需要的信息.

1 基于主元分析的故障辨识

PCA 的主要任务是对多维变量时间序列构成的有限长度随机矩阵进行降维. 虽然, 有时需要用数目巨大的系统变量才能较完善地描述一个复杂系统, 但是, 在通常情况下, 系统的大部分变异性仅需用少数几个主元变量便可以说明. 而利用这少数的主元变量就可以有效地代表或解释原始系统变量序列构成的随机矩阵的特征结构, 从而实现对信号的压缩与系统的故障检测与诊断^[4-5].

1.1 主元模型

对一个给定的多变量系统, 设用来描述系统的 n 维变量为

$$\mathbf{x}(k) = [x_1(k), x_2(k), \dots, x_n(k)]^T \in \mathbf{R}^n$$

而由它的一段长度有限的系统变量时间序列构成的随机阵为

$$X = X(k, k + N - 1) = [\mathbf{x}(k), \mathbf{x}(k + 1), \dots, \mathbf{x}(k + N - 1)]$$

若用 \mathbf{x}_i 表示系统的第 i 个分量序列构成的有限长度时间序列

$$\mathbf{x}_i = \mathbf{x}(k, k + N - 1) = [x_i(k), x_i(k + 1), \dots, x_i(k + N - 1)]$$

则矩阵 X 又可以被表示为

$$X = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} \quad (1)$$

若再记上述随机矩阵 X 的协方差阵为

$$\Sigma_X = E\{[X - E\{X\}][X - E\{X\}]^T\}$$

那么, 通过求解

$$|\lambda I - \Sigma_X| = 0 \quad (2)$$

和

$$(\lambda_i I - \Sigma_X)\mathbf{p}_i = 0, \quad i = 1, 2, \dots, n \quad (3)$$

就可以分别求出矩阵 Σ_X 的特征值 λ_i 和对应的特征向量 $\mathbf{p}_i = [p_{i1}, p_{i2}, \dots, p_{in}]^T$, 且有

$$\langle \mathbf{p}_i, \mathbf{p}_j \rangle = \mathbf{p}_i^T \mathbf{p}_j = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \quad (4)$$

$$i, j = 1, 2, \dots, n$$

为了便于描述, 假定 n 个特征值互不相等且有 $\lambda_1 > \lambda_2 > \dots > \lambda_n > 0$.

注 1. 若特征值有重根时, 式 (3) 的形式会有相应的变化^[6].

由式 (3) 求出特征向量 \mathbf{p}_i 并结合式 (1), 可得

$$\begin{cases} \mathbf{v}_1 = \mathbf{v}_1(k, k + N - 1) = (\mathbf{p}_1)^T X = \sum_{j=1}^n p_{1j} \mathbf{x}_j \\ \mathbf{v}_2 = \mathbf{v}_2(k, k + N - 1) = (\mathbf{p}_2)^T X = \sum_{j=1}^n p_{2j} \mathbf{x}_j \\ \vdots \\ \mathbf{v}_n = \mathbf{v}_n(k, k + N - 1) = (\mathbf{p}_n)^T X = \sum_{j=1}^n p_{nj} \mathbf{x}_j \end{cases}$$

因此

$$V = P^T X \quad (5)$$

其中

$$V = V(k, k + N - 1) = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_n \end{bmatrix}$$

$$P = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_n]$$

若再定义

$$\mathbf{v}(k) = [v_1(k), v_2(k), \dots, v_n(k)]^T$$

则随机矩阵 V 亦可表示为

$$V = [\mathbf{v}(k), \mathbf{v}(k+1), \dots, \mathbf{v}(k+N-1)]$$

由于 P 是正交矩阵, 因此分析和研究随机矩阵 X 的统计特性就等价地变为分析随机矩阵 V 的统计特性.

若顺序选取前 m ($m < n$) 个向量 $\mathbf{v}_1, \dots, \mathbf{v}_m$ 作为主(要)元(素), 那么, X 便可以分解为如下形式^[7]

$$X = PV = \sum_{i=1}^n \mathbf{p}_i \mathbf{v}_i = \sum_{i=1}^m \mathbf{p}_i \mathbf{v}_i + \sum_{i=m+1}^n \mathbf{p}_i \mathbf{v}_i = P_m V_m + P_e V_e = X_m + E \quad (6)$$

其中, $X_m = \sum_{i=1}^m \mathbf{p}_i \mathbf{v}_i$ 为主元模型, $E = \sum_{i=m+1}^n \mathbf{p}_i \mathbf{v}_i$ 为建模误差. 显然有, $P_m = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_m]$, $V_m = [\mathbf{v}_1^T, \mathbf{v}_2^T, \dots, \mathbf{v}_m^T]^T$, $P_e = [\mathbf{p}_{m+1}, \mathbf{p}_{m+2}, \dots, \mathbf{p}_n]$, $V_e = [\mathbf{v}_{m+1}^T, \mathbf{v}_{m+2}^T, \dots, \mathbf{v}_n^T]^T$.

通常采用方差累计贡献率 (Cumulative percent variance, CPV) 法来确定主元个数 m ^[8], 即

$$CPV(m) = 100 \left(\frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^n \lambda_i} \right) \% \geq CL \quad (7)$$

式中 CL 是人为设定的控制限, 一般取 85%.

性质 1. 若记

$$\text{Var}(\mathbf{v}_i) = E \left\{ [\mathbf{v}_i - E\{\mathbf{v}_i\}] [\mathbf{v}_i - E\{\mathbf{v}_i\}]^T \right\}$$

$$\text{Cov}(\mathbf{v}_i, \mathbf{v}_j) = E \left\{ [\mathbf{v}_i - E\{\mathbf{v}_i\}] [\mathbf{v}_j - E\{\mathbf{v}_j\}]^T \right\}$$

则有

$$\begin{aligned} \text{Var}(\mathbf{v}_i) &= (\mathbf{p}_i)^T \Sigma_X \mathbf{p}_i = \lambda_i, \quad i = 1, 2, \dots, n \\ \text{Cov}(\mathbf{v}_i, \mathbf{v}_j) &= (\mathbf{p}_i)^T \Sigma_X \mathbf{p}_j = 0, \quad i \neq j \end{aligned} \quad (8)$$

证明. 性质 1 的证明过程是平凡的. 由式 (3) 和式 (4), 有

$$\text{Cov}(\mathbf{v}_i, \mathbf{v}_j) = (\mathbf{p}_i)^T \Sigma_X \mathbf{p}_j = (\mathbf{p}_i)^T \lambda_j \mathbf{p}_j = \begin{cases} \lambda_i, & i = j \\ 0, & i \neq j \end{cases}$$

□

定义 1. 随机矩阵的范数定义为

$$\|X\|_2 = \left(\sum_{i=1}^n \sum_{q=0}^{N-1} E\{x_i^2(k+q)\} \right)^{\frac{1}{2}}$$

注 2. 容易验证 $\|X\|_2$ 满足范数的三个基本条件^[9].

性质 2. 主元分析前后系统随机矩阵能量保持守恒, 即 $\|X\|_2^2 = \|V\|_2^2$.

证明.

$$\begin{aligned} \|V\|_2^2 &= \sum_{i=1}^n \sum_{q=0}^{N-1} E\{v_i^2(k+q)\} = E\{\text{tr}(VV^T)\} = \\ &E\{\text{tr}(V^T V)\} = E\{\text{tr}(X^T P P^T X)\} = \\ &E\{\text{tr}(X^T X)\} = E\{\text{tr}(X X^T)\} = \\ &\sum_{i=1}^n \sum_{q=0}^{N-1} E\{x_i^2(k+q)\} = \|X\|_2^2 \end{aligned} \quad (9)$$

上式中 $\text{tr}(\cdot)$ 表示矩阵的迹. □

1.2 统计量的确定

根据式 (6) 中所确立的主元模型, 在新的时刻 k , 过程向量 $\mathbf{x}(k)$ 可被分解为

$$\mathbf{x}(k) = \hat{\mathbf{x}}(k) + \tilde{\mathbf{x}}(k) \quad (10)$$

其中 $\hat{\mathbf{x}}(k) = P_m P_m^T \mathbf{x}(k)$ 是 $\mathbf{x}(k)$ 在主元子空间上的投影, $\tilde{\mathbf{x}}(k) = I - P_m P_m^T \mathbf{x}(k)$ 是 $\mathbf{x}(k)$ 在残差子空间上的投影.

统计指标 T^2 表示 $\mathbf{x}(k)$ 在主元子空间上投影的大小, 即

$$T^2 = \|\hat{\mathbf{x}}(k)\|_2^2$$

Q -统计量也叫做平方预报误差 (Squared prediction error, SPE), 它度量了 $\mathbf{x}(k)$ 偏离主元模型的程度, 即

$$SPE = \|\tilde{\mathbf{x}}(k)\|_2^2$$

如果过程运行正常, 应同时满足

$$T^2 < UCL, \quad SPE < Q_\alpha$$

其中, UCL 和 Q_α 分别定义了 T^2 和 SPE 的判断门限^[10].

进一步, 通过对过程变量的 SPE 统计贡献图进行分析, 便可以找出是哪些变量的变化引起了控制限的超限. 多元变量系统中, 第 i 个分变量 \mathbf{x}_i 在第 k 时刻对 SPE 统计量的贡献度定义为^[11]

$$C_i^{[SPE]}(k) = \|x_i(k) - \hat{x}_i(k)\|_2^2$$

□

其中, $\hat{x}_i(k)$ 为 $\hat{\mathbf{x}}(k)$ 的第 i 个随机变量.

2 量纲标准化分析

例 1. 考虑一近似匀加速直线运动的目标, 其状态方程为

$$\begin{aligned} \mathbf{x}(k+1) &= A(k)\mathbf{x}(k) + \mathbf{B}(k)\mathbf{w}(k) \\ \mathbf{w}(k) &\sim N[0, 0.1] \end{aligned}$$

其中

$$\mathbf{x}(k) = [x_1(k), x_2(k), x_3(k)]^T$$

$$A(k) = \begin{bmatrix} 1 & t & \frac{t^2}{2} \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{B}(k) = [0, 0, 1]^T$$

x_1 、 x_2 和 x_3 分别表示位移、速度和加速度, 按照国际单位制, 其单位依次取作米 (m)、米每秒 (m/s) 和米每二次方秒 (m/s²). 选初始值为 $[0, 18, 6]^T$; $t = 0.5$ s.

经传统的 PCA 运算, 求得的第一主元为 $0.9955x_1 + 0.0940x_2 - 0.0065x_3$. 注意到, 位移变量 x_1 几乎完全确定了主元变量. 这是因为位移与速度或加速度相比, 它的量值会较大, 测量时产生的误差也就会较大, 然而在实际系统中, 速度或加速度的变化将对目标的运动状态有更大的影响.

另一方面, 量纲的不同使我们很难对主元的物理意义作出解释. 例如, 很难说明, 位移、速度和加速度的线性组合代表什么物理意义.

例 2. 假定对某一物体的重量 x_1 和长度 x_2 进行测量, 单位分别取作千克 (kg) 和厘米 (cm). 经 4 次观察, 求得的测量矩阵协方差阵为

$$\Sigma_X^{(1)} = \begin{bmatrix} 367.58 & 3.25 \\ 3.25 & 2.92 \end{bmatrix}$$

有时, 也可能需要以毫米 (mm) 来作为长度的基本单位, 那么, 测量矩阵协方差阵就变为

$$\Sigma_X^{(2)} = \begin{bmatrix} 367.58 & 32.5 \\ 32.5 & 292 \end{bmatrix}$$

针对上述两种情况分别进行传统的 PCA 运算, 可以得出对应 $\Sigma_X^{(1)}$ 的第一主元为 $1.0000x_1 + 0.0089x_2$, 对应 $\Sigma_X^{(2)}$ 的第一主元为 $0.9380x_1 + 0.3467x_2$. 进一步, 求得的前者第一主元贡献率为 99.2%, 而后者相应的贡献率仅占 57.6%.

因此, 可以说, 一个随机变量相对小的变化对主元的选取有很大的影响, 而这种变化正是由于量纲的不同引起的.

针对上述一类问题, 许多研究人员做了大量的工作^[12], 提出的主要方法是

$$x_i^*(k) = \frac{x_i(k) - E\{x_i(k)\}}{[\text{Var}\{x(k)\}]^{\frac{1}{2}}}, \quad i = 1, 2, \dots, n \quad (11)$$

称由随机变量 $x_i^*(k)$ 构成的随机矩阵 X^* 为 X 的标准化随机矩阵. 这样, 一维随机变量 $x_i^*(k)$ 就转化为均值为 0 方差为单位的随机变量, 而计算 n 维随机变量的协方差阵 Σ_X 就等价地变成计算相关矩阵 R_X .

注 3. 随机变量 $x_i(k), x_i(k+1), \dots, x_i(k+N-1)$ 具有相同的统计特性. 记 $z_i(k)$ 为 $x_i(k)$ 的一次观测值, 一般来说, 在实际中要想确定 $x_i(k)$ 的统计量 $E\{x_i(k)\}$ 和 $\text{Var}\{x_i(k)\}$, 其难度比较大, 常用其观测值来近似替代, 即

$$E\{x_i(k)\} \approx \frac{1}{N} \sum_{q=0}^{N-1} z_i(k+q)$$

$$\text{Var}\{x_i(k)\} \approx \frac{1}{N} \sum_{q=0}^{N-1} [z_i(k+q) - E\{x_i(k+q)\}]^2$$

性质 3. 量纲标准化过程 (11) 使各个随机向量 \mathbf{x}_i^* 处于“平等”的地位, 也就是说每个随机变量序列的能量统计相同, 即 $\|\mathbf{x}_i^*\|_2^2 = \|\mathbf{x}_j^*\|_2^2, i \neq j$.

证明. 由 $\|\mathbf{x}_i^*\|_2^2 = E\left\{\sum_{q=0}^{N-1} [x_i^*(k+q)]^2\right\}$ 和式 (11) 得

$$\|\mathbf{x}_i^*\|_2^2 = E\left\{\sum_{q=0}^{N-1} \left[\frac{x_i(k+q) - E\{x_i(k+q)\}}{[\text{Var}\{x_i(k+q)\}]^{\frac{1}{2}}}\right]^2\right\} = N$$

□

注 4. 利用定义 1, 容易验证量纲标准化前后系统能量并不一定守恒, 即 $\|X\|_2^2 \equiv \|X^*\|_2^2$ 不恒成立.

定义 2. 若由式 (2) 所确定的特征值 $\lambda_1, \lambda_2, \dots, \lambda_n$ 近似相等, 则称由序列 $\mathbf{x}(k)$ 构成的随机矩阵 X 是分布“均匀”的, 反之称为分布“突出”.

注 5. 本节所定义的分布“均匀”与概率中“均匀分布”是两个不同的概念^[13]. 分布“均匀”的随机矩阵具有如下性质.

性质 4. 满足分布“均匀”的随机矩阵 X 中的向量集 $[E\{\mathbf{x}(k)\}, E\{\mathbf{x}(k+1)\}, \dots, E\{\mathbf{x}(k+N-1)\}]$ 在 \mathbf{R}^n 中点的分布近似构成一个超球体.

证明. 该性质的证明过程将会在第 4 节给出.

□

综上所述, 引入量纲标准化过程消除了由于不同量纲所造成的虚假变异影响. 然而通过分析可知, 经去量纲标准化之后的 PCA 带来了新的问题:

- 1) 量纲标准化前后系统能量并不一定守恒;
- 2) 量纲标准化之后随机矩阵 X^* 大多是分布“均匀”的。

3 相对主元分析方法

为了解决传统 PCA 方法中存在的问题, 本节在引入相对化变换的基础上建立了一种相对主元分析方法。

3.1 相对化变换

记 $X = (\mathbf{x}_i(k)), k = 1, 2, \dots, N$ 为由长度为 N 的系统 n 维变量时间序列构成的随机矩阵。不失一般性, 令 $E\{X\} = \mathbf{0}$ 。

定义 3 (相对化变换). 若记

$$X^R = W \cdot X = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix} \times \begin{bmatrix} x_1(1) & x_1(2) & \cdots & x_1(N) \\ x_2(1) & x_2(2) & \cdots & x_2(N) \\ \vdots & \vdots & \ddots & \vdots \\ x_n(1) & x_n(2) & \cdots & x_n(N) \end{bmatrix} = \begin{bmatrix} x_1^R(1) & x_1^R(2) & \cdots & x_1^R(N) \\ x_2^R(1) & x_2^R(2) & \cdots & x_2^R(N) \\ \vdots & \vdots & \ddots & \vdots \\ x_n^R(1) & x_n^R(2) & \cdots & x_n^R(N) \end{bmatrix} \quad (12)$$

$$\mathbf{x}_i^R = w_i \mathbf{x}_i, \quad i = 1, 2, \dots, n \quad (13)$$

其中 $w_i = \mu_i m_i > 0$ 。则称式 (12) 是对原始系统随机矩阵 X 所做的相对化变换, W 是相应的相对化变换算子, X^R 是相对化变换后的随机矩阵。 μ_i 称为比重因子, 是一种根据实际系统而定的先验信息, 它分别作用在每个变量上, 其大小体现了相应分变量在系统中的相对重要程度; m_i 是对应随机变量的标准化因子, 事实上, 依据不同的系统, 有多种标准化选择方式, 如

$$m_i = \frac{1}{\max_{1 \leq i \leq n} \|x_i(k)\|_2} \quad (14)$$

而上节提到的式 (11) 是其中最常用的一种。

若定义随机变量的相关系数如下

$$\rho\{x_i(k), x_j(m)\} = \frac{\text{Cov}\{x_i(k), x_j(m)\}}{\sqrt{\text{Var}\{x_i(k)\}\text{Var}\{x_j(m)\}}}$$

则有

性质 5. 相对化变换不改变随机变量之间的相关性, 即 $\rho\{x_i^R(k), x_j^R(m)\} = \rho\{x_i(k), x_j(m)\}$ 。

证明.

$$\rho\{x_i^R(k), x_j^R(m)\} = \frac{\text{Cov}\{x_i^R(k), x_j^R(m)\}}{\sqrt{\text{Var}\{x_i^R(k)\}\text{Var}\{x_j^R(m)\}}}$$

将式 (13) 代入上式中, 则

$$\rho\{x_i^R(k), x_j^R(m)\} = \frac{w_i w_j \text{Cov}\{x_i(k), x_j(m)\}}{\sqrt{w_i^2 \text{Var}\{x_i^R(k)\} w_j^2 \text{Var}\{x_j^R(m)\}}} = \frac{\text{Cov}\{x_i(k), x_j(m)\}}{\sqrt{\text{Var}\{x_i(k)\}\text{Var}\{x_j(m)\}}}$$

□

条件 1. 根据信号分析前后能量需保持守恒的准则, 我们要求随机矩阵相对化变换前后也要保持能量守恒, 即 $\|X\|_2^2 = \|X^R\|_2^2$ 。

由于对系统有效分析的前提是系统经变换前后的能量保持守恒或是一定比例关系, 而未知的比重因子 μ_i 可依据实际系统用不同的方法进行获取, 如神经网络和线性规划等。参照文献 [14] 在强跟踪滤波中确定次优渐消因子的方法, 我们也可给出一个类似的方法, 即假定由系统的先验知识, 可以大致确定

$$\mu_1 : \mu_2 : \cdots : \mu_n = k_1 : k_2 : \cdots : k_n \quad (15)$$

此时, 可令

$$\mu_i = \alpha k_i, \quad i = 1, 2, \dots, n$$

其中, $k_i > 0$ 为预先确定的常数; α 为待定因子。则由条件 1 可得

$$\|X\|_2^2 = \|X^R\|_2^2 = E \left\{ \sum_{i=1}^n \sum_{k=1}^N x_i^2(k) \right\} = E \left\{ \sum_{i=1}^n \sum_{k=1}^N [x_i^R(k)]^2 \right\} = \alpha^2 \left(\sum_{i=1}^n \frac{k_i^2}{\text{Var}\{x_i(k)\}} E \left\{ \sum_{k=1}^N [x_i(k) - E\{x_i(k)\}]^2 \right\} \right)$$

从而

$$\alpha = \left(\frac{E \left\{ \sum_{i=1}^n \sum_{k=1}^N x_i^2(k) \right\}}{\sum_{i=1}^n \frac{k_i^2}{\text{Var}\{x_i(k)\}} E \left\{ \sum_{k=1}^N [x_i(k) - E\{x_i(k)\}]^2 \right\}} \right)^{\frac{1}{2}}$$

标准化因子 m_i 的作用是去除系统第 i 个变量的量纲, 从而使系统各分量达到“平等”的地位; 而比重因子 μ_i 则体现了达到“平等”后的系统第 i 个“相对”分量对整个系统的影响程度; 一般是根据下列原则来确定 μ_i 的:

1) 相对化变换将使得由系统变量一段长度有限的时间序列构成的随机矩阵或经去量纲标准化之后的标准化随机矩阵能尽量摆脱分布“均匀”的现象;

2) 经相对化变换后求取的相对主元更具有代表性和表征系统变化的能力;

3) 相对化变换前后系统的能量保持守恒, 即 $\|X\|_2^2 = M \|X^R\|_2^2$.

因此, 经相对化变换后的随机矩阵 X^R 不仅保持了原随机变量间的相关性、摆脱原有分布“均匀”的现象, 而且还将会更加方便主元的选取, 从而能进行更有效的特征提取、数据压缩和故障诊断等工作。

3.2 相对主元模型

通过对 X 进行相对化变换, 先计算 X^R 的协方差矩阵 Σ_X^R , 然后再求解 Σ_X^R 的特征值 $\lambda_1^R \geq \dots \geq \lambda_n^R$, 及相应特征向量 $\mathbf{p}_1^R, \dots, \mathbf{p}_n^R$ 等步骤, 就可以得到 m 个互不相关的相对主元 (Relative principal components, RPCs) $\mathbf{v}_1^R, \dots, \mathbf{v}_m^R, m < n$.

类似于 PCA, 在 RPCA 方法里也可度量每个 RPC 对系统贡献的百分比

$$CPV_i^R = 100 \left(\frac{\lambda_i^R}{\sum_{i=1}^n \lambda_i} \right) \% \quad (16)$$

并根据累计贡献率的大小及系统的近似表示程度来确定 RPCs 的数目 m , 分析和解释 Σ_X^R 的特征结构, 表征系统的变化特征. 这样, 一个由长度为 N 的 n 维系统随机变量时间序列构成的随机阵 $X \in \mathbf{R}^{n \times N}$ 就可以被分解为对应的相对主元子空间 $X_m^R = (W^{-1}P_m^R P_m^{R\top} W) X$ 和相对残差子空间 $E^R = (I - W^{-1}P_m^R P_m^{R\top} W) X$.

4 RPCA 的几何解释

设原始变量有限长度时间序列构成的随机矩阵 X 的分布为 $N[E\{X\}, \Sigma_X]$. 在以 $E\{X\}$ 为中心的超椭球上, X 的密度是常数^[8]

$$[X - E\{X\}]^T (\Sigma_X)^{-1} [X - E\{X\}] = c^2$$

而此超椭球的各轴分别为 $\pm c\sqrt{\lambda_i} \mathbf{p}_i, i = 1, \dots, n$. 其中 λ_i 和 \mathbf{p}_i 分别是协方差阵 Σ_X 的特征值及其对应的特征向量.

不妨令 $E\{X\} = \mathbf{0}$, 则 $X^T (\Sigma_X)^{-1} X = c^2$. 由 $\Sigma_X = \sum_{i=1}^n \lambda_i \mathbf{p}_i \mathbf{p}_i^T$ 知

$$(\Sigma_X)^{-1} = \sum_{i=1}^n \frac{1}{\lambda_i} \mathbf{p}_i \mathbf{p}_i^T \quad (17)$$

所以

$$X^T \left(\sum_{i=1}^n \frac{1}{\lambda_i} \mathbf{p}_i \mathbf{p}_i^T \right) X = \frac{1}{\lambda_1} (\mathbf{p}_1^T X)^T (\mathbf{p}_1^T X) + \dots + \frac{1}{\lambda_n} (\mathbf{p}_n^T X)^T (\mathbf{p}_n^T X) = c^2 \quad (18)$$

其中, $\mathbf{p}_1^T X, \dots, \mathbf{p}_n^T X$ 为 X 的主元. 那么, 式 (18) 也可以写为

$$c^2 = \frac{1}{\lambda_1} \mathbf{v}_1^T \mathbf{v}_1 + \dots + \frac{1}{\lambda_n} \mathbf{v}_n^T \mathbf{v}_n \quad (19)$$

如果 λ_1 是最大的特征值, 那么超椭球的主轴沿着 \mathbf{p}_1 的方向, 其余较次要的坐标轴依次沿着由 $\mathbf{p}_2, \dots, \mathbf{p}_n$ 所确定的方向. 若 $\lambda_1 \approx \dots \approx \lambda_n$, 则 X 在 \mathbf{R}^n 空间中点的分布近似构成一个超球体, 从而验证了性质 4.

而本文所阐述的 RPCA 方法可以改变随机矩阵在空间中点的“分布”. 比重因子 μ_i 的引入可使特征值的大小有明显差别, 即超椭球的长短轴变得明显, 从而将分布“均匀”的随机矩阵变为分布“突出”, 最终导致所提取的主元贡献率较大.

若令相对化变换前随机矩阵 X 为分布“均匀”的, 即 $\lambda_1 \approx \dots \approx \lambda_n \approx \lambda$, 则存在非奇异矩阵 T , 使得

$$T^{-1} \Sigma_X T = \Lambda \approx \lambda I \quad (20)$$

其中, $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$. 那么

$$\Sigma_X \approx \lambda I \quad (21)$$

而相对化变换之后 $\Sigma_{X^R} = E\{(WX)(WX)^T\} = W \Sigma_X W$, 将式 (21) 代入, 有

$$\Sigma_{X^R} \approx \lambda W^2 \quad (22)$$

λW^2 是以 Σ_{X^R} 的特征值 $\lambda_i^R, i = 1, \dots, n$ 为对角线元素构成的矩阵, 所以 $\lambda_i^R \approx w_i \lambda_i$. 也就是说, 如果原始随机矩阵协方差的特征值是近似相等的, 根据 w_i 大小的不同, 就可以得到分布“突出”的相对化随机矩阵.

针对分布“突出”的系统变量的 40 次实现, 如图 1(a) 所示, 本节将分别给出利用 PCA 方法、量纲标准化 PCA 方法和 RPCA 方法得出的几何解释. 表 1 为仿真环境和参数设置, 表 2 为计算结果. 对原始数据直接进行 PCA 处理, 计算出的最大特征

值为 24.77, 对应的第一主元贡献率为 95.24%. 然而, 传统 PCA 方法由于没有考虑量纲对特征值大小的影响, 因而不能准确反映变量对系统的主要贡献. 先利用式 (11) 去量纲标准化, 然后再运用 PCA 方法得出 $\lambda_1 \approx \lambda_2$, 表明经标准化之后得到的数据阵的特征值是近似相等的, 这由图 1(c) 很容易看出, 椭圆近似为圆形, 此时各主元没有明显的代表性. 图 1(d) 给出了当比重因子为 $\mu_1 : \mu_2 = 2 : 5$ 时, 经 RT 变换得到的 40 个相对化数据的分布. 由 RPCA 运算, 我们可以求得 $\lambda_1^R \approx \lambda_2^R$, 第一相对主元 RPC1 的贡献率 CPV_1^R 为 86.33%, 有明显的代表性. 从图 1(e) 可以看出, 椭圆图的长短轴分明, 相对分量的变化主要集中在 RPC1 方向上, 因此, 用 RPC1 就可以解释相对分量中的大部分变化. 同样地, 图 1(f) 给出当比重因子为 $\mu_1 : \mu_2 = 5 : 2$ 时关于 RPCs 的数据分布图.

表 1 仿真环境和参数设置

Table 1 Simulation environment and parameter setting

样本选择类型	样本序列长度 (N)	变量个数 (n)	标准化方法	c
分布“突出”	40	2	式 (11)	2

值得注意的有如下三点:

1) 标准化变换过程 (11) 使得每个变量都处于“平等”的地位. 从表 2 可以看出, 数据经标准化之后协方差矩阵的特征值 λ_i^* 基本上是在单位值附近变化, 其大小近似相等. 选取适当的比重因子

$$\mu_1 : \mu_2 = 5 : 2$$

通过计算机仿真实验可以计算出相应的特征值之间的比例关系近似为

$$\lambda_1 : \lambda_2 \approx 5^2 : 2^2$$

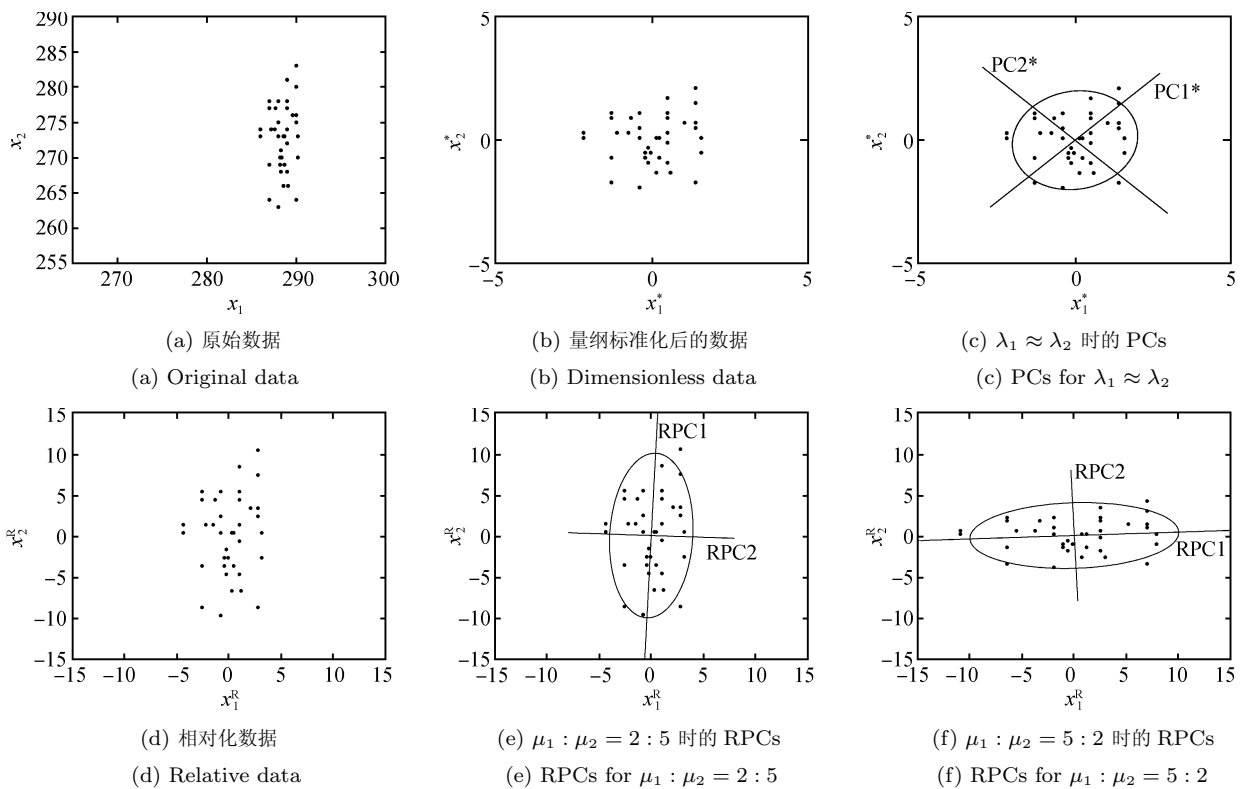


图 1 基于 PCA、量纲标准化 PCA 及 RPCA 方法的数据分布
Fig. 1 The distributions of data based on PCA, dimensionless PCA, and RPCA

表 2 仿真结果

Table 2 Simulation results

PCA		标准化之后的 PCA			$\mu_1 : \mu_2 = 2 : 5$ 或 $\mu_1 : \mu_2 = 5 : 2$ 时的 RPCA						
λ_1	24.77	CPV_1 (%)	95.24	λ_1^*	1.09	CPV_1^* (%)	54.35	λ_1^R	25.04	CPV_1^R (%)	86.33
λ_2	1.24	CPV_2 (%)	4.76	λ_2^*	0.91	CPV_2^* (%)	45.65	λ_2^R	3.97	CPV_2^R (%)	13.67

2) 当系统变量序列呈现分布“均匀”时, 就无法或者很难提取有代表意义的主元. 而本文建立的 RPCA 方法可通过相对化变换来改变原始随机矩阵的特征结构, 也就是可将分布“均匀”转化为分布“突出”. 例如将图 1 (b) 中点的分布在几何上进行相对化转换后, 得到图 1 (d). 因此, 可在多维随机系统中实现用少量几个主元来代替原来的系统变量, 从而达到特征提取的目的.

3) 考虑系统变量为二维时, 比重因子确定为 $\mu_1 : \mu_2 = k_1 : k_2, k_1, k_2 \in \mathbf{R}^+$, 与 $\mu_1 : \mu_2 = k_2 : k_1$ 得到的主元贡献率是相同的, 如表 2 所示. 这是因为

$$\Sigma_{X^R} = E \left\{ \begin{bmatrix} (w_1)^2 \mathbf{x}_1(\mathbf{x}_1)^T & w_1 w_2 \mathbf{x}_1(\mathbf{x}_2)^T \\ w_2 w_1 \mathbf{x}_2(\mathbf{x}_1)^T & (w_2)^2 \mathbf{x}_2(\mathbf{x}_2)^T \end{bmatrix} \right\} = E \left\{ \begin{bmatrix} (\mu_1)^2 N & \mu_1 \mu_2 E\{\mathbf{x}_1^*(\mathbf{x}_2^*)^T\} \\ \mu_2 \mu_1 E\{\mathbf{x}_2^*(\mathbf{x}_1^*)^T\} & (\mu_2)^2 N \end{bmatrix} \right\}$$

所以有

$$|\lambda^R I - \Sigma_{X^R}| = (\lambda^R)^2 - [(\mu_1)^2 N + (\mu_2)^2 N] \lambda^R + (\mu_1)^2 (\mu_2)^2 N^2 - (\mu_1)^2 (\mu_2)^2 E^2\{\mathbf{x}_1^*(\mathbf{x}_2^*)^T\}$$

从上式可知, 该特征方程与 μ_1, μ_2 的顺序无关, 即当 $\mu_1 : \mu_2 = k_1 : k_2$ 或 $\mu_1 : \mu_2 = k_2 : k_1$ 时所得到的特征值完全相同.

5 RPCA 方法在数据压缩中的应用

RPCA 方法可以用于高阶系统的降维、辅助变量的选择以及对系统变量样本序列构成的数据进行有效地压缩和特征提取. 本节将通过计算机仿真实验来分析在满足选取比重因子三原则的前提下, 比重因子的选取对系统主元变量的影响程度, 并验证与传统 PCA 相比, RPCA 方法在数据压缩中的优越性. 仿真环境设置如表 3.

表 3 仿真环境
Table 3 Simulation environment

样本类型	变量个数 (n)	样本序列长度 (N)	k_1	k_2	k_3	k_4
分布“均匀”	4	100	0.5	2	3	10

从表 4 可以看出, 由 4 个变量的 100 次实现所构成的原始矩阵协方差阵的特征值是近似相等的, 用传统 PCA 方法计算出第一主元的贡献百分比仅为 33.54%. 因此, 这里需要选取 3 个主元来代替原来的 4 个变量. 为了使获得的主元更具有代表性, 选择比重因子为 $\mu_1 : \mu_2 : \mu_3 : \mu_4 = 0.5 : 2 : 3 : 10$, 并采用 RPCA 新方法对同样的数据阵进行分析, 相对化变换后矩阵协方差阵的特征值就有明显的差异了, 第一相对主元 CPV_1^R 的贡献百分比为 88.37%.

因此, 在信息丢失最小的原则下, 仅用一个相对主元就可以合理解释原始数据阵中的大部分变化. 图 2 为传统 PCA 方法得出的主元的分布图. 图 3 为

表 4 传统 PCA 与 RPCA 方法的参数比较
Table 4 Parametric comparison for conventional PCA and RPCA

	传统 PCA			RPCA			
λ_1	9.48	$CPV_1(\%)$	33.54	λ_1^R	100.08	$CPV_1^R(\%)$	88.37
λ_2	7.83	$CPV_2(\%)$	27.70	λ_2^R	9.29	$CPV_2^R(\%)$	8.20
λ_3	6.91	$CPV_3(\%)$	24.45	λ_3^R	3.65	$CPV_3^R(\%)$	3.22
λ_4	4.04	$CPV_4(\%)$	4.04	λ_4^R	0.23	$CPV_4^R(\%)$	0.21

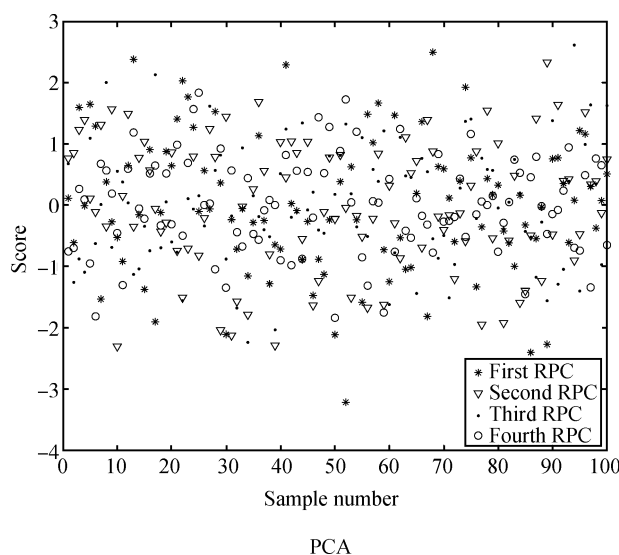


图 2 主元分布图

Fig. 2 The distribution plot of PCs

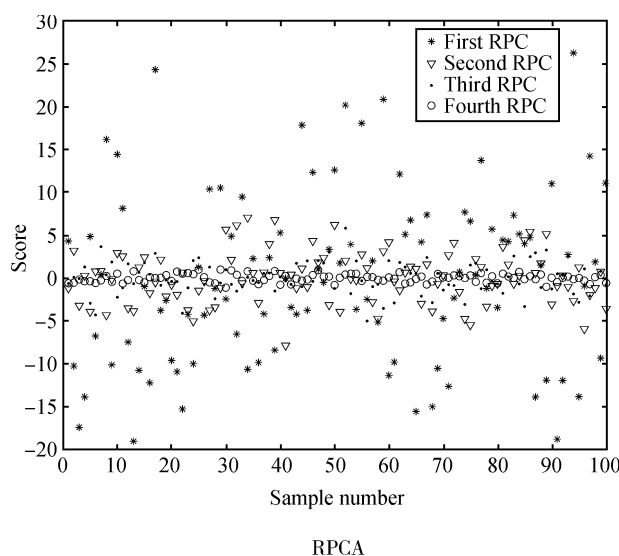


图 3 相对主元分布图

Fig. 3 The distribution plot of RPCs

RPCA 方法得出的相对主元的分布图. 比较两个图可以看出, 采用 RPCA 方法求取的 RPCs 彼此分离程度要比采用 PCA 方法求取的 PCs 间的分离程度大. 而且, 由于比重系数的引入, 使得相对主元的压缩能力要比主元的压缩能力强.

与传统 PCA 相比, 为了衡量 RPCA 方法的重构力, 我们定义如下关于“重构精度”的性能指标, 分别为: 重构误差平方和 $S^{[15]}$ 与重构率 c_m , 即

$$S = \sum_{k=1}^N \|\tilde{\mathbf{x}}(k)\|_2^2, c_m = \frac{\|X_m\|_2^2}{\|X\|_2^2}$$

这里, 从能量的角度定义了前 m 个主元重构率. 若重构误差平方和越小, 重构率越高, 表明重构精度越高. 表 5 依次给出了在选定相同主元数目时, 这两种方法在应用于数据重构时的比较结果.

表 5 传统 PCA 与 RPCA 的重构能力比较

Table 5 Comparison of restructuring abilities for conventional PCA and RPCA

选取的主元个数	数据重构方法	重构误差平方和	重构率
1	传统 PCA	12.8313	0.3335
	RPCA	0.0106	0.8837
2	传统 PCA	0.4804	0.6089
	RPCA	1.9403E-004	0.9657
3	传统 PCA	0.2461	0.8520
	RPCA	3.6550E-006	0.9980
4	传统 PCA	0.0793	0.9943
	RPCA	3.1554E-030	1.0000

以上结果表明:

1) 不管使用哪种方法, 相应的重构精度均随着选取的主元数目的增多而显著提高.

2) 比重因子的引入极大地提高了相对主元的代表能力. 因此, 当考虑相同主元数目的重构力时, RPCA 方法的重构精度将远远高于传统的 PCA 方法.

表 6 给出了在对分布“均匀”的随机矩阵进行特征提取时, 不同比重因子对 RT 后特征值的影响, 从中可以看出:

1) 由于所使用的 RPCA 新方法能充分地考虑到不同分量在实际应用中的相对重要程度, 那么, 原

始矩阵在经过相对化变换之后, 其特征结构也就变得较为突出, 因而更加方便了主元的提取, 同时也大大提高了压缩率.

2) 特别需要指出的是, 主元变量的选取与比重因子的取值排序无关. 即在式 (15) 中, 若设集合 $A = \{k_1, k_2, \dots, k_n | k_i \in \mathbf{R}^+\}$, 不妨令 $k_1 \geq k_2 \geq \dots \geq k_n$, 当 $k_i \in A, i = 1, 2, \dots, n$ 且 $\bigcup_{i=1}^n \{k_i\} = A$ 时, 都有 $\lambda_1^R : \dots : \lambda_n^R = k_1^2 : \dots : k_n^2$. 这里, $\alpha = 0.5009$, 且当比重因子大小为 $[0.5, 2, 3, 10]$ 时, 第一主元贡献率为 88.37%, 当比重因子大小改变为 $[2, 10, 0.5, 3]$ 时, 第一主元贡献率为 88.53%, 当比重因子大小改变为 $[3, 0.5, 2, 10]$ 时, 第一主元贡献率为 88.34%, 而当比重因子大小改变为 $[10, 2, 3, 0.5]$ 时, 第一主元贡献率为 88.56%.

6 基于 RPCA 的故障检测与辨识

一般情况下, 基于 RPCA 的故障检测与辨识应通过以下两个步骤来实现:

1) 采集正常生产过程中的测量数据, 用于 RPCA 建模并分别计算出 T^2 和 SPE 控制限.

2) 将新获取的过程数据经由 RPCA 模型分别投影到相对主元子空间和相对残差子空间, 运用统计分析的方法建立相应的统计量并进行假设检验. 在这个过程中, 由于引入了比重因子, 使得 RPCA 模型有可能更多地利用系统的先验信息, 从而更有效地辨识出系统的故障. 本节将用仿真实验来验证 RPCA 方法在故障检测与辨识中的效果.

考虑一个线性动态系统, 其状态方程和测量方程分别如下^[16]

$$\begin{aligned} \mathbf{x}(k) &= A\mathbf{x}(k-1) + B\mathbf{u}(k-1) \\ \mathbf{y}(k) &= \mathbf{x}(k) + \mathbf{v}(k) \end{aligned}$$

式中

$$A = \begin{bmatrix} 0.118 & -0.191 \\ 0.847 & 0.264 \end{bmatrix}, B = \begin{bmatrix} 1.0 & 2.01 \\ 3.0 & -4.0 \end{bmatrix}$$

$\mathbf{x}(k) \in \mathbf{R}^2$ 为状态向量, $\mathbf{y}(k) \in \mathbf{R}^2$ 为测量向量, $\mathbf{u}(k) \in \mathbf{R}^2$ 为过程输入, 定义为

$$\mathbf{u}(k) = G\mathbf{u}(k-1) + F\mathbf{w}(k-1)$$

表 6 比重因子不确定对相对主元贡献率影响的参数比较

Table 6 Parameter comparison of contribution rates for RPCs with uncertain proportion factors

k_1	k_2	k_3	k_4	λ_1^R	λ_2^R	λ_3^R	λ_4^R	$CPV_1^R(\%)$	$CPV_2^R(\%)$	$CPV_3^R(\%)$	$CPV_4^R(\%)$
0.5	2	3	10	100.0804	9.2890	3.6490	0.2316	88.37	8.20	3.22	0.21
2	10	0.5	3	100.2651	8.9413	3.8139	0.2298	88.53	7.90	3.37	0.20
3	0.5	2	10	100.0465	9.0733	3.9088	0.2214	88.34	8.01	3.45	0.19
10	2	3	0.5	100.2897	9.2797	3.4349	0.2458	88.56	8.19	3.03	0.22

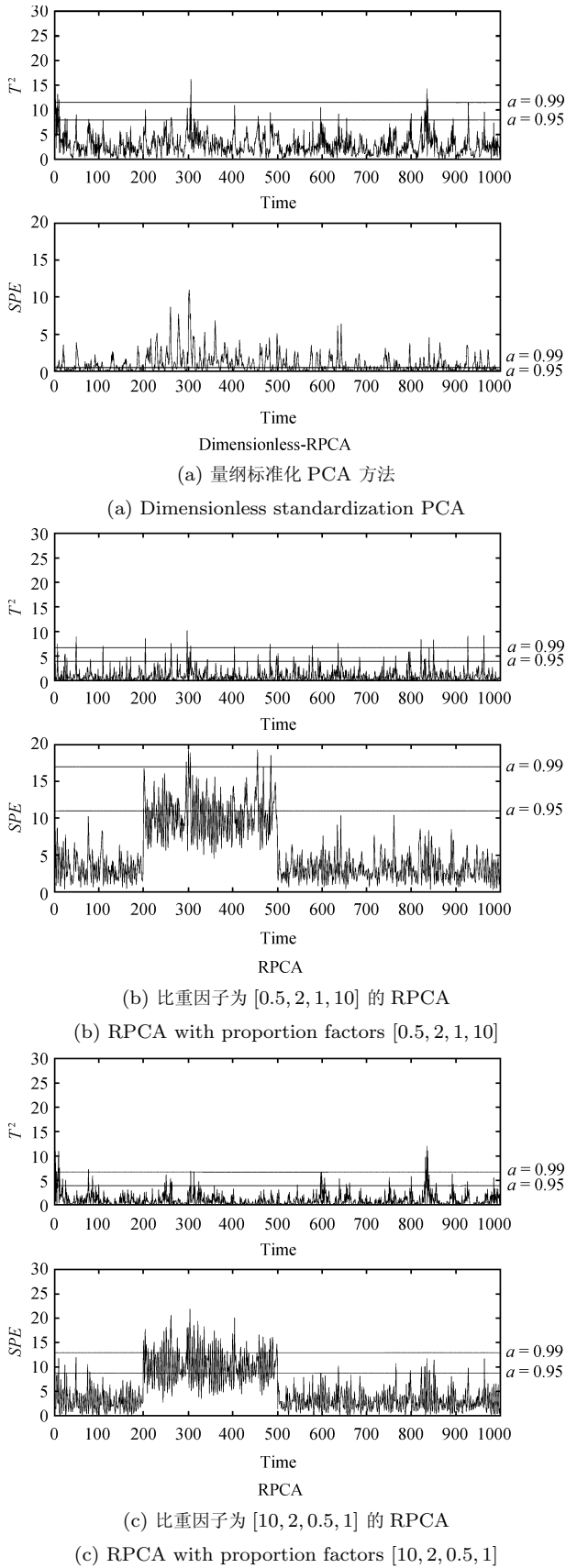


图 4 故障检测统计量的比较

Fig. 4 Comparison of statistics for fault detection

其中

$$G = \begin{bmatrix} 0.811 & -0.266 \\ 0.477 & 0.415 \end{bmatrix}, F = \begin{bmatrix} 0.193 & 0.689 \\ -0.320 & -0.749 \end{bmatrix}$$

$w(k) \in \mathbf{R}^2$ 为输入噪声, $v(k) \in \mathbf{R}^2$ 为测量噪声, 并具有如下的统计特性 $E\{w(k)\} = E\{v(k)\} = \mathbf{0}$ 、 $E\{w(k)w^T(j)\} = Q(k)\delta_{kj}$ 、 $E\{v(k)v^T(j)\} = R(k)\delta_{kj}$ 及 $Q(k) = R(k) = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}$.

首先取平稳运行的 1000 个正常工矿的样本用于建模, 检测变量为 $[y^T(k), u^T(k)]^T$, 求出相应的 T^2 和 SPE 控制限, 置信度分别取为 95% 和 99%. 接下来在第二个变量的 200~500 个测量区间加上均值漂移, 作为故障发生的假设. 最后分别采用量纲标准化 PCA 方法以及 RPCA 新方法进行实验. 仿真结果在表 7 和图 4 中给出.

表 7 $\mu_1 : \mu_2 : \mu_3 : \mu_4 = 0.5 : 2 : 1 : 10$ 时标准化 PCA 以及 RPCA 的参数比较

Table 7 Parameter comparison for standardization PCA and RPCA with $\mu_1 : \mu_2 : \mu_3 : \mu_4 = 0.5 : 2 : 1 : 10$

标准化之后的参数			RPCA 的参数		
λ_1	1.6417 $CPV_1(\%)$	41.04	λ_1^R	100.0144 $CPV_1^R(\%)$	95.03
λ_2	1.0135 $CPV_2(\%)$	25.34	λ_2^R	4.2113 $CPV_2^R(\%)$	4.00
λ_3	0.7494 $CPV_3(\%)$	18.73	λ_3^R	0.8055 $CPV_3^R(\%)$	0.77
λ_4	0.5955 $CPV_4(\%)$	14.89	λ_4^R	0.2188 $CPV_4^R(\%)$	0.21

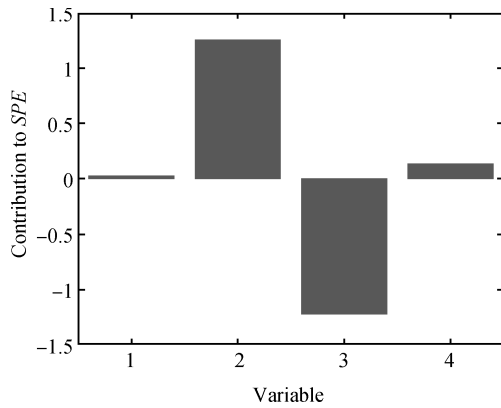
从上述的仿真结果中, 可以得知:

1) 量纲标准化 PCA 方法在改变原始数据特征结构的基础上, 也有可能缩小了系统的故障, 从而很难提取出有代表性的主元, 也就不容易检测出故障.

2) 从 T^2 统计图可以看出, 在置信度为 95% 的检测精度下, 两种方法均能检测出系统发生异常, 但是却不能指明出现异常的具体时间. 这是由于使用 T^2 方法容易丢失数据方向性信息的原因^[17].

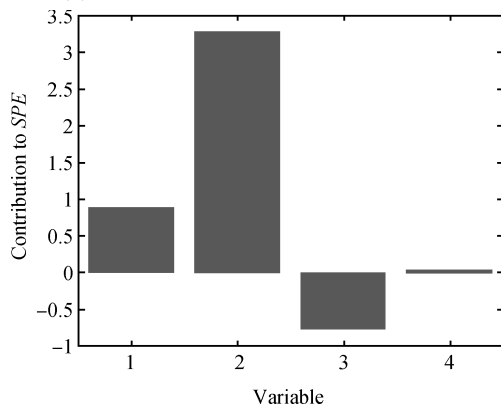
3) 通过选取比重因子 $\mu_1 : \mu_2 : \mu_3 : \mu_4 = 0.5 : 2 : 1 : 10$, 再运用 RPCA 新方法进行分析, 从 RPCA-SPE 图 (图 4(b)) 可以看出, 200~500 时刻的统计量明显超出控制限. 为了检测此方法关于不同比重因子的性能, 图 4(c) 给出了 $\mu_1 : \mu_2 : \mu_3 : \mu_4 = 10 : 2 : 0.5 : 1$ 时的仿真结果.

更进一步, 可以通过计算每个变量对 SPE 统计量的贡献率来确定故障变量, 以第 300 个采样点为例进行分析, 图 5 表明了 RPCA 方法能够准确地检测出是第二个变量发生了故障.



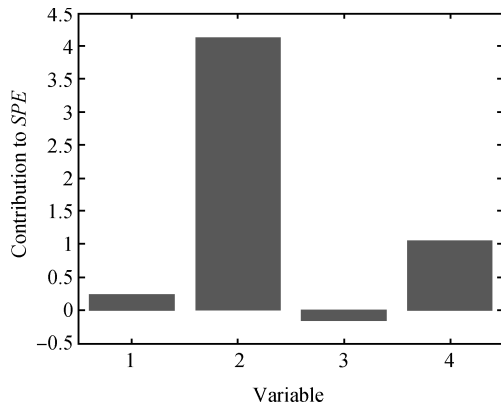
(a) 量纲标准化 PCA 方法

(a) Dimensionless standardization PCA



(b) 比重因子为 [0.5, 2, 1, 10] 的 RPCA

(b) RPCA with proportion factors [0.5, 2, 1, 10]



(c) 比重因子为 [10, 2, 0.5, 1] 的 RPCA

(c) RPCA with proportion factors [10, 2, 0.5, 1]

图 5 变量对 SPE 统计量的贡献图

Fig. 5 Contribution plots for SPE statistic

7 总结

通常情况下, 利用传统 PCA 方法可以对高阶系统进行数据压缩、特征提取及故障诊断等, 但当随机矩阵呈现分布“均匀”时, 由于很难选取主元, 或者选取主元时没有考虑到系统的实际需求及量纲对主元选取的影响, 使得主元没有代表性. 而本文建立的

RPCA 方法, 克服了采用 PCA 方法时存在的一些问题, 并具有如下优点:

1) 避免了在不同量纲意义下, 协方差大的变量在选取主元时起主要作用;

2) 考虑到不同分量在系统中的不同重要性. 从而能够更准确地辨识出系统故障, 仿真结果也显示了算法的优越性.

本文仅是提出了相对主元的概念, 并得到相应的 RPCA 方法, 虽然与传统方法相比具有出更多的优良特性, 但还有不少需进一步研究的工作, 例如:

1) 如何基于实际系统来求取比重因子;

2) 如何自适应地选取标准化因子;

3) 相对化变换使得系统的特征结构发生变化, 因此能否找到变换前后特征值之间的关系将是十分有意义的.

References

- Kresta J V, Macgregor J F. Multivariate statistical monitoring of process operating performance. *Canadian Journal of Chemical Engineering*, 1991, **69**: 35–47
- Lu Ning-Yun, Wang Fu-Li, Gao Fu-Rong, Wang Shu. Statistical modeling and online monitoring for batch processes. *Acta Automatica Sinica*, 2006, **32**(3): 400–410 (陆宁云, 王福利, 高福荣, 王姝. 间歇过程的统计建模与在线监测. *自动化学报*, 2006, **32**(3): 400–410)
- Wang Tian-Zhen. Intelligence Fusion Data Mining and Its Application [Ph.D. dissertation], Shanghai Maritime University, 2006. 4–67 (王天真. 智能融合数据挖掘方法及其应用 [博士学位论文], 上海海事大学, 2006. 4–67)
- He Zhao-Shui, Xie Sheng-Li, Zhang Jin-Long. Geometrical algorithm of blind source separation based on QR decomposition. *Control Theory and Applications*, 2005, **22**(1): 17–22 (何昭水, 谢胜利, 章晋龙. 基于 QR 分解的盲源分离几何算法. *控制理论与应用*, 2005, **22**(1): 17–22)
- Johnson R A, Wichern D W. *Applied Multivariate Statistical Analysis*. New Jersey: Prentice-Hall, 1998. 347–387
- Wilkinson J H [Author], Shi Zhong-Ci [Translator]. *The Algebraic Eigenvalue Problem*. Beijing: Science Press, 2001. 10–11 (Wilkinson J H [著], 石钟慈 [译]. 代数特征值问题. 北京: 北京科学出版社, 2001. 10–11)
- Dunia R, Qin S J, Edgar T F, McAvoy T J. Identification of faulty sensors using principal component analysis. *American Institute of Chemical Engineers Journal*, 1996, **42**(10): 2797–2812
- Jolliffe I T. *Principal Component Analysis (Second Edition)*. Berlin: Springer-Verlag, 2002. 10–77
- Gong Huai-Yun, Shou Ji-Lin, Wang Mian-Sen. *Applied Functional Analysis*. Xi'an: Xi'an Jiaotong University, 1985. 128–130 (龚怀云, 寿纪麟, 王绵森. 应用泛函分析. 西安: 西安交通大学出版社, 1985. 128–130)
- Jackson J E. *A User's Guide to Principal Components*. New York: Wiley-Interscience, 1991

- 11 Kano M, Hasebe S, Hashimoto I. Contribution plots for fault identification based on the dissimilarity of process data. In: Proceedings of American Institute of Chemical Engineers Annual Meeting. Los Angeles, USA: 2000. 255
- 12 Legendre L, Legendre P. *Numerical Ecology*. Amsterdam: Elsevier, 1983
- 13 Shen Heng-Fan. *Probability Theory and Mathematical Statistics*. Beijing: High Education Press, 2003. 73–75
(沈恒范. 概率论与数理统计教程. 北京: 高等教育出版社, 2003. 73–75)
- 14 Zhou Dong-Hua, Ye Yin-Zhong. *Modern Fault Diagnosis and Fault Tolerant Control*. Beijing: Tsinghua University Press, 2001. 26–27
(周东华, 叶银忠. 现代故障诊断与容错控制. 北京: 清华大学出版社, 2001. 26–27)
- 15 Wang Song, Xia Shao-Wei. Adaptive robust principal component analysis based on error modeling. *Acta Automatica Sinica*, 1999, **25**(4): 528–531
(王松, 夏绍玮. 基于误差模型的自适应鲁棒主成分分析. 自动化学报, 1999, **25**(4): 528–531)
- 16 Ku W, Store R H, Georgakis C. Disturbance detection and isolation by dynamic principal component analysis. *Chromometrics and Intelligent Laboratory Systems*, 1995, **30**(1): 179–196
- 17 Kourti T, MacGregor J F. Multivariate SPC methods for process and product monitoring. *Journal of Quality Technology*, 1996, **28**(4): 409–428

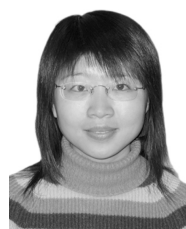


文成林 博士后, 教授. 主要研究方向为多尺度估计理论及其应用和信号处理与数据融合技术及应用和动态系统建模理论及其应用技术. 本文通信作者.
E-mail: wenc1@hdu.edu.cn
(**WEN Cheng-Lin** Post doctor, professor. His research interest covers multiscale estimation theory and its

applications, signal processing, data fusion technology and its applications, and dynamic system modeling theory and its application technology. Corresponding author of this paper.)



胡 静 河南大学硕士研究生. 主要研究方向为信号处理与多模态异常信号检测.
E-mail: hujing@hdu.edu.cn
(**HU Jing** Master student at Henan University. Her research interest covers signal processing and multi-mode anomaly detection.)



王天真 讲师, 上海海事大学博士. 主要研究方向为智能信息处理与故障诊断.
E-mail: tzwang@cle.shmtu.edu.cn
(**WANG Tian-Zhen** Lecturer, Ph.D. at Shanghai Maritime University. Her research interest covers intelligence information processing and fault diagnosis.)



陈志国 教授. 主要研究方向为计算机网络技术与信号处理.
E-mail: chen zg@henu.edu.cn
(**CHEN Zhi-Guo** Professor. His research interest covers computer network technology and signal processing.)