第 19 卷 第 1 期 天 文 学 进 展 Vol.19, No.1
2001 年 3 月 PROGRESS IN ASTRONOMY Mar., 2001

主成分分析法在天体物理中的应用

李 成[1] 孔 旭[1,2,3] 程福臻[1,2]

(1. 中国科学技术大学天体物理中心 合肥 230026)
(2. 中国科学院国家天文观测中心 北京 100012)
(3. 中国科学院 – 北京大学联合北京天体物理中心 北京 100871)

摘 要

主成分分析法是从观测数据中获取主要信息的一种多变量统计方法。它用数目少得多的新变量代替原有观测量，以寻找原观测量之间的相互关系，且不损失原始数据的主要信息，尤其是对于大样本、多参量的情况，该方法更为简捷而有效。目前，主成分分析法被广泛应用于天体物理的诸多研究领域中。介绍了主成分分析法的原理和它在天体物理中的广泛应用。

关 键 词 主成分分析法 — 方差 — 协方差矩阵 — 星系 — 光谱

分 类 号 P14

1 引 言

随着科学研究的逐步深入、观测设备的逐步改进，需要处理的数据也随之剧增。为了从这些看似纷繁复杂的数据中获取尽可能多的有用信息，寻找它们间内在的关系和规律，必须利用一些有效的方法，对这些数据进行简化。然而简化要有一个前提，即做到不损失主要信息。例如，天文学研究课题往往有大量的观测对象，对每个对象观测又可得到若干个参量，这就需要了解这些不同的观测量之间是否存在相互关联，如果存在，又是如何关联的？传统的解决办法是将每一个参量分别与其他所有参量进行分析，找出其中的关系。随着参量数目的增加，这一工作就显得极其复杂。一般只能应付两三个变量的情况，多于 3 个就需寻求其他途径了[1]。在现代天体物理学中，天文学家们常常面对的是具有大量的观测特征的大样本观测数据，简化就显得更为重要。于是一种被称作主成分分析 (Principal Component Analysis, PCA) 的统计方法便应运而生，并被广泛应用。

主成分分析法，或称作 Karhunen—Loève 变换、Hotelling 变换，是一种多变量数据的统计方法[2,3]。对于大样本多参量观测数据，它可以简捷而有效地寻求参量之间的相互关系。这种方法最早被应用于社会科学各研究领域，如描述人的智商高低的参量 IQ 便

国家自然科学基金资助项目 2000-06-12 收到

是运用 PCA 方法得到的 [4,5] 。但它真正被广泛应用于社会科学和自然科学诸多领域是在随着计算机出现和发展的 20 世纪 60 年代以后。 1964 年 Deeming 首次将主成分分析法应用于天体物理领域，他利用主成分分析法发现了一些可以对晚型巨星光谱进行有效分类的特征参数 [6] 。现在，随着 PCA 方法的逐步发展和观测样本数量的剧增，这种方法在天体物理领域中得到越来越广泛的应用，如在星系和恒星的光谱分类、特征参量的挑选、活动星系核光变的研究、大样本天体红移的测量等方面都有不俗的表现。在实践中， PCA 得到了进一步发展完善，推广的 PCA 方法 (Generalized PCA, GPCA) 可以从不完整的观测数据中找出尽可能多的信息 [7] 。事实上，实际观测得来的数据往往是不完整的，对这些数据的研究 GPCA 无疑是强有力的。

2 主成分分析法原理

主成分分析法是一种揭示大样本、多变量数据中各变量或样本之间内在关系的一种方法，其主要作用是降低观测空间的维数，以获取最主要的信息 [8] 。假设我们研究的对象是一个有 n 个天体组成的样本，每个天体有 m 个观测参量，则观测量可表示为矩阵 $\boldsymbol{X}=(x_{ij})_{n\times m}$ 。 PCA 方法类似于多元回归分析，即利用 m 个原始观测变量的线性组合得到的 $m'(m'\ll m)$ 个既能综合反映原来 m 个参量的信息且彼此间又相互独立的新变量来描述原始数据。这些新变量被称作主成分 (principal component, pc) ，与观测量 x 之间的关系可以表示为：

$$pc=\boldsymbol{e}\boldsymbol{X}=e_1x_{k1}+\cdots+e_ix_{ki}+\cdots+e_mx_{km} \tag{1}$$

其中 $\boldsymbol{X}=(x_{ij})_{n\times m}$ 为观测矩阵， pc 是主成分， $\boldsymbol{e}$ 为待求的 m 维特征向量 [17] 。当 $\boldsymbol{e}$ 给定后，对 m 个观测量就可求出一个主成分。主成分 pc 应尽可能多地反映原观测量具有的信息，且彼此互不相关；随机变量的信息可由其方差大小表示，而不同的特征向量 $\boldsymbol{e}$ 可以有不同的方差， PCA 就是寻求使 pc 的方差达到最大的一组特征向量 $\boldsymbol{e}$ 。

初始的观测矩阵 $\boldsymbol{X}$ 为：

$$\boldsymbol{X}=(x_{ij})_{n\times m}=\begin{vmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{vmatrix}$$

其中行矢量对应于同一天体的不同特征量，列矢量对应于不同天体的同一特征量。主成分分析法的任务是寻求使 pc 的方差达到最大的一组特征向量 $\boldsymbol{e}$ 。根据最小二乘法原理，此处的 $\boldsymbol{e}$ 为观测矩阵 $\boldsymbol{X}$ 的协方差矩阵 $\boldsymbol{C}=(c_{jk})_{m\times m}$ 的正交特征矢量，其中

$$c_{jk}=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)(x_{ik}-\overline{x}_k)\quad 1\le j,k\le m \tag{2}$$

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}; \qquad \overline{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik} \tag{3}$$

$\overline{x}_j, \overline{x}_k$ 为列矢量的平均值。通过求解行列式方程 $|\boldsymbol{C} - l\boldsymbol{I}| = 0$，我们将求得满足方程 (1) 的组合，其中 l 为行列式的特征根，$\boldsymbol{I}$ 为 $(m \times m)$ 的单位矩阵。求得特征根后，再由下式求得特征矢量 $\boldsymbol{e}$。

$$(\boldsymbol{C} - l\boldsymbol{I})\boldsymbol{e}_i = 0 \tag{4}$$

方程 $|\boldsymbol{C} - l\boldsymbol{I}| = 0$ 有 m 个特征根，根据它们的大小顺序，表示为 $l_1 \geq l_2 \geq l_3 \geq \cdots \geq l_m \geq 0$。对应于每一个特征根 l_i，有一个特征矢量 $\boldsymbol{e}_i$，可以得到一个主成分 pc_i，称作第 i 个主成分。其中最大的特征根 l_1 所对应的是第一主成分。因为 $l_1 + l_2 + l_3 + \cdots + l_m = \mathrm{tr}\boldsymbol{C}, \mathrm{tr}\boldsymbol{C}$ 为协方差矩阵 $\boldsymbol{C}$ 的迹，所以在主成分分析中，常称 $l_k / \sum_{i=1}^{m} l_i$ 为主成分 pc_k 的贡献率或相对贡献率，而称 $\sum_{j=1}^{k} l_j / \sum_{i=1}^{m} l_i$ 为主成分 $pc_1, pc_2, \cdots, pc_k (k \leq m)$ 的累积贡献率。

由于主成分分析的目的就在于用尽可能少的主成分来代替原有变量，并尽可能充分反映原变量的信息，所以主成分个数 k 的选取应按下列两条准则之一进行：(1) 使主成分的累积贡献率达 80% 以上；(2) 主成分对应于 $l \geq \bar{l}$(或 1)，其中 $\bar{l}$ 为特征根的均值。主成分分析法的计算步骤主要包括：(1) 计算观测样本矩阵 $\boldsymbol{X}$ 的协方差矩阵或相关矩阵；(2) 求协方差矩阵或相关矩阵的特征根和特征根所对应的特征矢量；(3) 由特征根数值，计算不同特征根的相对贡献率和累积贡献率，确定主成分的个数；(4) 对于选定的主成分，由特征矢量计算主成分。

在实际研究过程中，我们经常会利用各种不同的观测样本数据。由于不同的样本原初研究目的不同，所以观测的精度和观测量的种类都不尽相同，要想得到一个由不同样本构成的统一的、准确的观测数据样本基本上不可能。如何从这些不完备的数据中尽量多的获取有用信息呢？主成分分析法的一种推广形式 — 推广的主成分分析法 (Generalized PCA, GPCA) 正是可以完成这类工作的有效方法。GPCA 利用权重因子法来补充原始数据中缺省的观测量和观测精度较差的量，使得复合体系中多个变量的分布为最可几分布，且补充的数据不会改变原始数据的统计性质。再利用 PCA 方法分析补充后的数据样本，获得主成分信息 [7]。

3　主成分分析法的应用

3.1　星系谱的分类

随着大样本光谱巡天工作的开展，在不久的将来，就可收集到大量的星系的光谱。因为星系的光谱中包含丰富的信息，如果能够对这些光谱进行分类研究，这将有助于星系光谱特性的演化和与之相关的物理过程的研究。但由于星系光谱中有很多特征量，如吸收线、发射线和连续谱，使得对星系光谱进行有效的分类变得困难，如何从这些特征量中选择有效的分类判据就显得很重要。在众多的光谱分类方法中，PCA 方法是寻找这些判据并对星系光谱进行分类的最有效的手段之一。假如有 n 个星系光谱，每个光谱中有 m 个波长点，这样就组成一个 $(n \times m)$ 的矩阵，利用 PCA 方法，可以得到对星系光谱

分类所需的特征量数目和数目较小的正交组分 (特征谱)。研究发现，这些特征谱与星系的物理特性 (如恒星形成率、星族年龄等) 紧密相关。将观测的星系谱投影到这些正交的特征谱上，可以得到不同星族对星系的相对贡献，从而可以对星系光谱进行分类。因为星系的光谱是其中组成恒星特征的集体表现，所以这种对星系光谱进行分类的方法很自然 [9]。因此，现在 PCA 法被广泛应用于对星系光谱进行分类。

Sodré 和 Cuevas(1994) 利用 PCA 方法分析了 24 个正常星系的光谱，发现前面的 5 个主成分的累积贡献率达 90%，其中第一主成分与星系 Hubble 型 (形态) 的相关系数为 0.87。这表明星系的光谱和形态之间存在明显的相关性，可以利用星系的形态对星系光谱进行简单的分类。进一步研究表明，星系光谱中的一些与金属丰度或温度相关的特征量 (如 4000Åbreak 的幅值、G-band 的强度或 Mg2 谱指数) 与星系的 Hubble 型相关，它们也可以作为对星系光谱进行分类的特征量 [11]。

另外，Connolly 等人利用 PCA 方法分析了 31 个正常星系、39 个星暴星系、结果发现前 2 个特征谱的贡献重要，由它们可以对星系光谱进行分类；对于正常星系，可以用一个参数给予分类 [12]。利用 PCA 方法和神经网络技术，Folkes 等人发展了一套可以对低信噪比光谱进行自动分类的程序 [13]。利用星族合成方法和主成分分析法，Sodré 研究了 55 个星系光谱，发现只需单一参数即星系中恒星形成时标就可以对不同星系的光谱进行很好的分类 [14]。

同样的方法也可应用于类星体的光谱分类。为了描述类星体 (QSO) 的连续谱和发射线的特征，需要数十个参量。为了寻找可以作为对 QSO 进行分类的特征量，Francis 等人将 PCA 方法应用于 232 个 QSO 样本谱，结果发现了三个特征量：发射线线心强度、连续谱斜率和宽吸收线特征为主成分，并利用这三个主成分对 QSO 光谱进行了分类 [10]。

3.2 选取特征量

PCA 法在天体物理研究中另一个被广泛应用的方面是从大量的观测量中挑选出对天体特性起主导作用的特征量。利用 PCA 方法，知道了描述观测样本需要的主成分 (pc_j) 数目，也就可以知道确定样本最少所需的特征量数目。另外，第一主成分 pc_1 的作用最大，反映了样本的主要信息。根据公式 (1)，第一主成分 pc_1 可表示为：$pc_1 = \boldsymbol{e}_1\boldsymbol{X} = e_{11}x_{k1} + \cdots + e_{1i}x_{ki} + \cdots + e_{1m}x_{km}$，如果 e_{1i} 的值较大，则表明 x_{ki} 对第一主成分的贡献较大，就可以根据 e_{1i} 选取在观测样本中起主导作用的特征量。

Dultzin-Hacyan 和 Ruano 利用 PCA 法研究了 959 个 Seyfert 1 和 Seyfert 2 星系的多波段能谱分布 (SED) 特征。PCA 结果表明：对于 Seyfert 1 星系，第一主成分 pc_1 的贡献率 $\approx 90\%$，其它主成分的贡献较小；对于 Seyfert 2 星系，前三个主成分的贡献都比较重要。PCA 结果说明 Seyfert 1 星系的 SED 只需一个参量描述，而 Seyfert 2 星系的 SED 至少需要三个参量描述，他们认为导致这种差异的原因是因为 Seyfer 1 星系的 SED 仅受到中心大质量黑洞影响，而 Seyfert 2 星系除了受黑洞的影响，可能会受到核区周围的恒星形成区和星际介质的再辐射影响 [15]。

Francis 和 Wills 利用 PCA 方法分析了 22 个 QSOs 的 13 个特征量。将样本数据构造成 22×13 的矩阵，按照 PCA 方法的计算步骤，得到的结果如表 1 所示。表 1a 中，第 2~6

表 1a　前 5 个主成分的特征值和贡献率 [1]

	pc_1	pc_2	pc_3	pc_4	pc_5
特征值	6.4505	2.8157	1.5879	0.6257	0.5698
贡献率	0.496	0.217	0.122	0.048	0.044
累积贡献率	0.496	0.713	0.835	0.883	0.927

表 1b　前 5 个主成分与原始变量的相关系数 [1]

原始变量	pc_1	pc_2	pc_3	pc_4	pc_5
$\log L_{1216}$	0.053	0.535	−0.123	−0.029	−0.405
α_x	0.295	−0.198	0.079	0.485	−0.155
$FWHM$ Hβ	−0.330	0.077	−0.357	−0.082	−0.141
FeII/Hβ	0.341	−0.140	0.003	−0.487	−0.212
$\log EW$ [OIII]	−0.310	0.016	0.255	0.394	−0.095
$\log FWHM$ CIII]	−0.198	0.077	−0.623	0.054	0.402
$\log EW$ Lyα	−0.177	−0.502	−0.006	−0.143	0.033
$\log EW$ CIV	−0.336	−0.262	0.048	−0.050	−0.303
C IV/Lyα	−0.342	0.062	0.025	−0.074	−0.584
$\log EW$ C III]	−0.262	−0.413	−0.124	−0.176	−0.008
Si III]/C III]	0.342	−0.149	−0.018	−0.311	−0.116
N V/Ly α	0.231	−0.050	−0.573	0.107	−0.288
λ1400/Ly α	0.223	−0.351	−0.225	0.441	−0.216

列是所有 13 个主成分的前 5 个。第一行给出了各主成分的特征值，第二、三行分别是各主成分的贡献率和累积贡献率。由表中可以看出，第一主成分 pc_1 的贡献最大，达到近 50%；前 3 个主成分累积贡献达 84%，表明前 3 个主成分的作用很大。另外，表中还给出了 13 个原始变量与各主成分的相关系数，由方程 (1) 可以得到：$pc_1 = 0.053\times x_1 + 0.295\times x_2 - 0.330 x_3 + \cdots$，这里 $x_1, x_2, x_3, \cdots$ 分别是 13 个原始变量的值。对于每个主成分，13 个相关系数所构成的矢量就是其特征矢量，可以验证，13 个相关系数的平方和等于 1 。一般说来，相关系数绝对值越大，则表明该原始变量对相应的主成分越重要。由表中可以看出，对于第一主成分，Fe II/Hβ, C IV/Ly α 等特征量贡献很重要 [1] 。

近十多年中，星族合成方法在研究星系中的恒星成分、恒星形成历史、星系的形成和演化等方面取得了一些成果。但是在利用星族合成方法研究星系特征时，一个很棘手的问题是年龄和金属丰度存在耦合效应 [16] 。为解决这个问题，Kong 和 Cheng 利用 PCA 方法研究了大量简单星族 (SSP) 的谱指数，他们发现了一些可以较精确确定星族年龄的谱指数，如 Hβ 、 $C_2$4668 、 G4300 、 Fe4383 和 Mg b ，且谱指数对星族年龄的敏感程度与星族的金属丰度有关。 Hβ 、 G4300 适合确定金属丰度较低的星族年龄，其它谱指数适合确定金属丰度较高的星族年龄。这些谱指数将有助于我们区分年龄和金属丰度的耦合效应 [17] 。

3.3　活动星系核的光变研究

对活动星系核 (AGN) 的多波段国际联测数据分析研究，发现 AGN 发射线的变化是对连续谱变化的响应，这种响应关系表明 AGN 光谱中的发射线起源于中心电离源对其周围发射线云团的光致电离 [18] 。对 AGN 紫外谱线进一步详细观测和研究发现，UV 谱线中有多种独立的成分，每个成分随时间的变化模式不同 [19] 。为了深入理解 AGN 发射线区的结构及其运动，希望能将发射线分解成几个相互独立的成分。经典的方法是利用多高斯成分拟合谱线轮廓，将其分解成几个成分，但这种分解并不能确定这些成分是否独立变化。而 PCA 方法得到的主成分之间是正交的和互不相关的，所以利用 PCA 方法可以分解 AGN 光谱中的发射线。与 RM(reverberation mapping) 方法 [20] 和交叉相关分析法比较，PCA 法对观测数据的精度要求更低，可以研究那些观测点较少的资料 [21] 。

为了研究 NGC4151 紫外谱线中不同成分随连续谱的变化特征，Mittaz 等人首次利用 PCA 法来分解 AGN 的谱线，并发现了谱线中十个比较重要的主成分。由于大部分主成分的物理意义未能得到解释，所以只能对谱分析提供一些有限的应用 [19] 。利用相似的方法，Turler 和 Courvoisier 研究了 18 个 IUE 卫星观测得到的 AGN 长时间的监测谱 (每个源的观测多于 15 次) 。与 Mittaz 等人不同的是，Turler 和 Courvoisier 在利用 PCA 法时仅将 UV 谱中的 Ly α 和 C IVλ1549 线分解成两个互不相关的成分：主成分和剩余成分。这样做使得成分的物理解释变得容易得多，并且证实了 PCA 是研究发射线变化和发射线区结构和运动学的好方法。对主成分、剩余成分与 UV 连续谱进行相关分析发现，主成分的变化确实和连续谱存在很强的关联，而剩余成分则不然。对 NGC 5548 分析发现，剩余成分与 UV 连续谱之间的关系表明剩余成分对连续谱响应存在约 25d 的延迟。这个延迟是连续谱和 Ly α 、C IV 谱线的延迟 (8 $\sim$ 16d) 的两倍。这意味着 PCA 将谱线分解成两个部分：一个部分与连续谱相对应，只有很小的延迟 (小于 5d) ，另一部分则有较长延迟 ($\approx$ 25d) ，结果表明发射线的主成分和剩余成分来自宽线区 (BLR) 的内、外不同区域。另外，随星系的光度增加，主成分占有比下降，这种趋势的可能解释是光度较大的星系 BLR 的尺寸变大 [21] 。

NLS1 星系是具有活动星系核的 Seyfert 星系的一个子类，具有 Seyfert 1 或 1.5 型星系的特点，但有着不寻常的窄 HI 线。为了研究 NLS1 星系 Akn 564 的光变和物理特性，目前正在对它进行多波段特性的国际联测。PCA 方法在对其观测资料进行处理时，显然是大有帮助的。

3.4 大样本天体的红移测量

随着多光纤技术的发展和完善，使得同时观测数以千计的天体光谱成为可能。基于这种技术，现在很多大样本的天体光谱巡天工作正在或将要开始，如 2dF(two degree field) 红移巡天和我国的 LAMOST 巡天计划。这些大样本的红移巡天，将有助于我们更加深入地研究大尺度结构、星系的成团和演化 [22] 。随着这些巡天工作的开展，我们将得到大量星系的光谱，一个迫切需要解决的问题是发展一种能够快速、准确测量星系光谱红移的方法。

现在，测量星系光谱红移最常用的方法是交叉相关分析法 [23] 。这种方法是利用星系谱和一系列的星系模板谱进行交叉相关，交叉相关函数的最大的峰值表示星系谱和模板谱之间的匹配程度。峰的位置和宽度表征星系谱的红移值及其误差。如果星系谱和模板谱都足够好，交叉相关函数有一个尖锐的峰，确定的红移值就比较准确，但实际情况是星系谱一般不能用任何模板谱作很好拟合，使得确定的星系红移值的误差较大 [24] 。

为了准确测量大样本星系光谱的红移值， Glazebrook 等人利用了主成分分析法。与经典的交叉相关法不同， PCA 不是利用单个模板谱与星系光谱进行比较，而是利用 PCA 方法分析大量的星系模板谱，从中寻找不同类型星系谱的主成分，扣除噪声的影响；再利用多个主成分来分析红移待测的星系光谱，确定其红移值。为了检验 PCA 法和交叉相关分析法确定星系光谱红移的优劣， Glazebrook 将这两种方法运用到 104 个红移已知的星系谱，结果表明：对信噪比较高的星系谱，两种方法的结果一致；对信噪比较低的星系谱， PCA 法明显优于交叉相关法 [25] 。

目前我国正在建设大天区面积多目标光纤光谱天文望远镜 (LAMOST) ，其多达 4000 根光纤的光谱仪可以同时观测 4000 个天体的光谱， PCA 方法将有助于我们确定星系光谱的红移。

4　结　　论

综上所述，作为一种有效的多变量分析方法， PCA 可以从大样本多变量数据中发现一些主要特征量，寻找他们的相互关系，并用少得多的新变量来代替原有的观测量，使得我们可以从浩如烟海的观测数据中获取尽可能多的有用信息。本文虽只简要介绍了 PCA 方法在光谱分类、寻找特征量、 AGN 光变和星系红移确定等方面的应用，但 PCA 方法在天体物理中的应用远不只是这些，可以说，只要是多变量大样本 (小样本当然也行) 的观测数据，均可运用 PCA 方法进行处理、分析。

参　考　文　献

1　Francis P J, Wills B J. ASP Conf. Ser., 1999, 162: 363
2　Karhunen H. Ann. Acad. Science Fenn, 1947, Ser. A. I. 37
3　Loeve M. Processus Stochastiques et Mouvenent Brownien, Paris: Hermann, 1948
4　Hotelling H. J. Educ. Psych. 1933, 24: 417
5　Kendall M G. A Course in Multivate Analysis, London: Griffin, 1957
6　Deeming T J. M.N.R.A.S., 1964, 127: 493
7　Unno W, Yuasa M. Astrophys. Space Sci., 1992, 189: 271
8　Maćkiewicz A, Ratajczak W. Computers & Geosciences, 1993, 19: 303
9　Connolly A J, Szalay A S. A. J., 1999, 117: 2052
10　Francis P J *et al.* Ap. J., 1992, 398: 476
11　Sodré L, Cuevas H. Vistas Astron., 1994, 38: 287
12　Connolly A J *et al.* A. J., 1995, 110: 1071
13　Folkes S R, Lahav O, Maddox S J. M.N.R.A.S., 1996, 283: 651
14　Sodré L, Cuevas H. M.N.R.A.S., 1997, 287: 137

15 Dultzin-Hacyan D, Ruano C. Astron. Astrophys., 1996, 305: 719
16 Vazdekis A *et al.* Ap. J. Suppl. Ser., 1997, 111: 203
17 Kong X, Cheng F Z. Chin. Phys. Lett., 2000, 17: 700
18 Rodríguez-Pascual P M, Alloin D, Clavel J *et al.* Ap. J. Suppl. Ser., 1997, 110: 9
19 Mittaz J P D, Penston M V, Snijders M A J. M.N.R.A.S., 1990, 242: 370
20 Peterso B M. Publ. Astron. Soc. Pac., 1993, 105: 247
21 Turler M, Courvoisier T J L. Astron. Astrophys., 1998, 329: 863
22 Strauss M A. In: Dekel A, Ostriker J P eds. Structure Formation in the Universe, American Astronomical Society meeting 188, 1996, http://xxx.itp.ac.cn/abs/Astro-ph/9610033
23 Tonry J, Davis M. Ap. J., 1979, 84: 1511
24 Heavens A F. M.N.R.A.S., 1993, 263: 735
25 Glazebrook K, Offer A R, Deeley K. Ap. J., 1998, 492: 98

The Application of Principal Components Analysis to Astrophysics

Li Cheng[1] Kong Xu[1,2,3] Cheng Fuzhen[1,2]

(1. Center for Astrophysics, University of Science and Technology of China, Hefei 230026)

(2. National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012)

(3. Beijing Astrophysics Center (BAC), Beijing 100871)

Abstract

Principal Component Analysis (PCA) is a main multivariate statistical method for getting principal information from observational data. It uses few new variables instead of initial parameters, in order to find out the relations among the inital parameters, without losing the main information of intial data. Especially for the case of large sample and multivariate, this method is simpler and more efficient. In the present day, PCA is applied widely in many research fields of astrophysics. The main principle and applications to astrophysics of PCA are reviewed in this paper.

Key words principal component analysis—variance—covariance matrix—galaxy—spectrum