人类群体遗传结构的协方差阵主成分分析方法

薛付忠[1], 王洁贞[1], 郭亦寿[2], 胡 平[1]

(1. 山东大学公共卫生学院流行病学与医学统计学研究所, 济南 250012;
2. 山东大学医学院医学遗传学研究所, 济南 250012)

摘要: 目的: 探讨基因频率矩阵的中心化(或均值化)协方差阵主成分分析方法在人类群体遗传结构研究中的适用性和合理性。方法: 从基因频率矩阵的结构特征入手, 分析中心化、均值化协方差阵主成分分析与标准化相关阵主成分分析在特征根、特征向量以及降维效果等方面的差异, 并通过实例比较不同方法在解释群体遗传结构特征上合理性。结果: 中心化(或均值化)协方差阵的主成分不仅反映了基因变异程度的"方差信息量权", 而且反映了基因间相互影响程度的"相关信息量权"; 标准化相关阵的主成分反映的仅是"相关信息量权", 不包括"方差信息量权"。通过比较中国 26 个汉族人群 HLA-A 基因座中心化协方差阵和标准化相关阵 2 种主成分分析结果, 证实中心化协方差阵主成分分析方法在特征根与特征向量、保留主成分的个数和对主成分的群体遗传学解释的合理性等方面均优于标准化相关阵主成分分析方法。结论: 在对群体遗传结构进行主成分分析时, 应使用中心化(或均值化)变换消除基因频率矩阵中量级的影响, 然后在用其协方差阵提取主成分。

关键词: 人类群体遗传结构; 主成分分析; 中心化(或均值化)协方差阵; HLA-A
中图法分类号: Q987 文献标识码: A 文章编号: 1000-3193(2005)03-0221-11

某群体某基因座上各等位基因的频率即为该群体该基因座的遗传结构; 其全部基因座的基因频率即为该人群的整个遗传结构。所以, 人群遗传结构的差异, 实质上是基因频率的差异[1]。人类群体遗传的变异水平受人口迁移、人群融合、自然选择、遗传漂变、地理和社会隔离、突变等多因素的影响, 这些因素的综合作用构成了群体内或群体间某基因座或多基因座的遗传结构。遗传结构数据可用基因频率矩阵表示。由于群体内或群体间广泛存在基因多态性, 故一个基因座上可有多个等位基因, 多个基因座所包含的等位基因数目和组合方式更是复杂多样。因此, 如果仅用某基因座中的一个等位基因分析其群体遗传结构状态, 则不能充分利用其遗传变异信息, 结论是片面的, 须对每个基因进行分析。但是, 各基因分析的结果往往不同, 难以得出统一的结论。鉴于基因频率矩阵中各基因间常存在着某种联系, 利用这种联系可找出能基本反映原矩阵信息的少数几个综合指标, 进而由其反映基因座的群体遗传结构。主成分分析就是解决这类问题的多元统计方法, 已在群体遗传学研究中广泛应用[2-3]。在主成分分析中, 提取主成分的方法很多。例如, 在功能强大的 SAS 软件中, 除可用相关阵求取主成分外, 也可用协方差阵、偏相关阵、偏协方差阵等求取主成分。但是, 应用不同矩阵所得出的特征根与特征向量、所保留主成分的个数以及对主成分的解释均有很

收稿日期: 2004-09-13; 定稿日期: 2005-06-18
基金项目: 国家自然科学基金资助项目(30170527)
作者简介: 薛付忠(1964–), 男, 山东沂水县人, 医学博士, 副教授, 主要从事人类群体遗传空间结构统计分析方法的研究。

大区别。因此, 如果不考虑基因频率矩阵的结构特点, 简单套用主成分分析模型时, 其研究结果往往不能真实反映群体的遗传结构特点。作者从分析基因频率矩阵的结构特点入手, 曾提出了人类群体遗传结构的非线性主成分分析方法[4]。本文将进一步研究如何选择提取基因频率矩阵主成分的方法。

1 基因频率矩阵及其结构特点

1.1 基因频率矩阵的结构模式

设由 n 个群体的 m 个基因组成的基因频率矩阵为:

$$P_{n\times m}=\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nm} \end{pmatrix} \quad (1)$$

式中 p_{ij} 表示第 i 个群体第 j 个等位基因的频率, $i=1, 2, \cdots, n$ 个群体, 行向量 $P_i=(x_{i1}, x_{i2}, \cdots\cdots, x_{im})$ 为基因频率向量; $j=1, 2, \cdots, m$, 单一基因座时 m 为该基因座上的基因个数, 多基因座时 m 为多个基因座上的所有基因个数。

1.2 基因频率矩阵的结构特点

上述基因频率矩阵具有以下结构特点:

(1) 基因频率矩阵 $P_{n\times m}$ 的各元素均非负, 且每个群体的 m 个等位基因的频率之和为一常数(当基因频率矩阵由单一基因座的多态性数据组成时, 其和恒等于 1; 当基因频率矩阵由 l 个基因座的多态性数据组成时, 其和恒等于 l), 故称其为“ 成分数据” 或“ 定和数据” 或“ 闭合数据”;

(2) 因各群体的样本含量不够大或稀有基因的存在, 矩阵 $P_{n\times m}$ 内的某些元素(基因频率) 常常为零;

(3) 基因频率数据(闭合数据) 的取值范围[0, 1], 矩阵 $P_{n\times m}$ 所对应的空间在数学上称为“ 单形” (simplex)[5—7], 其维数是 $m-1$。通常是非线性的, 且多数情况下不服从正态分布, 而是服从 Dirichlet 分布。这种“ 单形性” 在矩阵维数较低时尤为突出;

(4) 基因频率数据矩阵 $P_{n\times m}$ 的行和列均无量纲, 但数据间的数量级差别很大。例如, 某群体中常见基因的频率为0.419 6, 而某稀有基因的频率为0.000 3, 二者在数量级上相差 599 倍;

(5) 基因频率矩阵 $P_{n\times m}$ 包含了 n 个群体间和各群体内在一个或多个基因座上的全部变异信息。其中, 各行间的差异反映的是 n 个群体间的遗传变异; 各列间的差异反映的是基因间或基因座间的变异和相互关系, 以及群体间的变异信息。

2 基因频率矩阵主成分分析方法的选择

2.1 基因频率矩阵的“ 闭合效应” 及其非线性主成分分析

由于基因频率矩阵存在“ 闭合效应”, 其协方差矩阵具有明显的负偏性, 因而影响线性主

成分分析的结果。作者曾基于“均值化对数比”变换提出了群体遗传结构的“中心化对数比”非线性主成分分析方法[4]。但通过大量资料分析后发现, 当基因频率矩阵中的基因数目较多(即某基因座的多态性极其丰富, 能检出为数众多的等位基因) 时, “闭合效应”的作用会大为减弱。因此, 当基因频率矩阵的“闭合效应”很弱或基因频率矩阵为“开放数据”时, 可不经“均值化对数比”变换, 直接进行线性主成分分析即可。

2.2　基因频率矩阵主成分提取方法的选择

提取基因频率矩阵主成分的关键是通过各基因的线性组合以概括原始基因频率矩阵中的变异信息, 即降维。例如, 由 n 个群体 m 个基因组成的基因频率矩阵 $P_{n\times m}$(式 1) 的线性组合为:

$$z_{ik} = c_1 p_{i1} + c_2 p_{i2} + \cdots + c_p p_{im} \tag{2}$$

其中, z_{ik} 是第 i 个群体($i = 1, 2, \cdots, n$) 第 k 个主成分($k = 1, 2, \cdots, K$) 的取值, $c_1, c_2, \cdots, c_m$ 是权重向量(表示每个基因对线性组合的贡献大小)。该过程可通过求取基因频率矩阵的“特征根”和“特征向量”来实现。但是, 在进行主成分分析时, 不同矩阵形式所得的特征根与特征向量、保留主成分的个数以及对主成分的解释都有很大区别。因此, 如何求取基因频率矩阵的特征根和特征向量却是十分复杂和有待深入分析的问题, 了解各种求取方法的原理和特点是正确求取主成分的前提。

前已述及, 基因频率矩阵 $P_{n\times m}$ 的行和列均无量纲, 但数据间的量级差别很大, 必然产生以大压小的现象, 从而导致其协方差阵的退化, 影响特征向量和特征根的提取。因此, 需对原始基因频率数据进行变量变换(如标准化变换, 均值化变换和中心化变换等) , 以消除量级的影响。由于变量变换中经常使用的除法实质上是几何上的相似变换, 改变了数据结构, 从而改变了数据的变异信息, 所以变换前后的方差之和不相等, 都会在一定程度上丢失基因频率矩阵所包含的遗传信息[8—9]。

设原始基因频率矩阵为 $P = (p_{ij})_{n\times m}$, 变换后的数据矩阵为 $Z = (z_{ij})_{n\times m}$, $D(p_j)$, $D(z_j)$ 分别是变换前后的方差, $r_{p_i p_j}$, $r_{z_i z_j}$ 分别表示变换前后的相关系数。

当采用标准化变换时, 取 $z_{ij} = (x_{ij} - \bar{p}_j)/\sqrt{D(p_j)}$, 则

$$D(z_j) = \frac{1}{D(p_j)} D(p_j) \tag{3}$$

当采用均值化变换时, 取 $z_{ij} = p_{ij}/\bar{p}_j$, 则

$$D(z_j) = \frac{1}{\bar{p}_j^2} D(p_j) \tag{4}$$

当采用中心化变换时, 取 $z_{ij} = p_{ij} - \bar{p}_j$, 则

$$D(z_j) = D(p_j) \tag{5}$$

可以证明: 三种变换前后的相关系数并未改变, 即 $r_{p_i p_j} = r_{z_i z_j}$, 但标准化变化和均值化变换的方差发生了变化, 二者都使原始数据的方差变小, 都会在一定程度上丢失原始基因频率矩阵所包含的遗传结构信息。

采取何种变换求取主成分, 必须从分析基因频率矩阵的结构特点入手, 并兼顾群体遗传结构分析的目的。基因频率矩阵 $P_{n\times m}$ 中包含两方面遗传结构的信息: 一是由各基因的频率方差大小来反映的基因在群体间变异信息; 二是由相关系数矩阵来体现来反映的各基因间

的相关信息。

(1) 若用标准化变换, 则必须从相关系数矩阵求取主成分。但是, 标准化变换使各基因的频率方差全部为 1, 消除了各基因在群体间变异程度上的差异。因此, 由相关阵提取的主成分, 只是包含了各基因间相互影响的信息, 不能准确反映原始基因频率矩阵所包含的全部遗传结构信息。由统计学知识可知, 聚集程度和离散程度是变量的基本性质。而离散程度常以方差的大小来反映, 即方差越大, 表示观察对象在该指标所体现的属性上差异越大, 方差是反映属性变异的最基本的信息, 相关只度量变量间联系的强度。基于相关阵提取的主成分中各基因的系数并不是由该基因的频率方差信息决定, 而是由该基因与其它基因的相关程度决定, 因此该种主成分仅包含"相关信息量权"。不包含"方差信息量权"。

(2) 相比之下, 均值化变换虽然也损失部分信息, 但有其优越性, 适于对基因频率矩阵进行降维处理。均值化变换后, 基因频率的协方差阵 $Cov=(u_{ij})_{m\times m}$ 的元素为: $u_{ij}=\frac{1}{n-1}\sum_{l=1}^{n}(z_{li}-\bar{z}_i)(z_{lj}-\bar{z}_j)$。由上述均值化变换的公式知[9—10], 均值化变换后各指标的均值为 1, 所以 $u_{ij}=\frac{1}{n-1}\sum_{l=1}^{n}(z_{li}-1)(z_{lj}-1)=\frac{1}{n-1}\sum_{l=1}^{n}\frac{(p_{li}-\bar{p}_i)(p_{lj}-\bar{p}_j)}{\bar{p}_i\bar{p}_j}=\frac{s_{ij}^2}{\bar{p}_i\bar{p}_j}$, s_{ij}^2 为原始数据的协方差。特别地, 当 $i=j$ 时为: $u_{ij}=\frac{s_{ii}^2}{\bar{p}_i^2}=\left(\frac{s_{ii}}{\bar{p}_i}\right)^2$, $s_{ii}^2=\frac{1}{n}\sum_{l=1}^{n}(p_{li}-\bar{p}_i)^2$。因此, 均值化变换后基因频率的协方差矩阵的对角元素是各基因频率的变异系数的平方, 它反映各基因变异程度上的差异。均值化变换前, 反映各基因间相互影响程度的相关系数 r_{ij} 为: $r_{ij}=\frac{s_{ij}^2}{\sqrt{s_{ii}^2}\cdot\sqrt{s_{jj}^2}}$, 均值化变换后相关系数 r'_{ij} 为: $r'_{ij}=\frac{u_{ij}^2}{\sqrt{u_{ii}^2}\cdot\sqrt{u_{jj}^2}}$, 将 $u_{ij}=\frac{1}{n-1}\sum_{l=1}^{n}(z_{li}-\bar{z}_i)(z_{lj}-\bar{z}_j)$ 代入, 可得

$$r'_{ij}=\frac{(s_{ij}^2/\bar{p}_i\bar{p}_j)}{\sqrt{s_{ii}^2/\bar{p}_i}\cdot\sqrt{s_{jj}^2/\bar{p}_j}}=\frac{s_{ij}^2}{\sqrt{s_{ii}^2}\cdot\sqrt{s_{jj}^2}}=r_{ij} \quad (6)$$

所以, 均值化变换不改变各基因间的相关系数, 相关系数矩阵的全部信息都在相应的协方差矩阵中得到反映。均值化变换后的协方差矩阵不仅消除了数量级的影响, 还同时包含了基因变异程度的差异信息和各基因间相互影响程度上的相关信息。

(3) 由于基因频率矩阵 $P_{n\times m}$ 无量纲, 因此, 采用中心化变换(即 $z_{ij}=p_{ij}-\bar{p}_j$) 即可消除数量级的影响。中心化变换较均值化变换更加优越, 因为在变化过程中没有使用除法, 所以变换后的方差不变, 原始基因频率矩阵中的遗传结构信息全部反映在中心化变换矩阵中。在求取基因频率矩阵的"特征值"和"特征向量"时宜使用中心化或均值化变换消除数量级的影响, 然后利用协方差阵求取"特征值"和特征向量"。

3 实例分析——中国 26 个汉族人群 HLA-A 基因座群体遗传空间结构的主成分分析

3.1 群体遗传学资料

根据不同地理环境, 收集中国 26 个汉族群体的 HLA-A 基因多态性群体遗传学调查数据(因参考文献过多, 文后未列出资料来源) , 以各基因的基因频率为指标构成中国汉族

HLA-A 基因座的基因频率矩阵。标准为: ①样本含量大于 100; ②分别对每个人群的基因频率数据进行 χ^2 检验, 剔除不符合 Hardy-Weinberg 定律者。该基因座中的等位基因包括 **A1, A2** (A203), **A3, A5, A9** (A23, A24, A 2403), **A10** (A25, A26, A34, A66), **A11** (A11.1, a11.2), **A19** (A29, A30, A32, A33, A34, A74), **A28** (A68, A69), **A36, A43**. (表 1)

表 **1**　中国 **26** 个汉族人群 **HLA-A** 基因座基因频率分布

The allele frequency of HLA-A locus in 26 Chinese Han populations

群体 Population	地点 Location	基因频率(Allele Frequency)									
		A1	A2	A3	A5	A9	A10	A11	A19	A28	A36
Anhui Han	蚌埠	0.0295	0.2656	0.0349	0.0000	0.1715	0.0298	0.1715	0.1656	0.0098	0.0000
Beijing Han	北京	0.0433	0.3275	0.0483	0.0000	0.1621	0.0367	0.1636	0.1431	0.0179	0.0000
Fujian Han	闽南	0.0000	0.2929	0.0000	0.0000	0.1515	0.0305	0.3144	0.1927	0.0000	0.0000
Gansu Han	泾川县	0.0646	0.2468	0.0697	0.0000	0.2342	0.0801	0.1622	0.1512	0.0293	0.0000
Guangdong Han	广州	0.0049	0.3078	0.0111	0.0000	0.1676	0.0312	0.3331	0.1291	0.0049	0.0000
Guangxi Han	广西	0.0051	0.3110	0.0102	0.0000	0.1169	0.0102	0.3644	0.1039	0.0051	0.0000
Guizhou Han	贵阳	0.0129	0.3113	0.0129	0.0000	0.1452	0.0194	0.3020	0.1603	0.0064	0.0000
Hainan Han	海南	0.0047	0.2995	0.0093	0.0000	0.1723	0.0426	0.2864	0.0140	0.0093	0.0000
Hebei Han	唐山	0.0308	0.2546	0.0955	0.0000	0.1652	0.0572	0.1472	0.0887	0.0051	0.0000
Henan Han	洛阳	0.0767	0.1882	0.1018	0.0000	0.2229	0.0057	0.1882	0.1190	0.0057	0.0057
Heilongjiang Han	佳木斯	0.0513	0.3072	0.0461	0.0000	0.1815	0.0251	0.0620	0.0610	0.0151	0.0000
Hubei Han	湖北	0.0269	0.3352	0.0429	0.0000	0.1415	0.0429	0.2530	0.0879	0.0055	0.0000
Hunan Han	长沙	0.0145	0.3349	0.0243	0.0000	0.1702	0.0493	0.2997	0.0392	0.0096	0.0000
Jilin Han	长春市	0.0461	0.3000	0.0461	0.0000	0.1876	0.0253	0.1515	0.1540	0.1010	0.0000
Jiangsu Han	徐州	0.0400	0.2141	0.0248	0.0000	0.1775	0.0451	0.1775	0.1438	0.0000	0.0000
Liaoning Han	沈阳	0.0202	0.3073	0.0101	0.0000	0.1754	0.0151	0.1693	0.0726	0.0461	0.0000
Nei Mongol Han	呼和浩特	0.0601	0.3809	0.0426	0.0000	0.1056	0.0084	0.1149	0.0084	0.0084	0.0000
Shandong Han	山东	0.0742	0.3186	0.0457	0.0000	0.1548	0.0457	0.1762	0.1470	0.0000	0.0000
Shanxi Han	太原	0.0501	0.2978	0.0379	0.0000	0.1816	0.0416	0.1816	0.1558	0.0188	0.0000
Shaanxi Han	西安	0.0396	0.3033	0.0296	0.0000	0.1524	0.0498	0.1756	0.1715	0.0098	0.0000
Shanghai Han	上海	0.0224	0.3114	0.0193	0.0000	0.1744	0.0290	0.2177	0.2026	0.0102	0.0000
Sichuan Han	什邡、绵竹	0.0332	0.3022	0.0044	0.0000	0.1556	0.0242	0.3341	0.1129	0.0000	0.0000
Hong Kong Han	香港	0.0083	0.3164	0.0111	0.0000	0.1487	0.0227	0.3384	0.1554	0.0030	0.0000
Yunan Han	云南	0.0559	0.2702	0.0331	0.0000	0.1992	0.0388	0.2555	0.0852	0.0000	0.0000
Zhejiang Han	杭州	0.0305	0.3072	0.0050	0.0050	0.1574	0.0253	0.2384	0.1506	0.0000	0.0000
Chongqing Han	重庆	0.0164	0.3277	0.0156	0.0000	0.1880	0.0275	0.3017	0.0960	0.0069	0.0000

3.2　中国 **26** 个汉族人群 **HLA-A** 基因座 **2** 种主成分分析结果的比较

由于各群体 HLA-A 基因频率资料中均存在空白基因, 因此表 1 所表示的 HLA-A 基因频率矩阵不存在"闭合效应"。故, 仅比较用原始基因频率的标准化相关系数阵和均值化协方差阵 2 种方法进行主成分分析的结果。

由两种主成分分析的特征根、贡献率和累积贡献率(表 2)可见, 用标准化相关阵所做的主成分分析的第 1 主成分的贡献率仅为 34.25%, 提供了 HLA-A 基因座遗传结构变异性的 34.25%的信息; 其第 2, 3 主成分的贡献率分别为 15.32% 和 13.96%; 前 3 个主成分的累积

贡献率仅为63.52%, 尚不能提供HLA-A 基因座遗传结构变异性的80% 以上的信息。表明该方法的降维效果不够理想。而用中心化协方差阵所做的主成分分析的第 1 主成分的贡献率为 54.36% , 提供了 HLA-A 基因座遗传结构变异性的 54.62% 的信息; 其第 2、3 主成分的贡献率分别为 22.72% 和 12.15% ; 前 3 个主成分的累积贡献率为 89.50%, 提供了 HLA-A 基因座遗传结构变异性近 90% 的信息。表明降维效果优于前者。

表 **2** 二种不同主成分分析方法的特征根及其贡献率

The eigenvalues and their cumulative proportion of different method

PC	标准化相关阵 standardized Correlation Matrix			中心化协方差阵 Centred Covariance Matrix		
	Eigenvalue	Proportion	Cumulative	Eigenvalue	Proportion	Cumulative
1	3.424612 13	0.3425	0.3425	0.007245 24	0.5463	0.5463
2	1.531581 43	0.1532	0.4956	0.003012 82	0.2272	017735
3	11396232 41	011396	01 6352	0.001 610 74	011215	018950

由两种方法的前 3 个主成分在各等位基因上的特征向量(表 3) 可见, 用标准化相关系数阵所做的主成分分析的第 1 主成分主要反映的是等位基因 A1、A2、A3、A9、A11、A36 对 HLA-A 基因座遗传结构的影响, 第 2 主成分主要反映的是等位基因 A2、A11、A19、A28、A36 对 HLA-A 基因座遗传结构的影响, 第 3 主成分主要反映的是等位基因 A9、A10、A19、A36 对 HLA-A 基因座遗传结构的影响; 而用中心化协方差阵所做的主成分分析的第 1 主成分主要反映的是等位基因 A1、A2、A3、A11 对 HLA-A 基因座遗传结构的影响, 第 2 主成分主要反映的是等位基因 A2、A9、A19 对 HLA-A 基因座遗传结构的影响, 第 3 主成分主要反映的是等位基因 A2、A3、A9、A19 对 HLA-A 基因座遗传结构的影响。

表 **3** 二种主成分分析方法的前 **3** 个主成分的特征向量

The eigenvectors of different method

Allele	标准化相关阵 standardized Correlation Matrix			中心化协方差阵 Centred Covariance Matrix		
	RPC1	RPC2	RPC3	CPC1	CPC2	CPC3
A1	01429250	- 01199196	- 01150154	- 01198134	01037817	- 01080503
A2	- 01377643	- 01457866	- 01136246	01126600	- 01444470	01671759
A3	01469472	- 01117326	- 01138598	- 01217780	01048893	- 01256568
A5	- 01086882	01177040	01007992	01000635	01001260	01003677
A9	01401494	01191830	01269258	- 01117841	01218560	- 01407634
A10	01159179	- 01155952	01661969	- 01040619	01061213	- 01071198
A11	- 01362693	01452700	01053599	01933800	01072805	- 01236647
A19	01044861	01348625	01441682	01041593	01861177	01492197
A28	01146563	- 01377700	01107154	- 01089593	01012735	01081804
A36	01321874	01418710	- 01468483	- 01002453	01004593	- 01015216

313 中国 **26** 个汉族人群 **HLA-A** 基因座 **2** 种主成分分析的 **Q** 型和 **R** 型散点图比较

图 1 和图 2 分别是用原始基因频率的标准化相关阵和中心化协方差阵所做的主成分分析的前 3 个主成分的三维 Q 型散点图。

比较图 1 和图 2 不难发现: 用标准化相关阵所做的主成分分析的前 3 个主成分的三维

Q 型散点图(图 1) 中, 南方民族与北方民族相互混杂, 特别是云南汉族混入北方汉族群体中, 难以给出合理的遗传学解释。由此表明标准化相关阵主成分分析对 HLA-A 基因座基因频率矩阵的降维效果较差, 其遗传结构不符合中华民族源与流的客观规律[1, 3, 10) 12]。而用中心化协方差阵所做的主成分分析的前 3 个主成分的三维 Q 型散点图(图 2) 以上海汉族群体为界, 将中国汉族群体区分为南方汉族和北方汉族两大群体。其分析结果符合中华民族源与流的客观规律[1) 3, 10) 12]。

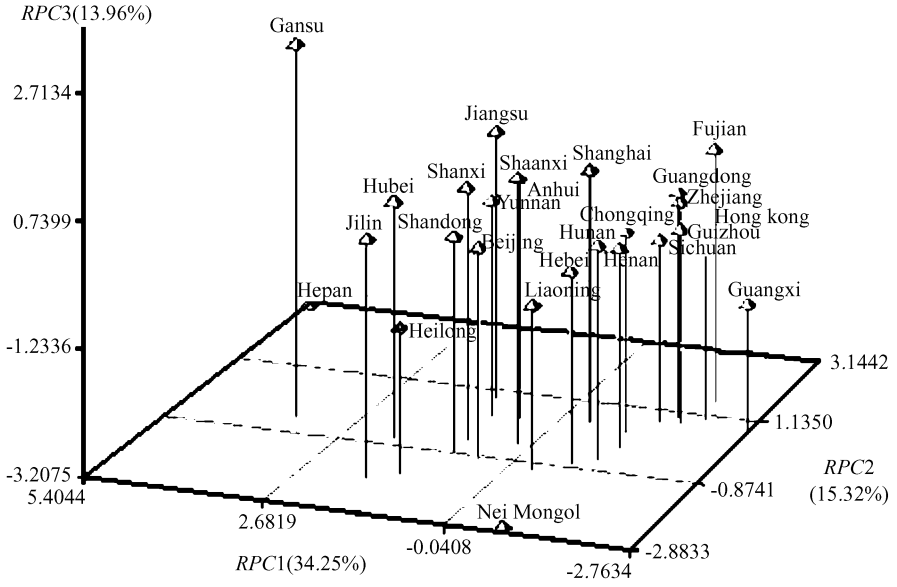

图 1　中国 26 个汉族群体 HLA-A 基因座标准化相关系数阵主成分分析的第 1、2、3 主成分散点图

Scallergram of 1st , 2rd and 3th scores base on standardized correlation matrix on HLA-A locus in 26 Chinese Han populations

Note: RPC denoted the principal component of standardized correlation Matrix

由图 1、图 2 还可看出, 两图所反映的中国北方和南方汉族群体遗传结构的差异相反: 图 1 显示各南方汉族群体间的空间距离较小, 各北方汉族群体间的空间距离较大; 图 2 却显示各北方汉族群体间的空间距离较小, 各南方汉族群体间的空间距离较大。结合中国人类群体遗传学的研究结论[3], 本研究认为图 2 符合中国汉族群体遗传学规律[1) 3, 10) 12]。

图 3 和图 4 分别是用标准化相关阵和均值化协方差阵所做的主成分分析的前 3 个主成分的三维 R 型因子载荷排序图。

由图 3 和图 4 可见, 在标准化相关阵主成分分析的前 3 个主成分三维 R 型因子载荷排序图中, 各等位基因的空间位置分散, HLA-A 基因座中各等位基因的关系不太明确; 在中心化协方差阵主成分分析的前 3 个主成分的三维 R 型因子载荷排序图中, HLA-A 基因座中各等位基因的关系明确, 呈现出明显的结构关系: 等位基因 A5、A10、A28、A36 聚集, 等位基因 A1、A3、A9 聚集, 等位基因 A2、A11、A19 各自独立, 自成一体。该图反映了中国汉族群体中因突变、婚配系统、基因流动、自然选择、随机漂变及地理隔离等众多因素对 HLA-A 基因座遗传座结构的综合作用特点。

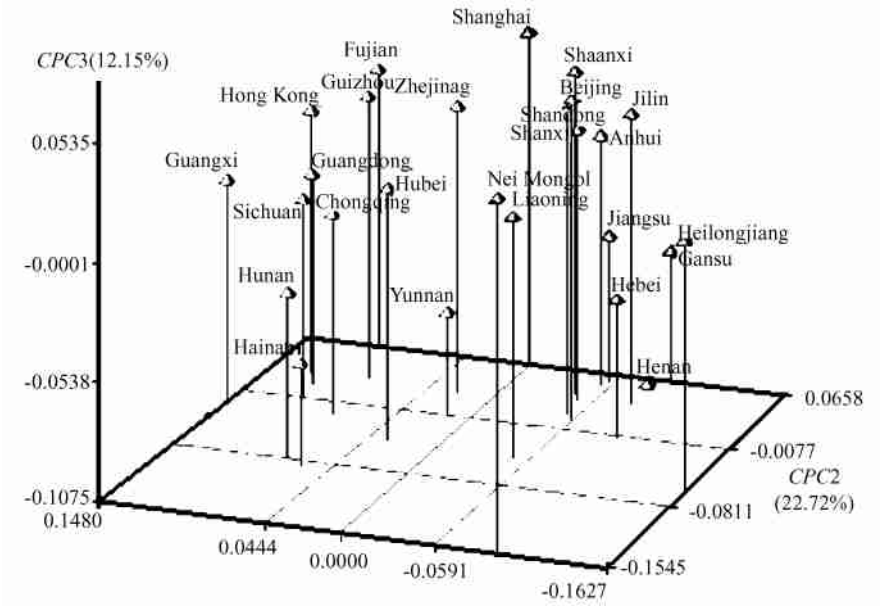

图 2 中国 26 个汉族群体 HLA-A 基因座中心化协方差阵主成分分析的第 1、2、3 主成分散点图

Scallergram of1st, 2rd and 3th scores base on centred covariance matrix on HLA-A locus in 26 Chinese Han populations

Note: *CPC* denoted the principal component of averaged covariance mtrix

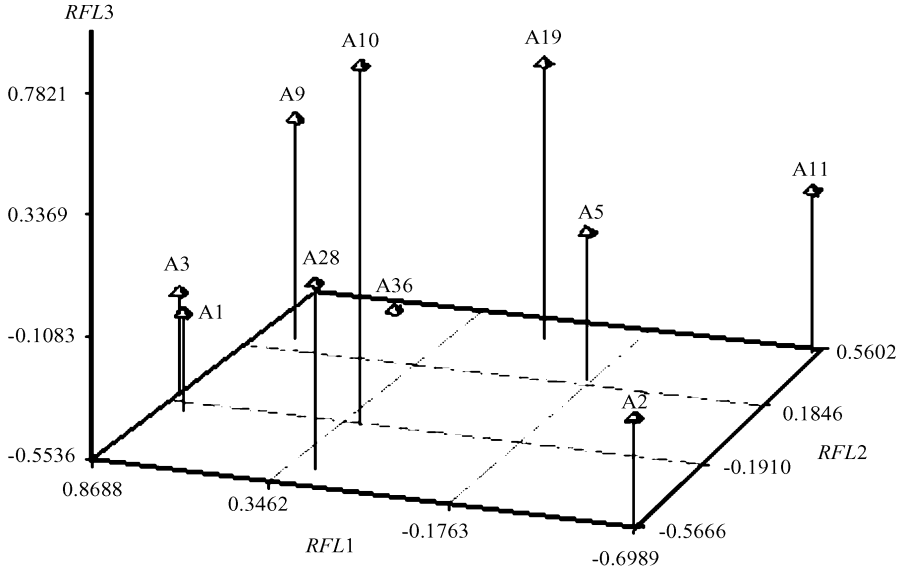

图 3 中国 26 个汉族群体 HLA-A 基因座标准化相关系数阵主成分分析的第 1、2、3 主成分因子载荷散点图

Scallergram of factor loading of 1st, 2rd and 3th base on standardized correlation matrix on HLA-A locus in 26 Chinese Han populations

Note: RFL denoted the factor loading of principal component based on standardized correlation matrix

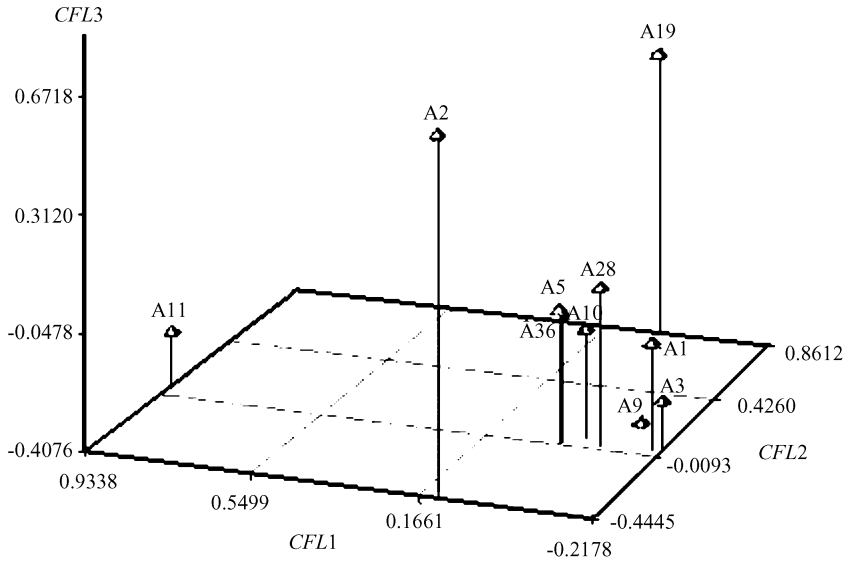

图 4　中国 26 个汉族群体 HLA-A 基因座中心化协方差阵主成分分析的第 1、2、3 主成分因子载荷散点图

Scallergram of factor loading of 1st, 2rd and 3th base on centred covariance matrix on HLA-A locus in 26 Chinese Han populations

Note: *CFL* denoted the factor loading of principal component based on averaged covariance matrix

4　讨　论

主成分分析是一种掌握主要矛盾的多元统计方法, 它能帮助研究者了解群体遗传结构的基本规律。因而, 在群体遗传学研究中广泛应用[2) 3]。然而, 由于等位基因多态性数据/闭合效应0的存在, 必然影响线性主成分分析, /闭合数据0使等位基因频率协方差矩阵具有明显的负偏性, 其每一特征向量的元素之和为 0, 即主分量也是/闭合数据0[13]。此时, 需进行非线性主成分分析[4]。另外, 在利用标准化变换消除基因频率量级差别的影响时, 标准化变换使各基因频率的方差变成 1, 消除了各基因变异程度上的差异, 不能正确反映原始基因频率矩阵所包含的全部信息, 其降维效果差。均值化或中心化变换后的协方差阵不仅包含了/方差信息量权0, 也包含了/相关信息量权0, 其降为效果好。本研究对中国 26 个汉族群体的 HLA-A 基因座的群体遗传结构进行中心化协方差阵主成分分析表明: 第 1 主成分的贡献率为 54163%, 即已提供了 HLA-A 基因座遗传结构变异性的 54163% 的信息, 前 3 个主成分的累积贡献率达 89150%, 提供了 HLA-A 基因座遗传结构变异性近 90% 的信息。因而, 其结果符合群体遗传学规律。

此外, 图 2 符合中国汉族群体遗传学规律[1) 3, 10) 12]。这是因为北方地区多为平原, 地理隔离少, 冬天黄河封冻可自由行走, 近 5 千年来由战乱灾荒所致的大规模人口迁移多发生在北方, 且迁移频繁。所以北方各地汉族人群在遗传结构上融和均匀, 各群体间的遗传差异理应较小。而长江以南地区由于南方蒙古人种的血缘越来越多, 北方蒙古人种的血缘越来越少, 且南方山多河流多, 地理隔离比北方多, 战乱与灾荒引起的大规模人口迁移也较北方少,

因此南方各群体之间的遗传结构理应较北方大。

参考文献:

[1] 杜若甫. 我国人类遗传学研究[J]. 生物学通报, 1997, 32(7): 9) 12.
[2] Barbujani G. Geographic patterns: how to identify them and why. Hum Biol, 2000, 72(1): 133) 53.
[3] 肖春杰, 杜若甫. 中国人类基因频率的主成分分析[J]. 中国科学, C 辑, 2000, 30(1): 434) 442.
[4] 薛付忠, 王洁贞, 郭亦寿, 等. 等位基因多态性群体遗传结构的多元非线性分析方法[J]. 遗传学报, 2004, 31, (20): 202) 211.
[5] Aitchison J. The statistical analysis of compositional data (with discussion)[J]. J Roy Stat Soc, Ser B, 1982, 44: 140) 177.
[6] Aitchison, J. The Statistical Analysis of Compositional Data, Chapman and Hall, 1986.
[7] 张尧庭, 成分数据多元分析引论[M]. 北京: 科学出版社, 2000, 8.
[8] 白雪梅, 赵松山. 对主成分分析综合评价方法若干问题的探讨[J]. 统计研究, 1995, 6: 47) 50.
[9] 叶双峰. 关于主成分分析做综合评价的改进[J]. 数理统计与管理, 2001, 20(2): 52) 55.
[10] 杜若甫, 肖春杰. 用 38 个基因座的基因频率计算中国人群间遗传距离[J]. 中国科学, C 辑, 1998, 28(1): 84) 89.
[11] 谭茜, 杜若甫. 中国 21 个人群的遗传拓扑学分析[J]. 人类学学报, 1993, 12(1): 80) 87.
[12] 杜若甫, 肖春杰. 从遗传学探讨中华民族的源与流[J]. 中国社会科学, 1997, 4: 139) 149.
[13] Chayes F, Trochimczyk J. An effect of closure on the structure of principal components[J]. J Math Geol, 1978, 10(4): 323) 333.

The Methodology of Principle Component Analysis Based on the Averaged Covariance Matrix for the Analysis of Human Populational Genetic Structures

XUE Fu-zhong[1], WANG Jie-zhen[1], GUO Yi-shou[2], HU Ping[1]

(1 *Dept. of Epidemiology and Biostatistics, School of Public Health, Shandong University, Jinan* 250012
2 *Dept. of Medical Genetics, Medical College, Shandong University, Jinan* 250012)

Abstract: **Objective**: To explore the applicability and rationale of principle component analysis based on the averaged covariance matrix for analyzing human populational genetic structure. **Methods**: Based on the structure of gene frequency matrix, we showed differences of eigenvalues, eigenvectors, and their effect in reducing the dimensionality between the standardized correlation matrix principle component analysis and the averaged covariance matrix principle component analysis. To validate and compare their use and rationale in human population genetics, we analyzed the genetic structure of HLA-A locus in 26 Chinese Han populations using both standardized correlation matrix principle component analysis and averaged covariance matrix principle component analysis methods. **Results**: The principle component of standardized correlation matrix does not represent the variance weight of gene frequency matrix. Instead it represents the correlation weight between the genes. The principle component of averaged covariance matrix not only reflects the variance weight of gene frequency matrix, but also identifies correlation weight between the genes in gene the matrix. From analyzing the genetic structure of HLA – A locus in 26 Chinese Han populations using the different two methods, we

discovered that the averaged covariance matrix principle component analysis is better than the standardized correlation matrix principle component analysis in reducing the dimensionality of gene frequency matrix. And using the principle method in reducing covariance matrix, the genetic structure of HLA-A locus in Chinese Han populations can be explained correctly. **Conclusion**: carry out the principle component analysis of human population genetic structure, one should calculate the PC using averaged covariance matrix rather than the standardized correlation matrix.

Key words: Human population; Genetic structure; Principle component analysis; Averaged covariance matrix; HLA-A