arXiv:2302.04596v1 [stat.ME] 9 Feb 2023

通过主成分分析或混合模型推断的群体结构评估

Jan van Waaij[1,3,†]、Song Li[1,†]、Genís Garcia-Erill[2], 、Anders Albrechtsen[2] 和 Carsten Wiuf[1,*]

[1]哥本哈根大学数学科学系，2100 哥本哈根，丹麦[2]哥本哈根大学生物学系，2100 哥本哈根，丹麦[3]现地址：丹麦技术大学健康技术系，2800 孔恩斯灵比，丹麦

†这些作者对本研究作出了同等贡献

*通讯作者：丹麦哥本哈根 2100，Universitetsparken 5，哥本哈根大学数学科学系。电子邮箱：wiuf@math.ku.dk

摘要

主成分分析（PCA）常用于遗传学，以推断和可视化群体结构以及群体之间的混合。PCA 的解读方式通常类似于推断混合比例，即假设个体属于若干可能群体之一，或由这些群体混合而成。我们提出一种新方法，用于评估 PCA（将其解读为由排名靠前的主成分张成的模型）的统计拟合度，并表明违反 PCA 假设会影响拟合效果。我们的方法利用选定的排名靠前的主成分预测基因型。通过评估残差（观测基因型与预测基因型之差）的协方差（及相关性），我们能够检测模型假设是否遭到违反。基于模拟和全基因组人类数据，我们表明，对拟合度的评估可用于指导数据解读，并找出无法由选定主成分充分表示的个体。我们的方法同样适用于其他类似模型，例如混合模型；在该模型中，数据的均值由线性矩阵分解表示。

关键词：主成分分析；残差残差；群体建模；古代DNA；统计　　度
群体建模；古代DNA；统计

引言

主成分分析（PCA）和基于模型的聚类方法是解析个体和群体祖先遗传史的常用方法。其中一种特定模型，即混合模型（Pritchard 等 2000），因其结构简单且在某些情况下易于解释而发挥了重要作用。PCA通常被认为是不依赖模型的，但正如Engelhardtand Stephens（2010）所指出的，这两种方法非常相似。对PCA分析结果的解释通常基于与混合模型相似的假设，即混合个体是代表未混合个体的特征向量的线性组合。这样，混合个体便位于主成分分析图中未混合个体之间。正如混合模型所示，许多人口史都可能导致相同的结果（Lawson 等2018a），并且许多人口史会违背混合模型的假设（Garcia-Erill 和 Albrechtsen，2020）。正如我们将要展示的，这对PCA同样成立，因为PCA具有相似的底层模型（Engelhardt and Stephens 2010）。

混合模型认为，每个个体的遗传物质由k 个不同祖先同质群体的贡献构成。然而，在真实数据分析中，这一假设常受到质疑，因为祖先群体结构可能比混合模型所设定的复杂得多。例如，k 个祖先群体本身可能具有异质性；由于存在许多贡献较小的群体，祖先群体的确切数量可能难以确定；或者，个体的遗传组成可能是连续迁移或近期回交所致，这同样违背了混合模型的假设。此外，混合模型假设个体之间互不相关，但实际情况未必如此。本文旨在评估主成分分析的拟合情况，并利用其与混合模型之间的特殊关系（Engelhardtand Stephens 2010）展开研究。具体而言，我们关注如何量化模型拟合，并在样本和个体两个层面评估模型的有效性。通过真实数据和模拟数据，我们表明，主成分分析的拟合情况会受到混合模型假设遭到违背的影响。

我们考虑个体的基因型数据 G，其中 n 个个体和 m 个单核苷酸多态性，且 $G_{si} \in \{0,1,2\}$ 表示个体i 在单核苷酸多态性 s上的参考等位基因数。通常假设 G_{si} 服从参数为 Π_{si}的二项分布，其中 Π_{si} 取决于祖先群体的数量 k、其混合比例以及祖先群体等位基因频率。对于 ADMIXTURE（Alexander and Lange 2011）等基于聚类的分析，k 表示聚类数；而在主成分分析中，它表示 $k-1$ 个主成分。下一节将详细介绍混合模型，并在 “ 材料与方法”一节中说明该模型与主成分分析的关系。

有几种方法旨在以某种意义上估计最佳k（Alexanderand Lange 2011；Evanno 等，2005；Pritchard 等，2000；Raj 等，2014；Wang 2019），但找到这样的k 并不意味着数据符合模型（Lawson 等，2018b；Janes 等，2017）。在统计学中，通常使用残差及其分布摘要来评估模型拟合（Box 等，2005）。观测值的残差定义为观测值与预测值（在某个模型下估计）之间的差。残差中呈现的视觉趋势（例如，群体间的差异）表明模型

拟合不佳，而残差绝对值较大则表明存在离群值（例如，由实验误差或亲缘关系导致）。如果模型正确，残差直方图预期呈以零为中心的单峰分布（Box 等，2005）。

在我们的语境中，Garcia-Erill 和 Albrechtsen (2020) 认为，残差相关矩阵中的趋势包含有关底层模型的信息，可用于直观地评估模型。为此，他们设计了一种方法，用于判断相关结构是否与所提出的模型一致，尤其是是否与所提出的同质祖先群体数量一致（Garcia-Erill 和 Albrechtsen，2020）。然而，即使模型设定正确，残差通常也彼此相关（Box 等 2005），因此，即使模型成立，也可能观察到趋势，导致模型评估有误。为校正这种相关性，他们开发了一种基于混合模型参数最大似然估计的留一法：当模型正确时，该方法会消除残差之间的相关性；而模型设定错误时则不会（Garcia-Erill 和 Albrechtsen，2020）。这一方法也可用于主成分分析，此时可使用概率主成分分析计算期望基因型（Meisner 等 2021）。不过，这种留一法的计算成本很高。

为解决计算方面的困难，我们采用不同的方法来研究相关结构。我们提出两种计算残差相关矩阵的方法。第一种方法是直接计算残差的经验相关矩阵。第二种方法可视为基于模型的估计相关矩阵。两者都易于计算。在温和的正则性假设下，如果模型正确且单核苷酸多态性数量较大，这两种度量会趋于一致。因此，当混合模型未被违反时，两者的差值预计接近于零。如果差值明显偏离零，则说明模型失配。

为了探究所提出方法的适用性，我们采用不同的方式进行主成分分析（PCA），以计算基因型的预测值（进而得到残差）。不过，我们也表明，该方法同样可用于估计的混合比例。具体而言，我们采用：1）一种不常见但非常实用的 PCA 方法（此处称为 PCA 1），基于未标准化的基因型（Cabreros 和 Storey 2019；Chen 和 Storey 2015年）；2）对均值中心化数据进行 PCA（PCA 2），参见 Patterson等人 (2006)；以及 3）对经过均值和方差标准化的数据进行 PCA（PCA 3）（Patterson 等人 2006）。这三种方法的计算速度都很快，而且不需要像 Garcia-Erill 和 Albrechtsen (2020) 那样分别估计祖先等位基因频率和群体比例。因此，计算残差的计算成本很低。此外，我们还表明，该方法也可应用于例如 ADMIXTURE 软件的输出（Alexander等人2009），以估计 Π_{si} 每个 s 和 i的值，并根据这些估计值计算残差。关于 PCA 的概述，参见 Jolliffe 和 Cadima (2016)。

我们证明了，所提出的方法在模拟数据和真实数据上效果良好，无论采用上述四种方式中的哪一种来计算预测值（及残差）。此外，我们还从数学上证明了：对于主成分分析 1 和主成分分析 2，在正确的混合模型下，只要单核苷酸多态性的数量足够大，这两种相关性度量就会一致。对于后者，还需要满足一些额外假设。在所提出的模型下，估计协方差（及相关系数）可视为群体结构的校正项。因此，从经验协方差中减去该项，便可得到一个在正确模型下基线为零、且不受群体结构影响的协方差估计量。我们自然会想到，在包含群体结构和亲缘关系的模型中或许也能采用类似方法；我们将在后续研究中对此进行探讨。

下一节将介绍模型、计算残差的统计方法，以及评估模型拟合的方法。此外，我们还将给出数学论述，说明该方法在理论上的表现。在“结果”一节中，我们将分别分析模拟数据和真实数据。最后，我们将进行讨论。数学证明汇集于附录中。

材料与方法

符号说明

对于一个 $\ell_1 \times \ell_2$ 矩阵 $A = (A_{ij})_{i,j}$，$A_{\star i}$ 表示 A 的第 i 列，$A_{i\star}$ 表示第 i 行，A^T 表示转置矩阵，而 $\operatorname{rank}(A)$ 表示秩。方阵 A 的弗罗贝尼乌斯范数为

$$\|A\|_F = \sqrt{\sum_{i=1}^{\ell}\sum_{j=1}^{\ell} A_{ij}^2}.$$

若方阵 A 满足 $A^2 = A$ 且 $A^T = A$，则它是正交投影。对称矩阵有 n 个实特征值（计重数），且特征向量可以选取为彼此正交。若矩阵是正（半）定的，则特征值为正（非负）。

对于随机变量/向量/矩阵 X，其期望记为 $\mathbb{E}[\mathrm{X}]$（假设其存在）。随机变量 X 的方差记为 $\operatorname{var}(\mathrm{X})$，两个随机变量 X、Y 之间的协方差记为 $\operatorname{cov}(\mathrm{X},\mathrm{Y})$（假设它们存在）。类似地，对于随机向量 $\mathrm{X} = (\mathrm{X}1,\ldots,\mathrm{X}_n)$，其协方差矩阵记为 $\operatorname{cov}(\mathrm{X})$。对于随机变量/向量/矩阵序列 X_m、$\mathrm{m} = 0, \ldots$，若当 $\mathrm{m} \to \infty$ 时 $\mathrm{X}_m \to \mathrm{X}_0$ 几乎必然成立（除一个概率为零的集合外，对所有实现均收敛），为方便起见，我们省略“几乎必然”，写作当 $\mathrm{m} \to \infty$ 时 $\mathrm{X}_m \to \mathrm{X}_0$。

主成分分析与混合模型

我们考虑一个模型，其中包含来自 n 个体的基因型观测值，以及 m 个双等位基因位点（单核苷酸多态性），其中假设 m 远大于 n，$m \geq n$。假设个体 i 在单核苷酸多态性位点 s 上的基因型 G_{si} 是一个二项随机变量

$$G_{si} \sim \operatorname{binomial}(2, \Pi_{si}).$$

用矩阵表示法，我们有 $G \sim$ 服从二项分布 $(2,\Pi)$，其期望为 $\mathbb{E}(G \mid \Pi) = 2\Pi$，其中 G 和 Π 是 $m \times n$ 维矩阵。给定 Π，我们假设 G 的各元素是相互独立的随机变量。

此外，我们假设矩阵 Π 具有如下形式 $\Pi = FQ$，其中 Q 是一个（可能不受约束的）$k \times n$ 阶矩阵，其秩为 $k \leq n$，且 F 是一个（可能不受约束的）$m \times k$ 阶矩阵，其秩也为 k（这意味着 Π 同样秩为 k，引理13）。逐元素表示为

$$\Pi_{si} = (FQ)_{si} = \sum_{k=1}^{k} F_{sk} Q_{ki}, \quad s = 1,\ldots,n, \quad i = 1,\ldots,m.$$

要使二项分布假设成立，必须要求 Π 的各元素介于零和一之间。

在文献中，这种模型通常以含有 k 个祖先群体的混合模型形式出现，参见例如 Pritchard 等人 (2000)； Garcia-Erill 和 Albrechtsen (2020)。适用于主成分分析的一般无约束设定也已有讨论（Cabreros 和 Storey 2019）。在混合模型中， Q 是一个祖先混合比例矩阵，其中个体 i的基因组中源自群体 j 的比例为 Q_{ji}。此外， F 是一个祖先 SNP 频率矩阵，

因此，群体 s 中单核苷酸多态性位点的参考等位基因频率为 j F_{sj}。在许多应用中，矩阵 Q 的各列之和为 1。

虽然我们倾向于将其解释为祖先群体比例和SNP频率，但我们的方法既不强制也不假定Q （混合比例）的各列之和为 1，而是允许这些列不受约束。这至少有两个优点。首先，提出的模型可能只包含主要祖先群体，而忽略较早出现或定义较不明确的群体。因此，祖先比例之和可能小于 1。其次，在拟合祖先群体数量少于真实模型的模型时，只需要求混合比例之和至多为 1。

残差

我们的目标是设计一种策略，以评估 Π 是否为两个矩阵的乘积。由于我们不知道真实的k，我们建议采用一个数值k' 作为祖先群体的数量，并在这一约束下估计模型参数。也就是说，我们假设模型形式如下：

$$G \sim \text{binomial}(2, \Pi_{k'}), \quad \Pi_{k'} = F_{k'} Q_{k'},$$

其中G 的每个元素都服从二项分布。$Q_{k'}$ 的维度为$k' \times n$，$F_{k'}$的维度为$m \times k'$，且秩$(Q_{k'}) =$秩$(F_{k'}) = k'$，因此秩$(\Pi_{k'}) = k'$。在全文中，我们使用索引k'表示施加的秩条件，并假设$k' \le k$，除非另有说明。后一假设仅用于保证某些论述在数学上的有效性，并非实际使用该方法所必需。

我们的方法建立在残差之上，即观测数据与预测数据之间的差异。为了定义残差，我们令 P: $\mathbb{R}^n \to \mathbb{R}^n$为投影到由（真实）Q 的 k 行张成的 k 维子空间上的正交投影，因此 $\mathrm{P} = \mathrm{Q}^T(\mathrm{QQ}^T)^{-1}\mathrm{Q}$，且 $\mathrm{QP} = \mathrm{Q}$。$\widehat{\ }$令 $\mathrm{P}_{k'}$ 为基于数据 G 得到的 P 的估计量，并假设 $\mathrm{P}_{k'}$ 是投影到 k′维子空间上的正交投影。在本节后文中，我们将说明如何根据 $\mathrm{Q}_{k'}$ 的估计量或 $\Pi_{k'}$的估计量得到 $\mathrm{P}_{k'}$ 的估计量。这些参数的估计量可以通过现有方法获得，例如基于最大似然分析的方法（Wang 2003；Alexander 等，2009；Garcia-Erill 和 Albrechtsen，2020）。此外，对于三种主成分分析方法，只需根据数据矩阵的奇异值分解（SVD）所得的特征向量，即可获得投影矩阵的估计量。

我们将残差的 $m \times n$ 矩阵定义为

$$R_{k'} = G - 2\widehat{\Pi} = G(I - \widehat{P}_{k'}),$$

其中，G 是观测数据，而$G\widehat{\ }P_{k'}$是预测值。后者也可以视为 2Π 的估计量，即G的期望值。这里对残差的定义与多线性回归模型中的残差定义一致：残差是观测数据（此处为G）与数据在回归变量张成的子空间上的投影（此处为$G\widehat{\ }P_{k'}$）之差。关键区别在于，在多线性回归模型中，回归变量是已知的，且$\widehat{\ }$不依赖于观测数据；而$P_{k'}$ 是根据数据估计得到的。

我们通过两种方式研究残差的相关矩阵，以评估模型拟合。首先，考虑条目为经验协方差$\widehat{\ }$矩阵 B 的矩阵，其中各项为

$$\begin{aligned}\widehat{B}_{ij} &= \frac{1}{m-1}\sum_{s=1}^{m}(R_{k',si} - \bar{R}_{k',i})(R_{k',sj} - \bar{R}_{k',j}) \\ &= \frac{1}{m-1}\sum_{s=1}^{m}(R_{k',si}R_{k',sj} - \bar{R}_{k',i}\bar{R}_{k',j}),\end{aligned}$$

其中

$$\bar{R}_{k'i} = \frac{1}{m}\sum_{s=1}^{m} R_{k',si},$$

以及相应的经验相关矩阵，其元素为

$$\widehat{b}_{ij} = \frac{\widehat{B}_{ij}}{\sqrt{\widehat{B}_{ii}\widehat{B}_{jj}}},$$

i、$j = 1, \ldots, n$。其次，我们考虑估计协方差矩阵

$$\widehat{C} = (I - \widehat{P}_{k'})\widehat{D}(I - \widehat{P}_{k'})$$

对应的估计相关矩阵，

$$\widehat{c}_{ij} = \frac{\widehat{C}_{ij}}{\sqrt{\widehat{C}_{ii}\widehat{C}_{jj}}},$$

i、$j = 1,\ldots,n$。其中，$\widehat{\ }D$是一个$n \times n$对角矩阵，其中包含每个个体的平均杂合度，

$$\widehat{D}_{ii} = \frac{1}{m}\sum_{s=1}^{m} G_{si}(2 - G_{si}), \quad i = 1, \ldots, n.$$

在合理的正则条件下，我们可以量化随着单核苷酸多态性数量增加，$\widehat{\ }B$ 和 $\widehat{\ }C$ 的行为。具体而言，我们假设F 的各行独立同分布，其分布为Dist(μ, Σ)，其中 μ 表示该分布的k维均值向量，Σ 表示$k \times k$协方差矩阵，即

$$F_{s\star} = (F_{s1}, \ldots, F_{sk}) \overset{\text{iid}}{\sim} \text{Dist}(\mu, \Sigma),$$

s $= 1,\ldots,$m。矩阵Q假设为非随机的，即固定不变。这些假设很常见，通常用于模拟遗传数据，例如参见Pickrell and Pritchard (2012); Cabreros and Storey (2019); Garcia-Erill and Albrechtsen (2020)。通常，dist(μ, Σ)取为k个独立均匀分布的乘积，此时 $\mu = 0.5(1,1,\ldots,1)$，且 Σ 是一个对角矩阵，其对角元素为1/12；不过也有其他取法，例如参见Balding and Nichols (1995); Conomos et al. (2016)。

设D为一个对角矩阵，其元素为

$$D_{ii} = 2\mathbb{E}[\Pi_{si}(1 - \Pi_{si})], \quad i = 1, \ldots, n. \tag{1}$$

$\widehat{\ }$由附录中的引理7 可知，当D 收敛于D 时，$m \to \infty$。此外，由于D_{ii} 是G_{si} 的方差（其服从二项分布），因此$\widehat{\ }D_{ii}$ 可被视为该方差的一个估计量。这些陈述的证明见附录。

定理1。设 $k' \le k$。在给定假设下，进一步假设当 $m \to \infty$时，$\widehat{\ }P_{k'} \to P_{k'}$，其中 $P_{k'}$为某个矩阵。则 $P_{k'}$ 是一个正交投影。此外，下式成立，

$$\begin{aligned}\widehat{B} &\to (I - P_{k'})(D + 4Q^T\Sigma Q)(I - P_{k'}), \\ \widehat{C} &\to (I - P_{k'})D(I - P_{k'}),\end{aligned}$$

当 $m \to \infty$时。因此，还有

$$\begin{aligned}\widehat{B} - \widehat{C} &\to 4(I - P_{k'})Q^T\Sigma Q(I - P_{k'}) \\ &= 4(P - P_{k'})Q^T\Sigma Q(P - P_{k'}),\end{aligned}$$

当 $m \to \infty$时。对于 $k' = k$，如果 $P_k = P$，则右端为零矩阵；而对于 $k' < k$，一般情况并非如此。

定理 2。假设 $k' = k$ 和 $P_k = P$。此外，假设与定理 *1* 中的情况相同，并且所有元素均等于 *1* 的向量位于 Q 的行所张成的空间中（例如，当每个个体的混合比例之和为 *1* 时，即属于这种情况）。那么，

$$\frac{\sum_{i=1}^{n}\sum_{j=1, i\neq j}^{n}\widehat{B}_{ij}}{\sum_{i=1}^{n}\widehat{B}_{ii}} \to -1, \quad as \quad m \to \infty. \tag{2}$$

此外，如果 Q 具有以下形式

$$Q = \begin{pmatrix} Q_1 & 0 & \cdots & 0 \\ 0 & Q_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q_r \end{pmatrix}$$

其中 Q_ℓ 的维度为 $k_\ell \times n_\ell$，$\sum_{\ell=1}^{r} k_\ell = k$ 且 $\sum_{\ell=1}^{r} n_\ell = n$，那么 (2) 对 n_ℓ 个体的每个组分都成立。如果 $Q_\ell = (1 \ldots 1)$，那么

$$\widehat{b}_{ij} \to -\frac{1}{n_\ell - 1}, \quad as \quad m \to \infty,$$

对于 ℓ 个组分中的所有个体 i，j，无论 $Q_\ell, \ell' \neq \ell$ 的形式如何。

定理 3。假设 $k' = k$ 且 $P_k = P$。此外，假设与定理 *1* 相同，并且 Q 具有如下形式

$$Q = \begin{pmatrix} Q_1 & Q_2 \\ 0 & Q_3 \end{pmatrix},$$

其中 $Q_1 = (1 \ldots 1)$ 的维度为 $1 \times n_1, n_1 \leq n$。那么，$\widehat{b}_{ij}$ 随着 $m \to \infty$ 收敛到一个大于或等于 $-\frac{1}{n}$ 的值，对于所有 $i, j, \ldots, n_1$。

$\widehat{}$ $\widehat{}$最后两个定理中的相同陈述，在分别将B 和b替换为 $\widehat{}C$ 和 $\widehat{}c$后仍然成立。

这三个定理为评估模型提供了方法。具体而言，定理1 可用于评估所提出的k'是否正确（或恰当）；定理2 和定理3 则可用于评估一组个体（例如当代群体）的数据是否源自同一个祖先群体，而不论其余个体的来源如何。我们在结果部分给出示例。

工作流程如算法1所示。我们使用主成分分析 1、主成分分析 2、主成分分析 3和ADMIXTURE 软件处理真实及模拟的基因型数据，并评估模型的拟合情况。

算法1 所提出方法的工作流程

1. 选择k'，2. 计算投影P的估计量 $\widehat{}P_{k'}$，3. 计算残差$R_{k'} = G(I - \widehat{}P_{k'})$，4. 计算相关系数，$\widehat{}b$ 和 $\widehat{}c$，5. 绘制 $\widehat{}b$ 及其与校正相关系数的差值，$\widehat{}b - \widehat{}c$，6. 目视评估模型的拟合情况。

估计$P_{k'}$

估计Q，F和 Π 在文献中受到广泛关注，例如采用最大似然（Wang 2003；Alexander 等 2009）、贝叶斯方法（Pritchard 等 2000）或主成分分析（En- gelhardt and Stephens 2010）。

我们讨论了获得估计量的不同方法

主成分分析或混合模型推断的人群结构评估$_{k'}$ 的主成分分析或混合模型推断的人群结构评估。

$\widehat{}$ $\widehat{}$使用估计量 $Q_{k'}$ 的$Q_{k'}$ 估计量 $P_{k'}$ 可以通过投影到由k' 个 $\widehat{}Q_{k'}$的行张成的子空间上获得，

$$\widehat{P}_{k'} = \widehat{Q}_{k'}^T (\widehat{Q}_{k'} \widehat{Q}_{k'}^T)^{-1} \widehat{Q}_{k'},$$

假设秩($\widehat{}Q_{k'}$) $= k'$，以确保计算有效。

我们应用此方法，利用 ADMIXTURE 软件的输出估计投影矩阵。

使用估计量 $\widehat{\Pi}_{k'}$ 的 $\Pi_{k'}$ 设 $\widetilde{\Pi}_{k'}$ 为从 $\widehat{\Pi}_{k'}$ 中选出的k' 个线性无关行（共m行）。那么，$P_{k'}$ 的估计量 $\widehat{}P_{k'}$ 为

$$\widehat{P}_{k'} = \widetilde{\Pi}_{k'}^T (\widetilde{\Pi}_{k'} \widetilde{\Pi}_{k'}^T)^{-1} \widetilde{\Pi}_{k'},$$

假设秩($\widehat{\Pi}_{k'}$) $= k'$，则计算有效。或者，也可以使用 Gram–Schmidt 方法，此时向量按构造即为标准正交，且 $\widehat{}P_{k'} = \widetilde{\Pi}_{k'}^T \widetilde{\Pi}_{k'}$。只要秩($\widehat{\Pi}_{k'}$) $= k'$，估计量 $\widehat{}P_{k'}$ 就与所选的k' 行无关。

使用主成分*1*我们采用一种主成分分析方法，最初由 Chen 和 Storey (2015年提出，用于估计由 Q的各行张成的空间。我们遵循 Cabreros 和 Storey (2019) 中所述的步骤。

$\widehat{}$令H 为对称矩阵

$$\widehat{H} = \frac{1}{m} G^T G - \widehat{D}.$$

$\widehat{}$由于H是对称矩阵，因此所有特征值均为实数，且该矩阵可对角化。此外，$\widehat{}H$是 $\frac{1}{m} G^T G$ 的方差调整版本，见（†）。令$u1, \ldots, u_{k'}$为属于k'个最大特征值（按重数计）的$k' \leq k$个正交特征向量。定义$n \times k'$矩阵$U_{k'} = (u1, \ldots, u_{k'})$ 以及$n \times n$正交投影矩阵

$$\widehat{P}_{k'} = U_{k'} (U_{k'}^T U_{k'})^{-1} U_{k'}^T = U_{k'} U_{k'}^T$$

投影到由向量$u1$、…、$u_{k'}$的张成空间所给出的子空间上。

在H这可转换精确定义收敛性矩阵 $H \widehat{=} 4Q^T(\Sigma + \mu\mu^T)Q$。那么，$H$是对称且半正定的，因为 Σ和 $\mu\mu^T$均为半正定矩阵。因此，H具有非负特征值。此外，根据附录中的引理 8，$\widehat{}H$收敛于 H，当 $m \to \infty$。

定理 4。设 $k' \leq k$。令 $\lambda_1 \geq \ldots \geq \lambda_n \geq 0$ 为 H 的特征值，其对应的正交特征向量为 $v1$、…、v_n。特别地，$\lambda_{k+1} = \ldots = \lambda_n = 0$，因为 Q 的秩为 k。令 $P_{k'}$ 为投影到由 $v1$、…、$v_{k'}$张成空间上的正交投影，即

$$P_{k'} = V_{k'} (V_{k'}^T V_{k'})^{-1} V_{k'}^T = V_{k'} V_{k'}^T,$$

其中 $V_{k'} = (v1, \ldots, v_{k'})$。

假设 $k' = n$ 或 $\lambda_{k'} > \lambda_{k'+1}$，称为特征值条件。那么，当 $m \to \infty$ 时，$\widehat{}P_{k'} \to P_{k'}$。如果 $k' = k$ 满足特征值条件，那么 $P_k = P$，也就是说，P_k 是 Q 的行向量张成空间上的正交投影。特别地，当且仅当 $\Sigma + \mu\mu^T$ 是正定矩阵时，$k' = k$ 满足特征值条件。若 Σ 是正定矩阵，则后者成立。

对于 $k' = k$，最终可以找到 Q 的正确行空间，但无法找到 Q 本身。如果 $k' < k$，则可以找到该行空间的一个子空间，对应于 k' 个最大特征值。由于数据未进行均值中心化，我们舍弃第一个主成分，并使用后续的 $k' - 1$ 个特征向量和特征值。

使用主成分分析 **2**（均值中心化数据）在混合模型中估计 Π 的一种常见方法是基于均值中心化数据或均值和方差归一化数据进行主成分分析（Pritchard等人 2000；Engelhardt and Stephens 2010；Patterson 等人 2006）。

令 $G_1 = G - \frac{1}{n}nGE = G(I - \frac{1}{n}E)$ 为按 SNP 均值中心化的基因型，其中 E 是一个所有元素均为 1 的 $n \times n$ 矩阵。按照 Cabreros and Storey (2019) 中的阐述和符号说明，令 $G_1 = U\Delta V^T$ 为 G_1，的奇异值分解，其中 ΔV^T 由 G_1，的行主成分组成，并按奇异值排序。定义

$$S_{k'} = \begin{pmatrix} U^T_{1:(k'-1)} \\ e \end{pmatrix},$$

其中 $e = (11 \ldots 1)$ 是一个所有元素均为 1 的向量，而 $U^T_{1:(k'-1)}$ 包含 U^T 的前 $k'-1$ 行。那么，投影的估计量为

$$\widehat{P}_{k'} = S^T_{k'}(S_{k'}S^T_{k'})^{-1}S_{k'}.$$

G_1 的奇异值分解中的奇异值平方，与下列矩阵的特征值相同

$$\widehat{H}_1 = \frac{1}{m}G_1^T G_1 = \frac{1}{m}\left(I - \frac{1}{n}E\right)G^T G\left(I - \frac{1}{n}E\right)$$

(Jolliffe2002)。我们有

$$\begin{aligned}\mathbb{E}[\widehat{H}_1] &= \frac{1}{m}\left(I - \frac{1}{n}E\right)\mathbb{E}[G^T G]\left(I - \frac{1}{n}E\right) \\ &= \left(I - \frac{1}{n}E\right)(D + 4Q^T(\Sigma + \mu\mu^T)Q)\left(I - \frac{1}{n}E\right). \quad (3)\end{aligned}$$

令H_1 表示（3）右侧的表达式。

定理 5。设 $\lambda_1 \geq \ldots \geq \lambda_n$ 为 H_1，的特征值，对应的正交特征向量为 v_1，...，v_n。特别地，$v_n = e$ 且 $\lambda_n = 0$。如果 D 的所有对角元素均为正，则 $\lambda_{n-1} > 0$。

设 $k' \leq n$，令 $P_{k'}$ 为投影到 v_1，...，$v_{k'-1}$，e 的张成空间上的正交投影，即

$$P_{k'} = V_{k'}(V^T_{k'}V_{k'})^{-1}V^T_{k'},$$

其中 $V_{k'} = (v_1, \ldots, v_{k'-1}, e)$。如果 $k' = n$ 或 $\lambda_{k'} > \lambda_{k'+1}$，则当 $m \to \infty$ 时，$\widehat{P}_{k'} \to P_{k'}$。

不能保证当 $k' = k$ 时，我们有 $P_k = P$，也不能保证 $\widehat{B}$ 与 $\widehat{C}$ 之间的差在 m 较大时趋于零。然而，在某些额外条件下确实如此，而且在许多实际情形中似乎也是如此，参见结果部分。

定理 6。假设 $D = dI$，其中某个 $d > 0$。此外，假设向量 e 属于 Q 的行空间（例如，当每个个体的混合比例之和为 *1* 时便是这种情况）。那么，$\lambda_k = \ldots = \lambda_{n-1} = d$，且 $\lambda_n = 0$。

如果 $\Sigma + \mu\mu^T$ 是正定的，则 $\lambda_{k+1} > \lambda_k$ 以及 $P_k = P$，其中 P_k 如定理 *4*中所述。作为推论，对于定理 *1*中的 $k' = k$，$\widehat{B} - \widehat{C} \to 0$ 当 $m \to \infty$时。

使用主成分分析 ***3***（均值和方差标准化数据）设$G_2 = W^{-1}G_1$ 为经过均值和方差标准化的单核苷酸多态性基因型，其中W 是一个$m' \times m'$ 对角矩阵，其s阶元素为单核苷酸多态性s基因型的观测标准差。移除所有未观测到变异的单核苷酸多态性，因此单核苷酸多态性的数量可能小于原始数量$m' \leq m$。按照主成分分析 2 中的相同步骤，设$G_2 = U\Delta V^T$ 为G_2，的奇异值分解，其中 ΔV^T 由G_2，按行排列的主成分构成，并按奇异值排序。定义

$$S_{k'} = \begin{pmatrix} V^T_{1:(k'-1)} \\ e \end{pmatrix},$$

其中，$e = (11 \ldots 1)$，并且 $V_{T1:(k'-1)}$ 包含 V^T 的前 $k'-1$ 行。于是，投影的估计量为 $\widehat{P}_{k'} = S^T_{k'}(S_{k'}S^T_{k'})^{-1}S_{k'}$。

据我们所知，目前没有与定理1类似的理论依据来支持这一过程，但根据我们的模拟结果，它在许多实际情形中表现良好。

基因型数据模拟

我们采用不同的抽样策略，从不同的人口统计情景中模拟基因型数据。我们特意选择了不同的抽样策略，以考验该方法。首先，我们进行了简单的模拟，既说明了模型拟合问题，也展示了有限数量的个体和大量单核苷酸多态性数据所导致的残差相关性的理论和实际特性。模拟概况见表1。

在前两个情景中，我们针对每个祖先群体，独立地从均匀分布中模拟祖先等位基因频率：对于每个位点$F_{si} \sim \text{Unif}(0,1)$，$s = 1,\ldots,m$ 以及每个祖先群体$i = 1,\ldots,k$。在情景1中，我们模拟了来自三个群体的未混合个体，每个群体的抽样个体数量相同或不同。在情景2中，我们模拟了两个祖先群体，以及一个祖源各有一半来自这两个祖先群体之一的混合群体。

在情景 3 中，我们设定 $F_{si} \sim \text{Unif}(0.01,0.99)$，并以一种模拟长而狭窄岛屿上各群体间连续基因流随空间距离递减的方式，模拟了空间混合。我们首先模拟了长岛中央的一个群体。随后，从岛屿两侧开始，根据参数为 $F_{st} = 0.001$ 的 Balding-Nichols 分布，使用 R 软件包“bnpsd”（Ochoa 和 Storey，2019）递归模拟新的群体。这样，岛上每一对相邻群体的 F_{st} 均为 0.001。关于模拟的更多细节以及示意图，请参见Garcia-Erill 和 Albrechtsen (2020) 中的图 2。

在情景4中，我们首先从形状参数为 0.03 的对称贝塔分布中，为祖先群体模拟等位基因频率 $F_{si} \sim \text{Beta}(0.3,0.3)$，由此得到富含罕见变异的等位基因频谱，模拟人类等位基因频谱。随后，我们从二分树 (((群体1:0.1,popGhost:0.2):0.05,群体2:0.3):0.1,群体3:0.5) 中抽取等位基因频率，其中群体1和 popGhost 是姐妹群体，群体3是外群。使用 Balding-Nichols 分布和树的 F_{st} 枝长（见图 5），我们在四个叶节点中抽取等位基因频率。随后，我们创建了一个混合群体，其祖源有 30% 来自 popGhost，70% 来自群体2。我们从每个群体中为 50 名个体各抽取了 1000 万个基因型，但幽灵群体未纳入分析；之后移除了样本次要等位基因频率低于 0.05 的位点，最终共得到 694,285 个位点。

在情景5中，我们模拟了一个祖先群体，其等位基因频率来自均匀分布 $F_{si} \sim \text{Unif}(0.05,0.95)$；随后，我们使用 'bnpsd' 从该群体中抽样得到两个子群体的等位基因频率，这些频率服从 Balding Nicholds 分布，且 $F_{st} = 0$=0.3。接着，我们根据一个家系谱创建了近期杂交群体，其中除一位奠基者外，其他所有奠基者的祖源都来自第一个群体。家系谱中的世代数决定了混合比例和混合年代，其中 F1

个体分别从两个群体各有一位未混合亲本，而回交个体的一位亲本未混合，另一位是 F1。双重回交个体的一位亲本未混合，另一位是回交个体。依此类推，直至四重回交个体，其一位亲本未混合，另一位是三重回交个体。请注意，对于近期杂交群体，每个位点上等位基因对的祖源不再相互独立，这违反了混合模型的假设。

结果

情景1

在这一组初始设置中，我们仅使用主成分分析1来展示该方法。我们模拟了来自$k = 3$ 个祖先群体的未混合个体

$$Q = \begin{pmatrix} 1_{n_1} & 0 & 0 \\ 0 & 1_{n_2} & 0 \\ 0 & 0 & 1_{n_3} \end{pmatrix},$$

其中，1_{n_i}是一个所有元素均为 1 的行向量，且$n_1 + n_2 + n_3 = n$。我们按照上一节所述，为$n = 60$ 个体模拟了基因型，样本量分别为n_1,n_2 和n_3,。图1（A）展示了$k' = 2$时的残差相关系数，并绘制了相应的主要主成分。对于主成分分析1方法，由于数据未进行均值中心化，第一个主成分与群体结构无关，因此我们使用以下$k' - 1$ 个主成分。

假设群体数只有两个时，$k' = 2$，我们注意到各群体样本内的经验相关系数总体上较为一致，但校正相关系数通常不为零且符号各异，这表明模型失配。相反，当假设群体数为正确值$k' = 3$时，经验相关系数与理论值 $-\ ^1_{n_i-1}$吻合良好，这与定理2一致（见表2）。在所有样本中，校正相关系数均呈现出围绕零值的相当均一的模式。这很好地表明模型拟合良好，并且基于主成分2和3绘制的主成分分析图能很好地反映数据。

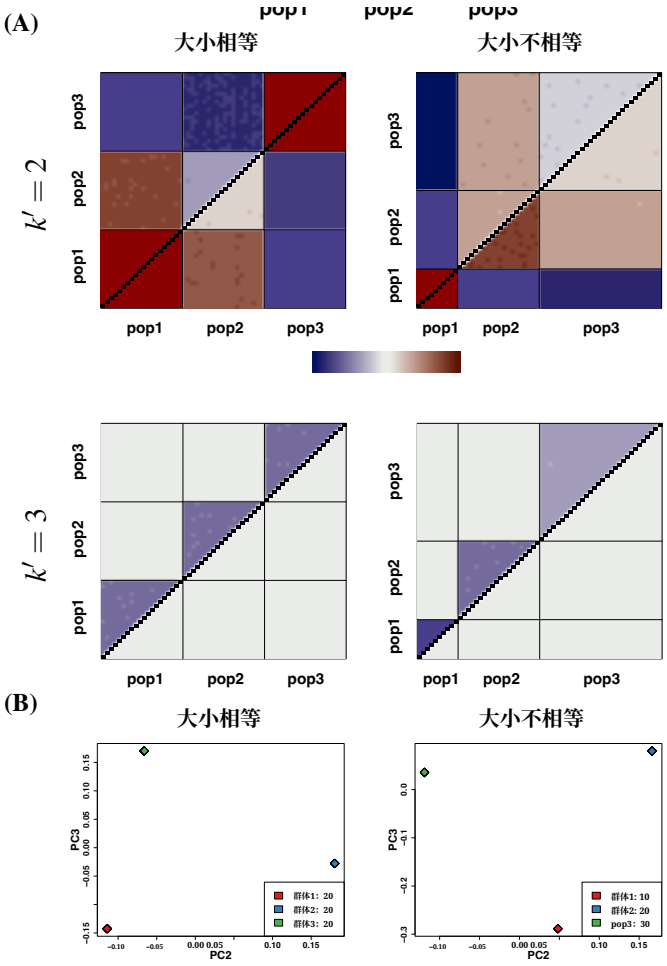


图1 情景1模拟的结果。(A) 图中上三角部分^显示经验相关系数b，下三角部分显示校正相关系数^$b - \hat{}c$。(B) 主要主成分（$k' = 3$）使三个样本明显分离（每个样本内的所有数据点几乎完全相同）。

情景2

在此设置中，我们还纳入了混合个体。我们从两个祖先群体以及由这两个群体混合而成的个体中模拟了样本。随后，我们将全部三种主成分分析方法和 ADMIXTURE 软件应用于这些数据。具体而言，我们选择

$$Q = \begin{pmatrix} 1_{n_1} & \frac{1}{2}1_{n_2} & 0 \\ 0 & \frac{1}{2}1_{n_2} & 1_{n_3} \end{pmatrix},$$

具有 $k = 2$ 个真实祖先群体，并选择 $(n_1,n_2,n_3) = (20,20,20)$ 或 $(n_1,n_2,n_3) = (10,20,30)$，详情请参阅上一节。我们使用 $k' = 1$、2、3 对数据进行了分析，得到了图 2 和 3以及表 2所示的相关结构。这两种标准方法

主成分分析 2 和主成分分析 3 的结果几乎完全相同，因此图中只展示主成分分析 2。主成分分析 2 和主成分分析 3 都使用排名靠前的主成分，而主成分分析 1 则忽略第一个主成分，因此图2(b) 和 3/b) 中的坐标轴标注有所不同。对于 $k' = 1$，不使用任何主成分，预测的标准化基因型简单地设为 0。四种方法的结果一致，尤其是在正确的 k' (= 2) 下；而对于错误的 $k' = 1$，各方法之间的差异较小。这种差异在主成分分析 1 和 ADMIXTURE 中最为明显。我们注意到，$\hat{b}$ 在每个群体样本中的平均相关系数符合定理 1 （见表 2）。

对于 $k' = 2$，所有样本的校正相关系数都呈现出围绕零值的相当均匀的分布模式，与情景1相同，这表明模型拟合良好。然而，与情景1不同，经验相关系数的偏差并非样本量的简单函数（见表2）。

在这种情况下，与其他所有考察的情况类似，我们没有发现这四种方法之间存在任何显著差异。因此，我们只展示主成分分析1的结果，因为我们有理论依据支持这些结果。

场景 3

我们模拟了 $n = 500$ 个体的基因型，在个体之间 $m = 88{,}082$ 个位点上存在连续的基因流，因此不存在真正的 k。我们假设 $k' = 2{,}3$ 对数据进行了分析，见图 4。图中，个体按照估计的祖先群体比例排序，因此经验相关系数和校正相关系数呈现出颜色波浪模式，见图 4(A)。正如预期，对于 $k' = 3$，校正相关系数比 $k' = 2$更接近零，不过与零的偏差仍然很大。因此，对于 k'的任一取值，我们都没有发现支持该模型的证据。这与主要主成分图一致，图中显示

表 1 模拟概述。

情景	k	n	m	描述	FFF_{is}[a]
1	3	20,20,20	$500K$	未混合	均匀（0,1）
1	3	10,20,30	$500K$	未混合	均匀（0,1）
2	2	20,20,20	$500K$	混合群体	Unif(0,1)
2	2	10,20,30	$500K$	混合群体	Unif(0,1)
3		500	$100K$[b]	空间的 $F_{st} = 0.001$ 相邻群体之间	Unif(0.01,0.99)
4	4	50,50,50,50,0[c]	$10M$[d]	幽灵混合	Beta(0.3,0.3)
5	2	20,20,50	$500K$	近期杂交群体	均匀分布(0.05,0.95)

[a] 祖先等位基因频率，$i = 1$、…、k [b] 应用 MAF> 5% 筛选后，剩余 88,082 个。[c] 未提供幽灵群体的参考样本。
[d] 应用 MAF> 5% 筛选后，剩余 694,285 个。

表2 使用主成分分析1 得到的每个群体中b 和$b - c$ 的均值（标准差）

情景1	kk'	n		pop1	pop2	pop3
	3	(20,20,20)	$\hat{b}$ [a]	-0.0526 (0.0015)	-0.0526 (0.0016)	-0.0526 (0.0016)
				-0.0526	-0.0526	-0.0526
			$\hat{b} - \hat{c}$	0e-04 (0.0015)	0e-04 (0.0016)	0e-04 (0.0016)
		(10,20,30)	$\hat{b}$	-0.1111（0.0011）	-0.0526（0.0016）	-0.0345 (0.0016)
				-0.1111	-0.0526	-0.0345
			$\hat{b} - \hat{c}$	0e-04 (0.0012)	0e-04 (0.0016)	0e-04 (0.0016)
情景2	kk'	n		**pop1**	**混合群体**	**pop3**
	2	(20,20,20)	$\hat{b}$	-0.0419 (0.0015)	-0.0192 (0.0015)	-0.0420 (0.0015)
				-0.0420	-0.0193	-0.0420
			$\hat{b} - \hat{c}$	0e-04 (0.0015)	0e-04 (0.0015)	0e-04 (0.0015)
		(10,20,30)	$\hat{b}$	-0.0701 (0.0018)	-0.0228 (0.0014)	-0.0304 (0.0016)
				-0.0701	-0.0229	-0.0304
			$\hat{b} - \hat{c}$	0e-04 (0.0017)	0e-04 (0.0014)	0e-04 (0.0016)
情景4	kk'	n		**pop1**	**pop2**	**pop3**	**pop4**
	3	(50,50,50,50)	$\hat{b}$	-0.0190 (0.0015)	0.0027 (0.0015)	-0.0204 (0.0017)	0.0122 (0.0013)
			$\hat{b} - \hat{c}$	0.0009 (0.0015)	0.0147 (0.0015)	0e-04 (0.0017)	0.0208 (0.0013)
	4		$\hat{b}$	-0.0204 (0.0015)	-0.0204 (0.0015)	-0.0204 (0.0017)	-0.0204（0.0014）
			$\hat{b} - \hat{c}$	0e-04 (0.0015)	0e-04 (0.0015)	0e-04 (0.0017)	0e-04 (0.0013)

[a] 每种情况下的第二行$\hat{b}$显示的是根据定理1极限得到的理论值。

数据连续变化，并未被划分为两个或三个聚类，见图 4(B)。

情景4

本例基于图5中的树，其中包括一个未采样的（所谓的）幽灵群体 popGhost。popGhost 是群体1的姐妹群体。

我们模拟了$n = 200$ 个个体的基因型：分别来自群体1、群体2和群体3的150个未混合样本；以及50个混合比例为0.3的样本来自 popGhost 的祖源以及来自群体2的 0.7 祖源（作为 pop4），如上一节所述。由于群体之间存在漂变，因此也存在遗传差异，正确的 $k = 4$ (群体1、群体2、群体3、popGhost）。我们的方法能够捕捉到这一点，并清楚地表明 $k' = 3$是错误的，因为校正相关系数明显偏离零。相比之下，对于 $k' = 4$，校正相关系数几乎为零（图6）。

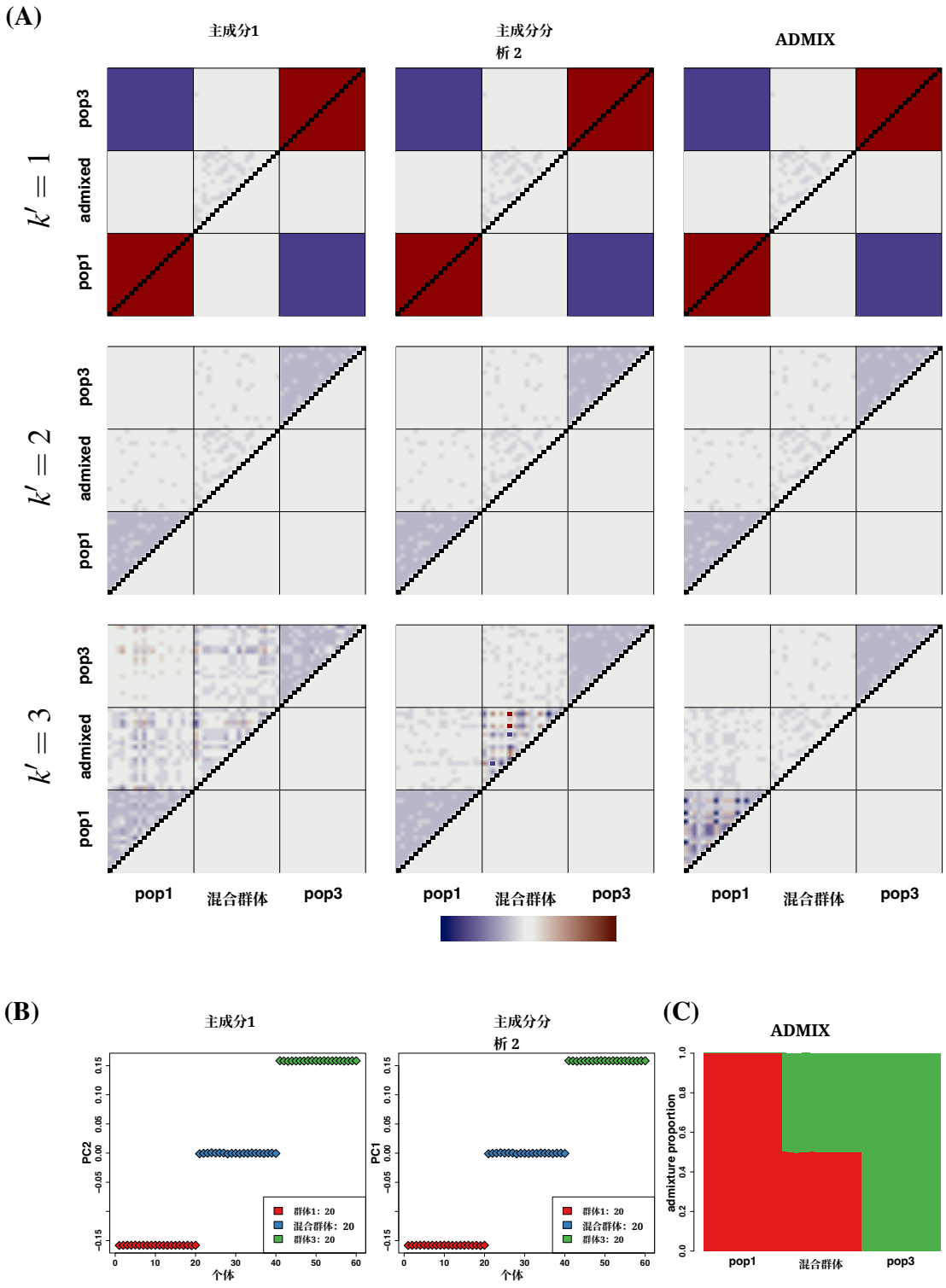


图 2 样本量相等的模拟情景2结果。（A）对于主成分1、主成分分析 2 和 ADMIXTURE，图中的左上三角显示经验相关性 $\hat{}b$，右下三角显示差值 $\hat{}b - \hat{}c$，样本量为 $(n_1,n_2,n_3) = (20,20,20)$。（B）对于 $k' = 2$ 的基于主成分的方法，其主要主成分（此时只有一个主成分）。每个样本中的个体使用相同的颜色。（C）ADMIXTURE 情况下的混合比例估计值。

情景5

在上一个示例中，我们模拟了两个群体（源自一个共同的祖先群体），并通过回交创建了混合群体，具体如上一节所述。因此，该模型确实不满足混合模型的假设，因为参考等位基因的数量并不服从二项分布，而是取决于特定的回交以及亲本群体的频率。

我们模拟了$n = 90$ 个体在$m = 500{,}000$个位点上的基因型。

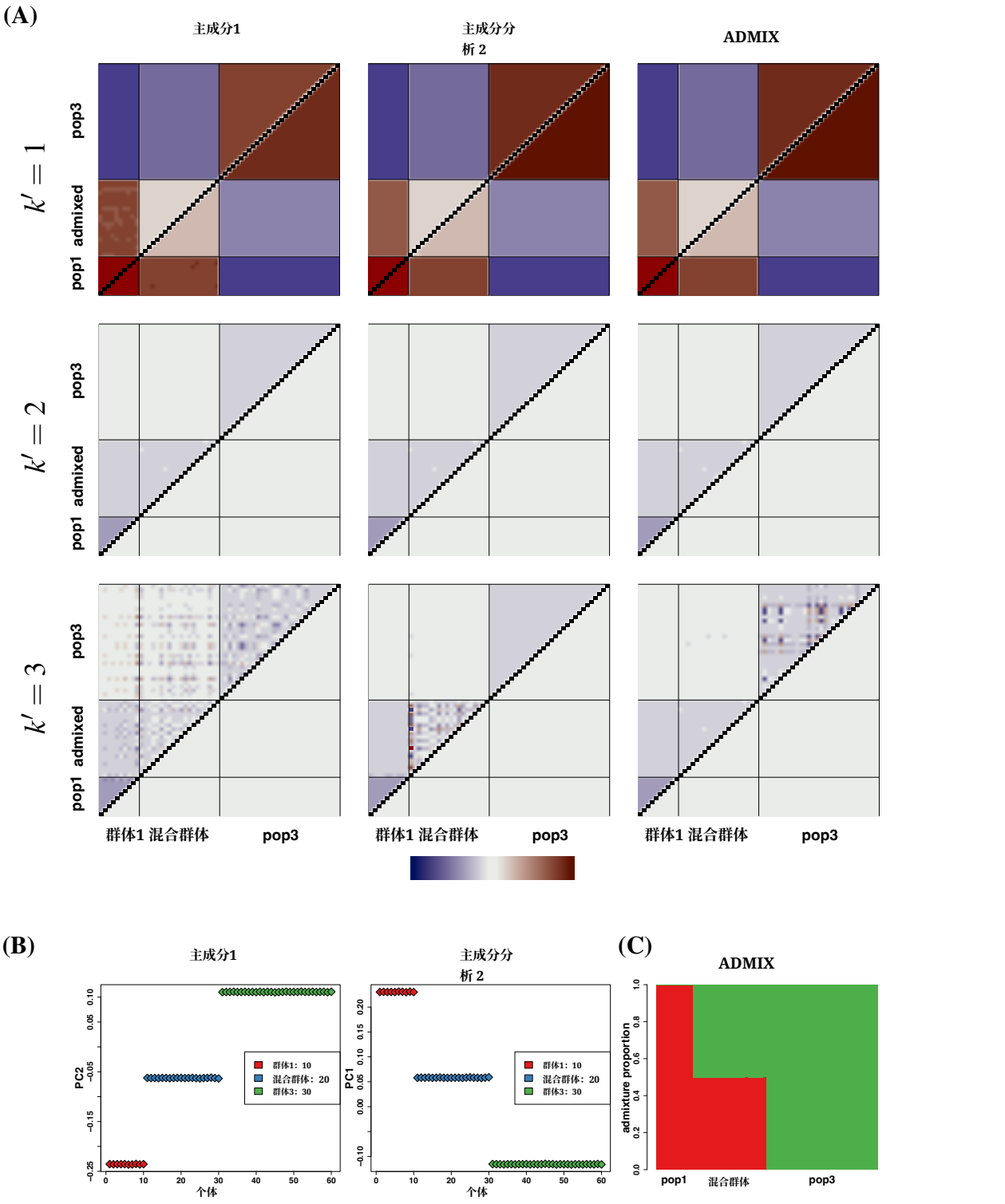


图 3 样本量不相等的模拟情景2结果。(A) 对于主成分1、主成分分析 2 和 ADMIXTURE，图中的左上三角显示经验相关性 $\hat{}b$，右下三角显示差值 $\hat{}b - \hat{}c$，样本量为 $(n_1,n_2,n_3) = (20,20,20)$。(B) 对于 $k' = 2$ 的情况，基于主成分分析的方法所得到的主要主成分（此时只有一个主成分）。每个样本中的个体均以相同颜色表示。(C) ADMIXTURE 情况下估计的混合比例。

每个亲本群体中有20个同质个体，此外，每个不同的近期混合类别中有10个不同个体。随后，我们使用$k' = 2$分析了数据，发现校正相关系数持续偏离零，尤其是对于其中一个亲本群体（图7）。因此，我们可以断定，混合模型无法提供合理的拟合。

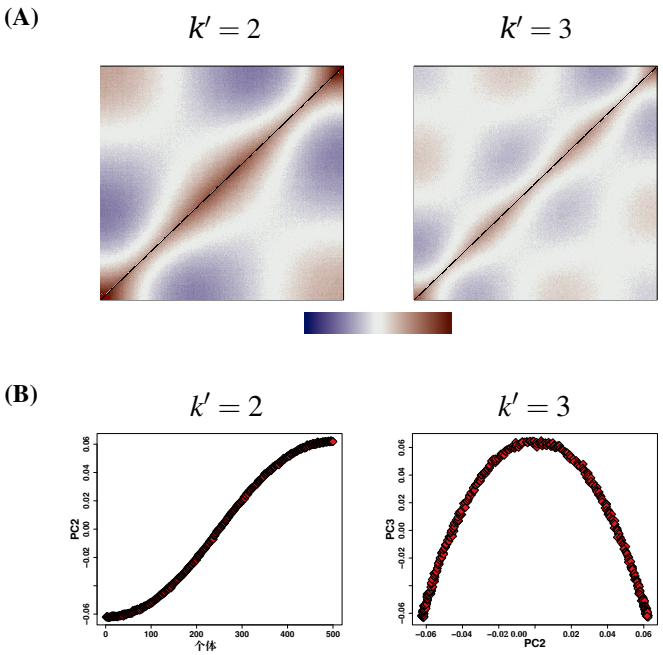


图 4 模拟情景3的结果。(A) 图中的上三角部分 ˆ 显示经验相关性b，下三角部分显示差异 $\hat{b} - \hat{c}$。(B) 主要主成分（在$k' = 2$的情况下只有一个）。

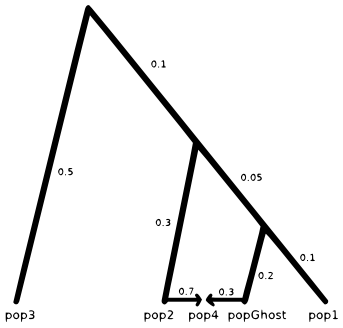


图5 用于模拟情景4中群体等位基因频率的树示意图，包括5个群体：群体1、群体2、群体3、pop4和popGhost。群体pop4由群体2和群体popGhost混合而成，后者没有采集任何个体的样本，因此是一个幽灵群体。分支上的数值表示以F_{ST}为单位的漂变。两条混合边上的数值表示分别来自各群体的混合比例。

真实数据

我们分析了来自千人基因组计划的一组全基因组测序数据（Auton 等，2015年），另见Garcia-Erill 和 Al- brechtsen (2020)，该研究也使用了相同的数据集。数据集包含来自五个不同祖源群体的数据：来自尼日利亚伊巴丹的约鲁巴族群体（YRI）、具有非洲血统的美国西南部居民（ASW）、具有北欧和西欧血统的犹他州居民（CEU）、来自加利福尼亚州洛杉矶的墨西哥血统群体（MXL），以及来自中国北京的汉族群体（CHB）；样本量分别为108、61、99、63和103，总计$n = 434$。我们仅保留人类起源SNP面板中包含的位点（Lazaridis 等2014），共计$m = 406$个SNP在经过0.05的次要等位基因频率过滤后保留下来。

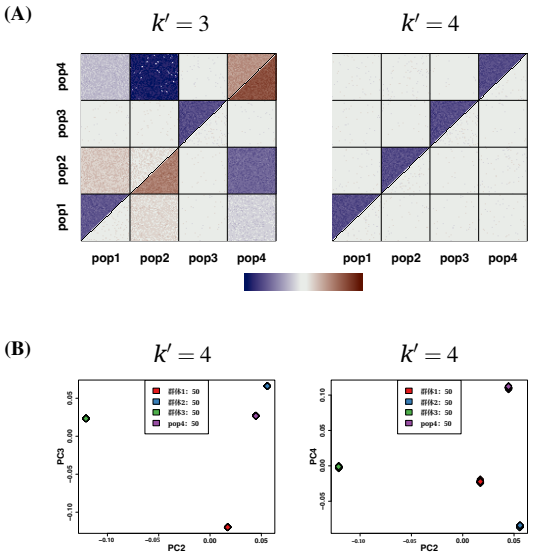


图6 模拟情景4的结果。（A）图中上三角显示经验相关性b，下三角显示差值$\hat{b} - \hat{c}$。（B）$k' = 4$个主要主成分导致4个样本明显分离（每个样本中的所有数据点几乎完全相同）。

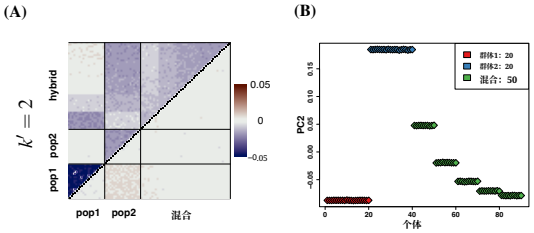


图7 模拟情景5（近期混合）的结果。（A）图中上三角显示经验相关性b，下三角显示差值$\hat{b} - \hat{c}$。（B）$k' = 2$个主要主成分。

我们使用$k' = 3$，4 进行了数据分析。对于 $k' = 3$，图 8 显示，无法解释 MXL、CEU 和 CHB 之间的关系，这表明 MXL 无法很好地解释为两者的混合体。对于 $k' = 4$，每个群体内部校正相关系数的颜色深浅几乎没有差异，表明美洲原住民群体有所贡献。图 8（D）进一步佐证了这一点，该图显示了使用 ADMIXTURE 软件估计的四个祖先群体的比例。

讨论

我们提出了一种新方法，可根据残差相关矩阵的结构评估主成分分析和混合模型的模型拟合情况。我们已证明，该方法在模拟数据和真实数据上均表现良好；评估过程中采用了文献中常用的一系列不同主成分分析方法，并使用 ADMIXTURE 软件估计模型参数。通过直观地评估残差相关结构，可以检测模型失配以及建模假设是否遭到违反。

通过直观比较两个残差相关系数矩阵来评估模型拟合。我们的方法在理论和实践上的优势体现在三个方面。首先，该方法计算简单、速度快。计算两个残差相关矩阵及其差值的成本很低。其次，

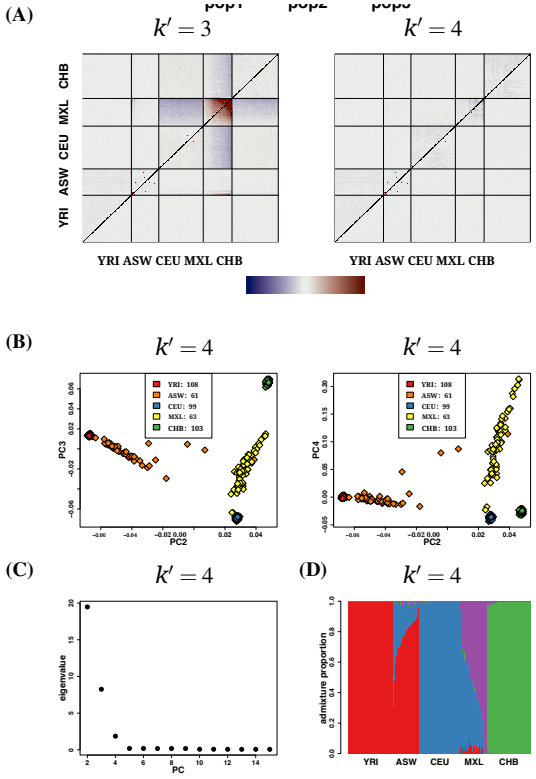


图8 真实人类数据（来自千人基因组计划）的残差相关系数、推断的群体结构和混合比例。（A）图中的上三角显示经验相关系数 $\hat{}b$，下三角显示差异 $\hat{}b-\hat{}c$。（B）PCA 1 的三个主要主成分$k'=4$。（C）移除第一个主成分的特征值后，其余主成分对应的特征值在第四个主成分之后接近于0。（D）使用 ADMIXTURE 估计的混合比例。

我们的方法为基于主成分分析和聚类方法（如 ADMIXTURE）的模型拟合提供了统一的方法。具体而言，它提供了一种简单的方式，用于评估所选主成分数量是否足以描述数据结构。通过绘制主成分之间的关系来评估其充分性，可能会造成虚假的信心。相比之下，我们的方法通过绘制两个残差相关系数矩阵之差，揭示模型失配。第三，在某些情况下，该方法具有理论保证。在无法提供理论有效性的情况下，这些保证还得到了模拟的进一步支持。最后，我们的方法或许可以调整后用于新一代测序数据，无需先估计基因型，而是直接使用基因型似然。

数据可用性

本研究所使用的数据集均可公开获取，包括模拟数据和真实数据。用于分析和模拟数据的 R代码相关信息可在 https://github.com/Ginwaitthreebody/ evalPCA获取。所使用的千人基因组计划数据的变异位点调用结果可在以下地址公开获取：ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/ 20130502/。

致谢

作者获得丹麦独立研究基金（资助编号：8021-00360B）和哥本哈根大学通过 Data+ initiative 提供的支持。SL 感谢中国国家留学基金管理委员会提供的经费支持。GGE 和 AA 获得丹麦独立研究基金的支持（资助编号分别为：8049-00098B 和 DFF-0135-00211B）。

附录A

在给定的分布假设下，我们首先分别给出$G_{s\star}$ 和 $\Pi_{s\star}$的期望和协方差矩阵，

$$\mathbb{E}[\Pi_{s\star}]=\mu^T Q, \qquad \operatorname{cov}(\Pi_{s\star})=Q^T\Sigma Q,$$
$$\mathbb{E}[G_{s\star}]=2\,\mathbb{E}[F_{s\star}Q]=2\mu^T Q,$$
$$\operatorname{cov}(G_{s\star})=\mathbb{E}[\operatorname{cov}(G_{s\star}\mid\Pi)]+4\operatorname{cov}(\Pi_{s\star})=D+4Q^T\Sigma Q,$$

对于$s=1,\ldots,m$，其中

$$D=2\,\mathbb{E}[\operatorname{diag}(\Pi_{s1}(1-\Pi_{s1}),\ldots,\Pi_{sn}(1-\Pi_{sn}))],$$

and

$$\mathbb{E}[\Pi_{si}(1-\Pi_{si})]=\mu^T Q_{\star i}-(\mu^T Q_{\star i})^2-(Q^T\Sigma Q)_{ii}.$$

根据构造，G 的无条件列$G_{s\star}$、$s=1$、…、m是独立随机向量。

以上意味着

$$\frac{1}{m}\mathbb{E}[G^TG]=D+4Q^T(\Sigma+\mu\mu^T)Q. \tag{4}$$

辅助结果见附录 B。

引理 7. 估计量 $\hat{D}$ 是 D 的无偏估计量，即 $\mathbb{E}[\hat{D}]=D$。此外，当 $m\to\infty$ 时，有 $\hat{D}\to D$。

证明。 在给定 Π_{si}的条件下，利用二项性可得 $\mathbb{E}[G_{si}(2-G_{si})\mid\Pi_{si}]=2\Pi_{si}(1-\Pi_{si})$，第一个结论由此得出。对于收敛性，注意到 $G_{si}(2-G_{si})$，$s=1$，…，m，在无条件下构成一列具有有限方差的独立同分布随机变量，因此由强大数定律可得收敛性结论（Jacod 和 Protter 2004）。 □

引理 8. 估计量 $\hat{H}=\frac{1}{m}G^TG-\hat{D}$ 是 $H=4Q^T(\Sigma+\mu\mu^T)Q$ 的无偏估计量，即 $\mathbb{E}[\hat{H}]=H$。此外，当 $m\to\infty$ 时，有 $\hat{H}\to 4Q^T(\Sigma+\mu^T\mu)Q$，且

$$\mathbb{E}\left[\|\hat{H}-4Q^T(\Sigma+\mu\mu^T)Q\|_F^2\right]\le\frac{16n^2}{m}.$$

证明。由 (4) 和引理 7 可得无偏性。考虑 $\frac{1}{m}G^TG$ 的第 (i, j) 个元素，即 $\frac{1}{m}\sum_{s=1}^m G_{si}G_{sj}$。序列 $G_{si}G_{sj}$, $s=1,\ldots,m$ 是具有有限方差的独立同分布序列，因此根据强大数定律（Jacod 和 Protter 2004），当 $m\to\infty$ 时，$\frac{1}{m}G^TG$ 收敛于 $\mathbb{E}[G^TG]$。结合引理 7 可得，当 $m\to\infty$ 时，$\hat{H}$ 收敛于 H。

剩下需要证明该不等式。定义

$$A_{s,ij}=\begin{cases}G_{si}G_{sj}-4(Q^T(\Sigma+\mu\mu^T)Q)_{ij} & \text{if } i\neq j,\\ 2G_{si}(G_{si}-1)-4(Q^T(\Sigma+\mu\mu^T)Q)_{ii} & \text{if } i=j.\end{cases}$$

于是，

$$(\hat{H}_{ij}-\mathbb{E}[\hat{H}_{ij}])^2=\left(\frac{1}{m}\sum_{s=1}^m A_{s,ij}\right)^2=\frac{1}{m^2}\sum_{s=1}^m\sum_{t=1}^m A_{s,ij}A_{t,ij},$$
$$\|\hat{H}-\mathbb{E}[\hat{H}]\|_F^2=\frac{1}{m^2}\sum_{i=1}^n\sum_{j=1}^n\sum_{s=1}^m\sum_{t=1}^m A_{s,ij}A_{t,ij}.$$

利用 $\mathbb{E}[A_{s,ij}]=0$，$A_{s,ij}$ 与 $A_{t,ij}$ 在 $s\neq t$ 时的独立性，以及 $|A_{s,ij}|\le 4$，我们有

$$\mathbb{E}[\|\hat{H}-\mathbb{E}[\hat{H}]\|_F^2]=\frac{1}{m^2}\sum_{i=1}^n\sum_{j=1}^n\sum_{s=1}^m\mathbb{E}[A_{s,ij}^2]\le\frac{1}{m^2}16mn^2=\frac{16n^2}{m},$$

由此即可证明该论断。 □

收敛结果也见于Chen 和 Storey （2015年，定理 2）。第二部分给出了 $\widehat{H}$ 在 L^2-范数下的收敛速率。收敛取决于m足够大，而非n足够大，并要求m 至少按n的平方增长。

定理1的证明。由于$P_{k'}$ 被假设为正交投影，即 $\widehat{P}_{k'}^2 = \widehat{P}_{k'}$ 且 $\widehat{P}_{k'}^T = \widehat{P}_{k'}$，那么其极限也为正交投影，$P_{k'}^2 = P_{k'}$ 且$P_{k'}^T = P_{k'}$。

考虑经验协方差 $\widehat{B}$。定义变量$T_{k'} = G(I - P_{k'})$，其中将 $\widehat{P}_{k'}$ 替换为$P_{k'}$，以及经验协方差

$$\begin{aligned}\widetilde{B}_{ij} &= \frac{1}{m-1}\sum_{s=1}^{m}(T_{k',si}T_{k',sj} - \overline{T}_{k',i}\overline{T}_{k',j}) \\ &= \frac{1}{m-1}\sum_{s=1}^{m}T_{k',si}T_{k',sj} - \frac{m}{m-1}\overline{T}_{k',i}\overline{T}_{k',j},\end{aligned}$$

其定义方式与 $\widehat{B}_{ij}$类似，其中 $\overline{T}_{k',i} = \frac{1}{m}\sum_{s=1}^{m}T_{k',si}$。序列 $T_{k',si}$，$s = 1,2,\ldots$，以及 $T_{k',sj}$，$s = 1,2,\ldots$，都是独立同分布随机变量，这是由关于 G 的分布假设得出的。此外，由于 $P_{k'}$是正交投影，因此 $\|I - P_{k'}\|_F^2 \le n$ 有界（引理12）。因此，$T_{k',si}$ 关于 s、i 一致有界，界为 $2\sqrt{n} \le 2n$。

利用有界性、独立性和强大数定律（Jacodand Protter 2004），

$$\widetilde{B}_{ij} \to \mathbb{E}[T_{k',1i}T_{k',1j}] - \mathbb{E}[T_{k',1i}]\,\mathbb{E}[T_{k',1j}] = \operatorname{cov}(T_{k',1i}, T_{k',1j}), \quad (5)$$

对于 $m \to \infty$，且 $\operatorname{cov}(T_{k',1i}, T_{k',1j}) = (I - P_{k'})(D + 4Q^T\Sigma Q)(I - P_{k'})$。后一等式由（4）得出。

考虑 $R_{\widehat{}} = \quad _{k'} = \quad - \quad _{k'} \quad _{k'} = \quad _{k'} -$ $G(I - \widehat{P})\,G(IP) + G(P_{k'} - \widehat{P})\,T + G(PP_{k'})$。因此，

$$\begin{aligned}|\overline{R}_{k',i} - \overline{T}_{k',i}| &\le \frac{1}{m}\sum_{s=1}^{m}\sum_{i'=1}^{n}\sum_{j'=1}^{n}2|(P_{k'} - \widehat{P}_{k'})_{i'j'}| \\ &= 2\sum_{i'=1}^{n}\sum_{j'=1}^{n}|(P_{k'} - \widehat{P}_{k'})_{i'j'}| \to 0,\end{aligned}$$

根据定理的假设，当 $m \to \infty$ 时。由此可知$R_{k',i}$ 收敛于 $\mathbb{E}[T_{k'}, 1i]$，当 $m \to \infty$时。此外，

$$\begin{aligned}&\frac{1}{m-1}\sum_{s=1}^{m}R_{k',si}R_{k',sj} - \frac{1}{m-1}\sum_{s=1}^{m}T_{k',si}T_{k',sj} \\ &= \frac{1}{m-1}\sum_{s=1}^{m}(T_{si} + (G(P_{k'} - \widehat{P}_{k'}))_{si})(T_{sj} + (G(P_{k'} - \widehat{P}_{k'}))_{sj}) \\ &\quad - \frac{1}{m-1}\sum_{s=1}^{m}T_{k',si}T_{k',sj} \\ &= \frac{1}{m-1}\sum_{s=1}^{m}T_{si}(G(P_{k'} - \widehat{P}_{k'}))_{sj} + \frac{1}{m-1}\sum_{s=1}^{m}(G(P_{k'} - \widehat{P}_{k'}))_{si}T_{sj} \\ &\quad + \frac{1}{m-1}\sum_{s=1}^{m}(G(P_{k'} - \widehat{P}_{k'}))_{si}(G(P_{k'} - \widehat{P}_{k'}))_{sj}.\end{aligned}$$

绝对 值的绝对值在上面最后一行中有界 由

$$\frac{4nm}{m-1}\sum_{j'=1}^{n}|(P_{k'} - \widehat{P}_{k'})_{j'j}|,$$

第二项也同理。第三项有界 由

$$\frac{4m}{m-1}\sum_{i'=1}^{n}\sum_{j'=1}^{n}|(P_{k'} - \widehat{P}_{k'})_{i'i}||(P_{k'} - \widehat{P}_{k'})_{j'j}|.$$

当 $m \to \infty$ 时，这三项均趋于零，因此由（5）可得 $\widehat{B}_{ij} \to$ 协方差$(T_{k'}, 1i, T_{k'}, 1j)$ 当 $m \to \infty$ 时。

$\widehat{}$估计协方差C 的结果由 $\widehat{D}$ 的收敛性以及定理的假设得出。其余部分由 $\widehat{B}$ 和 $\widehat{C}$ 的收敛性得出。注意，$QP = Q$，因此第二个等式成立。定理的最后一个结论可直接得出。

定理 2 的证明。考虑 $T_k = G(I - P_k) = G(1 - P)$，其中 $P = Q^T(QQ^T)^{-1}Q$ 是投影到 Q 的行空间上的投影。于是，T_k包含将 G 的 m 行对 Q 的 k 行进行多元回归后得到的残差（Box 等人，2005 年）。由于 e 属于 Q 的行空间，因此对于每个 $s = 1,\ldots,m$，残差之和为零：$\sum_{i=1}^{n}T_{k,si} = 0$（假设 e 属于行空间等价于回归模型中包含截距）（Box 等人，2005 年）。对于 $s = 1,\ldots,m$，有：

$$\begin{aligned}0 = \operatorname{var}\left(\sum_{i=1}^{n}T_{k,si}\right) &= \sum_{i=1}^{n}\operatorname{var}(T_{k,si}) + \sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}\operatorname{cov}(T_{k,si}, T_{k,sj}) \\ &= \sum_{i=1}^{n}\operatorname{var}(T_{k,1i}) + \sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}\operatorname{cov}(T_{k,1i}, T_{k,1j}),\end{aligned}$$

因为$T_{k,si}$ 的分布与s无关。由定理4的证明可知，$\widehat{B}$ 收敛于 $\operatorname{cov}(T_{k,1\star})$，当$m \to \infty$时。因此，

$$\sum_{i=1}^{n}\widehat{B}_{ii} + \sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}\widehat{B}_{ij} \to 0, \quad \text{as} \quad m \to \infty,$$

通过重新排列即可得到所需结果。

若Q 具有给定形式，则在多元回归中，各区室之间的残差相互独立，因为投影为

$$P = \begin{pmatrix} P_1 & 0 & \cdots & 0 \\ 0 & P_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_r \end{pmatrix},$$

其中$P_\ell = Q_\ell^T(Q_\ell Q_\ell^T)^{-1}Q_\ell$ 的维数为$n_\ell \times n_\ell$。由此可知，上述计算对每个区室都成立。最后，若$Q_\ell = (1\ldots1)$，则随机向量$T_{k,1\star}$ 的分布具有可交换性，从而得到

$$\begin{aligned}0 = \operatorname{var}\left(\sum_{i=1}^{n_\ell}T_{k,1i}\right) &= \sum_{i=1}^{n_\ell}\operatorname{var}(T_{k,1i}) + \sum_{i=1}^{n_\ell}\sum_{j=1,i\neq j}^{n_\ell}\operatorname{cov}(T_{k,1i}, T_{k,1j}) \\ &= n_\ell\operatorname{var}(T_{k,11}) + n_\ell(n_\ell - 1)\operatorname{cov}(T_{k,11}, T_{k,12})\end{aligned}$$

假设第 ℓ个区室中的个体编号为1到n_ℓ。$\widehat{}$重新排列各项，并将b^{ij} 代入$T_{k,i\star}$ 的矩，即可得到所需结果。

定理的证明 3。考虑$T_k = G(I - P_k) = G(1 - P)$，其中$P = Q^T(QQ^T)^{-1}Q$ 是到Q的行空间的投影。若$Q_1 = (1\ldots1)$，则随机变量$T_{k,11},\ldots,T_{k,1n_1}$的分布具有可交换性，从而得到

$$\begin{aligned}0 \le \operatorname{var}\left(\sum_{i=1}^{n_1}T_{k,1i}\right) &= \sum_{i=1}^{n_1}\operatorname{var}(T_{k,1i}) + \sum_{i=1}^{n_1}\sum_{j=1,i\neq j}^{n_1}\operatorname{cov}(T_{k,1i}, T_{k,1j}) \\ &= n_1\operatorname{var}(T_{k,11}) + n_1(n_1 - 1)\operatorname{cov}(T_{k,11}, T_{k,12}).\end{aligned}$$

$\widehat{}$重新排列各项，并将 b^{ij} 代入 T_k的矩，$i\star$ 即可得到所需结果。

定理 4 的证明。定理的收敛性结论是附录 B 中定理 9 的特例。取 $A_m = \widehat{H}$ (其取决于

单核苷酸多态性数量 m 以及具体实现），A = H 和定理中的 k = k′（k 在定理 9 中用作通用索引）。那么，$E_k E_k^T = P_{k'}$ 和 $F_{m,k} F_{m,k}^T = \hat{P}_{k'}$，因此定理 4 的结论成立。弗罗贝尼乌斯范数意义下的收敛等价于逐点收敛（因为 n 固定）$\hat{P}_{k'} \to P_{k'}$，这是由 $m \to \infty$ 时的定义得到的。

如果 $\Sigma + \mu\mu^T$ 是正定的，那么它的秩为 k。根据假设，秩 $(Q) = k$，由引理 13 可得，秩$(H) = k$。因此，H 有 k 个正特征值，并且 $\lambda_{k+1} = 0$，特征值条件成立。反过来，假设特征值条件成立。根据定义，秩$(H) \le k$。由于根据假设 $\lambda_k > \lambda_{k+1} \ge 0$，因此秩$(H) \ge k$，由此可得秩$(H) = k$。因此，$\Sigma + \mu\mu^T$ 的秩为 k；所以它是正定的。

如果 $k' = k = n$，那么 $P_k = V_k V_k^T = I$，且 $P = I$（因为 $k = n$），并且 $P_k = P$。因此，假设 $k' = k < n$。由于特征值条件成立，由上文可知，秩$(Q^T(\Sigma + \mu\mu^T)) = k$，而引理 13 表明 H 和 Q 的行空间相同。同理，我们有 $H = V_k \operatorname{diag}(\lambda_1,\ldots,\lambda_k) V^T k$，引理 13 表明 H 和 V_k^T 的行空间相同。这意味着 Q 和 V_k^T 的行空间相同。因此，$P_k = Q^T(QQ^T)^{-1}Q = P$，结论成立。

定理 5 的证明。 显然，e 是 H_1 的特征向量，特征值为 0。如果 D 的所有元素均为正，则其为正定矩阵，并且 $D + 4Q^T(\Sigma + \mu\mu^T)Q$ 也为正定矩阵，因此其秩为 n。根据引理 13 可知，H_1 的秩为 $n-1$，因此 $\lambda_{n-1} > 0$。

类似于附录 B 中引理 8 的证明，可以证明 $\mathbb{E}[\hat{H}_1] = H_1$ 和 $\hat{H}_1 \to H_1$ 分别为 $m \to \infty$，其中 H_1 表示（3）的右端。定理的其余部分类似于定理 4的证明。

定理 6 的证明。 注意，e 是 H_1 的特征向量，特征值为 0。考虑 v，它是 H_1, 的一个特征向量，并且与 e 正交，特征值为 λ。则下面两个方程等价:

$$\left(I - \frac{1}{n}E\right)(D + 4Q^T(\Sigma + \mu\mu^T)Q)\left(I - \frac{1}{n}E\right)v = \lambda v,$$
$$4\left(I - \frac{1}{n}E\right)Q^T(\Sigma + \mu\mu^T)Q\left(I - \frac{1}{n}E\right)v = (\lambda - d)v, \quad (6)$$

这里用到了 D = dI 和 v ⊥ e。这表明 v 是 K = 4(I − 1n E)Q^T(Σ + μμ^T)Q(I − 1n E) 的特征向量，其特征值为 μ = λ − d。由于 Q 的秩为 k，且向量 e 属于 Q 的行所张成的空间，因此 Q(I − 1 n E) 的秩为 k − 1。由此可知，K 至多有 k − 1 个正特征值，即 H1 中满足 λ > d 的特征值至多有 k − 1 个。此外，若 $\Sigma + \mu\mu^T$ 为正定矩阵，则恰有 k − 1 个正特征值（引理 13）。K 的其余特征值均为零，也就是说，H1 的对应特征值为 λ = d。

假设 $\Sigma + \mu\mu^T$ 为正定矩阵，则根据上述论证，H1 恰有 $k-1$ 个满足 $\lambda > d$ 的特征值，其对应的正交特征向量为 $v1,\ldots,v_{k-1}$。由（6）可知，$v1,\ldots,v_{k-1}$ 属于$Q(I - \frac{1}{n}E)$ 的行所张成的空间，因此这些特征向量属于Q 的行所张成的空间。根据假设，e 也属于该行空间。因此，由于Q 的秩为k，$v1,\ldots,v_{k-1},e$ 构成Q 的行空间的一组正交基。于是，$P_k = P$。

附录 B

定理 9。 设 A_m 是一列对称 $n \times n$ 矩阵，并在弗罗贝尼乌斯范数意义下收敛到一个对称 $n \times n$ 矩阵 A，即 $\|A_m - A\|_F \to 0$，当 $m \to \infty$ 时。设 $\lambda_1 \ge \ldots \ge \lambda_n$ 为 A 的特征值（按重数计，且不一定非负）。给定 $k \le n$ 并假设 $k = n$ 或 $\lambda_k > \lambda_{k+1}$。此外，设 $e1,\ldots,e_k$ 为分别对应于特征值 $\lambda_1,\ldots,\lambda_k$ 的正交特征向量，并设 $f_{m,1},\ldots,f_{m,k}$ 为对应于 A_m 的 k 个最大特征值（按重数计）的正交特征向量。于是，投影到 $f_{m,1},\ldots,f_{m,k}$ 的张成空间上的正交投影，在弗罗贝尼乌斯范数意义下收敛到投影到 $e1,\ldots,e_k$ 的张成空间上的正交投影。也就是说，定义 $E_k = (e1,\ldots,e_k)$ 和 $F_{m,k} = (f_{m,1},\ldots,f_{m,k})$，则当 $m \to \infty$ 时，$\|F_{m,k}F_{m,k}^T - E_k E_k^T\|_F \to 0$。

证明。如果 $k = n$，则 $E_n E_n^T = I$ 且 $F_{m,n}F_{m,n}^T = I$，命题显然成立。因此，假设 $k < n$。设 $e1$、…、e_n 是 A 的特征向量，分别对应于特征值 λ_1、…、λ_n。设 f_m、1、…、f_m、$_n$ 是 A $_m$ $\mu_{m\,1} \ge$ 的特征向量，$\ldots \ge \mu_{m\,n}$ 对应于特征值、、。可以假设所有特征向量均为标准正交。

由于当 m → ∞ 时 ‖A − A_m‖_F^2 → 0，因此 A_m 的每个元素都收敛于 A 中相应的元素。于是，A_m 的特征函数收敛于 A 的特征函数，并且 A_m 的特征值收敛于 A 的特征值，即 μ_m，当 j → λ_j 时 j = 1,…, n，且 m → ∞。令 T_m 满足 E_n = F_{m,n}T_m。由于 E_n 和 F_{m,n} 是正交矩阵，因此 T_m 也是正交的。对第一行和第三行应用引理 10 得到

$$\begin{aligned}\|A - A_m\|_F^2 &= \|AE - A_m E_n\|_F^2 = \|E\operatorname{diag}(\lambda_1,\ldots,\lambda_n) - A_m F_{m,n} T_m\|_F^2 \\ &= \|F_{m,n}T_m \operatorname{diag}(\lambda_1,\ldots,\lambda_n) - F_{m,n}\operatorname{diag}(\mu_{m,1},\ldots,\mu_{m,n})T_m\|_F^2 \\ &= \|T_m \operatorname{diag}(\lambda_1,\ldots,\lambda_n) - \operatorname{diag}(\mu_{m,1},\ldots,\mu_{m,n})T_m\|_F^2 \\ &= \sum_{i=1}^n\sum_{j=1}^n (\lambda_j T_{m,ij} - \mu_{m,i}T_{m,ij})^2 = \sum_{i=1}^n\sum_{j=1}^n T_{m,ij}^2(\lambda_j - \mu_{m,i})^2.\end{aligned}$$

根据假设，$\lambda_k > \lambda_{k+1}$。因此，根据特征值的收敛性，对于 $j \le k$、$i \ge k+1$，或 $j \ge k+1$、$i \le k$，当 $m \to \infty$ 时，我们有 $T_{m,ij} \to 0$。

此外，

$$\begin{aligned}E_k E_k^T - F_{m,k}F_{m,k}^T &= \sum_{\ell=1}^k \left(e_\ell e_\ell^T - f_{m,\ell}f_{m,\ell}^T\right) \\ &= \sum_{\ell=1}^k \left(\left(\sum_{a=1}^n f_{m,a}T_{m,a\ell}\right)\left(\sum_{a=1}^n f_{m,a}T_{m,a\ell}\right)^T - f_{m,\ell}f_{m,\ell}^T\right) \\ &= \sum_{\ell=1}^k \left(\sum_{a=1}^n\sum_{b=1}^n T_{m,a\ell}T_{m,b\ell}f_{m,a}f_{m,b}^T - f_{m,\ell}f_{m,\ell}^T\right) \\ &= \sum_{a=1}^n\sum_{b=1}^n\sum_{\ell=1}^k T_{m,a\ell}T_{m,b\ell}f_{m,a}f_{m,b}^T - \sum_{\ell=1}^k f_{m,\ell}f_{m,\ell}^T. \\ &= \sum_{(a,b)\in\{1,\ldots,n\}^2\setminus A_{1,k}} \left(\sum_{i=1}^k T_{m,ai}T_{m,bi}\right) f_{m,a}f_{m,b}^T \\ &\quad + \sum_{(a,a)\in A_{1,k}} \left(\sum_{i=1}^k T_{m,ai}T_{m,ai} - 1\right) f_{m,a}f_{m,a}^T,\end{aligned}$$

其中 $A_{i,j} = \{(a、a) : i \le a \le j\}$。

由引理 11，在弗罗贝尼乌斯内积下，对于 $(a,b) \ne (c,d)$，有 f_m, $_a f_{Tm,b} \perp f_{m,c} f_{Tm,d}$。此外，对于所有 a,b，$\|f_{m,a}f_{m,b}^T\|_F = 1$。因此，

$$\begin{aligned}&\|E_k E_k^T - F_{m,k}F_{m,k}^T\|_F^2 \\ &= \sum_{(a,b)\in\{1,\ldots,n\}^2\setminus A_{1,k}} \left(\sum_{i=1}^k T_{m,ai}T_{m,bi}\right)^2 + \sum_{(a,a)\in A_{1,k}} \left(\sum_{i=1}^k T_{m,ai}T_{m,ai} - 1\right)^2. \\ &= \sum_{(a,b)\in\{1,\ldots,n\}^2\setminus A_{1,n}} \left(\sum_{i=1}^k T_{m,ai}T_{m,bi}\right)^2 + \sum_{(a,a)\in A_{k+1,n}} \left(\sum_{i=1}^k T_{m,ai}T_{m,ai}\right)^2 \\ &\quad + \sum_{(a,a)\in A_{1,k}} \left(\sum_{i=1}^k T_{m,ai}T_{m,ai} - 1\right)^2. \quad (7)\end{aligned}$$

如上所述，当 $j \le k$、$i \ge k+1$，或 $j \ge k+1$、$i \le k$ 时，$T_{m,ij} \to 0$ 等于 $m \to \infty$。利用这一点以及 T_m 的正交性可得

$$\sum_{i=1}^{k} T_{m,ai}T_{m,bi} = \sum_{i=1}^{n} T_{m,ai}T_{m,bi} - \sum_{i=k+1}^{n} T_{m,ai}T_{m,bi} \to \begin{cases} 0 & \text{if } a \neq b, \\ 1 & \text{if } a = b, \end{cases}.$$

代入（7）得 $\|E_k E_k^T - F_{m,k} F_{m,k}^T\|_F^2 \to 0$，因为 $m \to \infty$。 □

引理 10。设 A 是一个 $a \times b$ 矩阵。设 U 为一个 $b \times b$ 正交矩阵，V 为一个 $a \times a$ 正交矩阵。那么，

$$\|A\|_F = \|VA\|_F = \|AU\|_F = \|VAU\|_F.$$

证明。参见 Golub 和 Loan(2013)。 □

引理 11。设 w，x，y，$z \in \mathbb{R}^b$。对于 $a \times b$ 矩阵 A 和 B，令 $\langle A, B \rangle_F = \sum_{i=1}^{a}\sum_{j=1}^{b} A_{ij}B_{ij}$ 表示 A 和 B 的弗罗贝尼乌斯内积，并令 $\langle \cdot, \cdot \rangle$ 表示 $\mathbb{R}^b$ 上的标准内积。那么，$\langle wx^T, yz^T \rangle_F = \langle w, y \rangle \langle x, z \rangle$。特别地，$\|wx^T\|_F = \|w\|_2 \|x\|_2$，且 $wx^T \perp yz^T$，如果 $w \perp y$ 或 $x \perp z$。

证明。注意到

$$\left\langle wx^T, yz^T \right\rangle = \sum_{i=1}^{b}\sum_{j=1}^{b} w_i x_j y_i z_j = \sum_{i=1}^{b} w_i y_i \sum_{j=1}^{b} x_j z_j = \langle w, y \rangle \langle x, z \rangle.$$

因此，如果 $w \perp y$ 或 $x \perp z$，则 $\langle wx^T, yz^T \rangle = \langle w, y \rangle \langle x, z \rangle = \|w\|_2^2 \|x\|_2^2$，且 $wxyz$，且 $wxwx$、wxw、wx、xwx，使得 $\|wx^T\|_F = \|w\|_2 \|x\|_2$。 □

引理12。设 $v_1,\ldots,v_\ell$ 是线性无关向量。在张成空间 $(v_1,\ldots,v_\ell)$ 上的正交投影矩阵的弗罗贝尼乌斯范数为 $\sqrt{\ell}$。

证明。不妨设 $v_1,\ldots,v_\ell$ 是标准正交的。于是，我们可以将投影矩阵写成 $P = v_1 v_1^T + \ldots + v_\ell v_\ell^T$。由引理 11，$v_i v_i^T \perp_F v_j v_j^T \neq 11 \|_F^2 = \|_1 v_1^T\|_F^2 + vvv$ 对于 ij。因此，再 $\ldots + \|_\ell v_\ell^T\|_F^2 = \ell$ 由引理，$Pvvvv$。 □

引理 13. 设 A 是一个 $a \times b$ 矩阵，B 是一个 $b \times c$ 矩阵，且二者的秩均为 b，并满足 $a,c \ge b$。令 $C = AB$。则 C 的秩为 b，且 C 的行空间与 B 的行空间相同。

证明。首先证明 rank(C) = b。注意，A 有 b 个线性无关行 $1 \le i_1 < \ldots < i_b \le b$，而 B 有 b 个线性无关列 $1 \le j_1 < \ldots < j_b \le b$。令 $\tilde{}$A 和 $\tilde{}$B 为 b × b 矩阵，其中 $\tilde{}A_{cd} = A_{i_c d}$ 且 $\tilde{}B_{cd} = B_{c j_d}$。那么 $\tilde{}$A 和 $\tilde{}$B 都是可逆矩阵。因此，$\tilde{}C = \tilde{}A\,\tilde{}B = (C_{i_a j_b})_{a,b}$ 也是可逆的，且秩为 k。由此可知 C 的秩为 k。由于 $\tilde{}$A 可逆，$\tilde{}$AB 的行张成空间等于 B 的行张成空间。也就是说，AB 的行张成空间等于 B 的行张成空间。 □

参考文献

Alexander DH、Lange K。2011年。增强用于个体祖源估计的混合算法。BMC Bioinformatics。12:246。Alexander DH、Novembre J、Lange K。2009年。对无亲缘关系个体进行祖源快速模型估计。Genome Res。19:1655–1664。Auton A、Brooks LD、Durbin RM、Garrison EP、Kang HM、Korbel JO、Marchini JL、McCarthy S、McVean GA、Abecasis GR 等。2015年。人类遗传变异的全球参考。Nature。526:68–74。Balding DJ、Nichols RA。1995年。一种量化多等位基因位点群体间分化的方法及其对身份和亲子关系研究的影响。Genetica。96:3–12。Box G、Hunter J、Hunter W。2005年。实验者统计学：设计、创新与发现。Wiley概率与统计学丛书。Wiley。Cabreros I、Storey J。2019年。连接混合模型与主成分分析的无似然群体结构估计量。Genetics。212:1009–1029。Chen X、Storey J。2015年。高维数据中低维潜在结构的一致估计。Conomos M、Reiner A、Weir B、Thornton T。2016年。近期遗传相关性的无模型估计。Am J Hum Genet。98:127–148。Engelhardt B、Stephens M。2010年。群体结构分析：统一框架及基于稀疏因子分析的新方法。PLOS Genetics。6。Evanno G、Regaut S、Goudet J。2005年。使用structure软件检测个体聚类数量：一项模拟研究。Mol Ecol。14:2622–2620。Garcia-Erill G、Albrechtsen A。2020年。推断的混合比例的模型拟合评估。Molecular Ecology Resources。20:936–949。Golub GH、Loan CF。2013年。矩阵计算。约翰斯·霍普金斯数学科学研究丛书。JHU Press。Jacod J、Protter P。2004年。概率学基础。Universitext。Springer。Janes JK、Miller JM、Dupuis JR、Malenfant RM、Gorrell JC、Culling-ham CI、Andrew RL。2017年。k02难题。Mol Ecol。26:3594–3602。Jolliffe IT。2002年。主成分分析（第2版）。Springer统计学丛书。Springer。Jolliffe T、Cadima J。2016年。主成分分析：综述与最新进展。Phil. Trans. R. Soc. A。374:0150202。Lawson D、van Dorp L、Falush D。2018a年。关于如何避免过度解读结构和混合条形图的教程。Nature Communications。9。Lawson DJ、van Dorp L、Falush D。2018b年。关于如何避免过度解读结构和混合条形图的教程。Nat Comm。19:3258。Lazaridis I、Patterson N、Mittnik A、Renaud G、Mallick S、Kirsanow K、Sudmant PH、Schraiber JG、Castellano S、Lipson M 等。2014年。古代人类基因组表明当今欧洲人拥有三个祖先群体。Nature。513:409–413。Meisner J、Liu S、Huang M、Albrechtsen A。2021年。在存在缺失数据的情况下使用主成分分析大规模推断群体结构。Bioinformatics。37:1868–1875。Ochoa A、Storey JD。2019年。f_{ST} 和任意群体结构的亲缘关系I：广义定义。bioRxiv。。Patterson N、Price AL、Reich D。2006年。群体结构与特征值分析。PLOS Genetics。2:e190。Pickrell J、Pritchard J。2012年。从全基因组等位基因频率数据推断群体分裂与混合。PLOS Genetics。8:1–17。Pritchard J、Stephens M、Donnelly P。2000年。使用多位点基因型数据推断群体结构。Genetics。155:945–959。Raj A、Stephens M、Pritchard J。2014年。Faststructure：大型单核苷酸多态性数据集中的群体结构变分推断。Genetics。197:573–589。Wang J。2003年。从遗传数据中最大似然估计混合比例。Genetics。154:747–765。Wang J。2019年。从类似结构的分析中估计群体数量的简约估计量。Mol Ecol Res。19:970–981。