arXiv:2301.06164v1 [stat.AP] 15 Jan 2023

用于探索矩阵间相似性的基于Procrustes的距离

Angela Andreella[1*]、Riccardo De Santis[2], 、Anna Vesely[3], 和 Livio Finos[3]

[1]经济系，威尼斯卡福斯卡里大学，意大利[2]统计系，帕多瓦大学，意大利[3]发展心理学与社会化系，帕多瓦大学，意大利[*]通讯作者：angela.andreella@unive.it

摘要

称为*Procrustes* 分析的统计形状分析通过相似变换最小化矩阵间的距离。该方法返回一组最优正交矩阵，将每个矩阵投影到公共空间中。本文介绍了两种由*Procrustes* 分析推导出的距离，用于探索矩阵间相似性。第一种关注*Procrustes* 分析的残差，即基于残差的距离度量。相比之下，第二种利用拟合得到的正交矩阵，即基于旋转的距离度量。借助这些距离，可以应用多维尺度分析方法等基于相似性的技术，以可视化和探索观测值之间的模式与相似性。事实证明，所提出的距离有助于功能性磁共振成像（fMRI）数据分析。对空间和时间上的脑激活进行测量后，可将其表示为矩阵。将所提出的距离应用于一组受试者（即矩阵）后，发现了一些具有相似神经脑激活模式的个体群组。

关键词： Procrustes 方法、ProMises 模型、正交变换、相似性、功能性磁共振成像组分析、功能性磁共振成像数据

1 引言

其应用领域包括生态学（Saito等人，2015年）、生物学（Rohlf和Slice，1990年）、分析化学计量学（Andrade等人，2004年）、心理计量学（Green，1952年；McCrae等人，1996年）和神经科学（Haxby等人，

2011年），这些领域需要比较以任意坐标系表示的矩阵所描述的信息。与该任意坐标系相对应的矩阵维度会导致功能失配。在此背景下，统计形状分析（Dryden 和 Mardia，2016年）中的*Procrustes* 分析（Gower等人，2004年）可以发挥作用。简而言之，*Procrustes*分析通过相似变换（即旋转、反射、平移和缩放变换）将矩阵对齐到公共参考空间。最优相似变换是使矩阵之间的平方距离最小的变换。

文献中可以找到多种基于*Procrustes*的功能对齐方法；其中最常用的两种是正交*Procrustes*问题（OPP）（Berge，1977）和广义*Procrustes* 分析（GPA）（Gower，1975）。前者处理两个矩阵的对齐，后者则在分析两个以上矩阵时寻找最优相似变换。OPP 有闭式解，而 GPA 基于 Gower（1975）提出的迭代算法。由于*Procrustes* 问题可视为最小二乘问题，Goodall（1991）将其转化为统计模型，即扰动模型，其中误差项服从矩阵正态分布（Gupta 和 Nagar，2018年）。

神经科学是广泛应用基于*Procrustes*方法的领域之一。尤其是，功能性磁共振成像（fMRI）是研究人类认知神经基础最常用的技术。脑激活表现为认知刺激序列与测得的血氧水平序列（血氧水平依赖，BOLD）之间的相关性。神经活动会引起脑血流动力学变化，从而影响磁共振信号的局部强度，即体素强度（单个体积元素）。然而，在受试者之间进行分析（例如分类分析、推断分析）时，会出现各种关键问题。即使向参与者呈现时间同步刺激，受试者之间的大脑解剖结构和功能结构仍存在显著差异（Watson 等人，1993；Tootell 等人，1995；Hasson 等人，2004）。因此，对齐是功能性磁共振成像组水平分析预处理流程中的重要环节。解剖标准化（例如 Talairach（1988年）；Jenkinson 等人（2002）；Fischl 等人（1999））通过仿射变换校正解剖错位，将脑图像对齐到标准解剖模板（例如 Talairach模板（Talairach，1988年）、蒙特利尔成像研究所（MNI）模板（Collins 等人，1994））。然而，解剖对齐无法捕捉受试者之间的功能变异，这是神经科学文献中众所周知的问题（Watson 等人，1993；Hasson 等人，2004；Tootell 等人，1995）。

一个受试者的脑激活可以用一个矩阵来描述，其中行表示时间点/刺激，列表示体素。因此，每一行都展示了对某一刺激在所有体素上的响应激活，而每一列则表示各体素的激活时间序列。功能错位可以聚焦于矩阵之间的列，即

不同受试者的激活时间序列彼此不对应，而响应激活彼此对应，因为刺激通常在时间上同步（Haxby 等人，2011年；Andreella 和 Finos，2022）。在功能性磁共振成像数据的背景下，最流行的基于 $Procrustes$ 的功能对齐方法之一，是 Haxby 等人（2011）提出的超对齐 技术，它是 OPP 的一种序贯方法。然而，OPP、GPA 和超对齐 都存在对齐矩阵（即功能性磁共振成像图像）及相关结果（例如统计 t 检验、分类器系数）可解释性较低的问题，也不适用于高维数据。具体而言，在功能性磁共振成像数据分析中，第一个问题会导致最终对齐图像失去解剖学解释，而第二个问题则使得对整个大脑应用对齐方法成为不可能。可解释性较低的原因在于基于 $Procrustes$ 方法的结构不适定：这类方法无法为最优正交变换返回唯一解。有关功能性磁共振成像数据分析框架下功能对齐问题的更多详情，请参见 Andreella 等人（2022）。

为此，Andreella 和 Finos（2022年）提出了由 Goodall（1991）开发的扰动模型的扩展，即 ProMises 模型。具体而言，扰动模型将 $Procrustes$ 问题重新表述为统计模型。Andreella 和 Finos（2022年）的扩展着重于在正交矩阵的估计过程中引入惩罚项，并为正交矩阵参数指定适当的先验分布。von Mises-Fisher 分布（Downs，1972年）被用于引入有关公共空间最终结构的先验信息。借助这一点，基于 $Procrustes$ 的方法所存在的非唯一性问题得以解决，从而获得了具有可解释性的正交矩阵变换估计量。这使得能够得到唯一的对齐矩阵以及相关的统计推断结果。由于 ProMises 模型可以被视为一种根据先验信息对零假设进行排序的程序，因此对齐过程不会影响第一类错误（Blinded，2022）。最大后验估计的计算十分直接；事实上，von Mises-Fisher 分布是矩阵正态分布（Gupta 和 Nagar，2018年）的共轭先验，而矩阵正态分布正是 ProMises 模型和扰动模型中误差项的分布。

在这项工作中，我们提出一种方法，利用基于$Procrustes$的方法（例如 GPA、超对齐 和 ProMises 模型）所产生的功能失配信息。我们在此提出两种距离度量（Deza 和 Deza，2006），用以刻画矩阵之间相似性/不相似性的不同视角，例如功能性磁共振成像中的受试者。$Procrustes$方法所求解的最小化问题，也可以定义为对象之间的距离（Dryden 和 Mardia，2016）。这里介绍的第一种距离度量基于求解 Procrustes 问题所得的残差。基于残差的 距离由此表示矩阵/受试者在功能对齐后的差异/相似程度。在这种情况下，该距离度量所刻画的是以下方面的相似性/不相似性：

噪声，因为矩阵在功能对齐后具有相同的方向。相比之下，第二种距离利用 Procrustes 问题的解中的正交矩阵参数。基于旋转的 距离计算这些估计所得正交矩阵之间的平方距离。正如我们将看到的，这种度量衡量的是矩阵/受试者在功能对齐前方向上的不相似/相似程度。

在本文中，我们展示了如何将这些度量用于基于距离的技术，例如多维尺度分析方法（Carroll 和 Arabie，1998）、层次聚类（Murtagh 和 Contreras，2012）以及 t 分布随机邻域嵌入（t-SNE）（Van der Maaten 和 Hinton，2008），以可视化并量化矩阵之间的模式和共有特征（即由多个维度描述的个体）。

本文结构如下。第 2 节介绍基于$Procrustes$的方法。本文的核心内容见第 3 节，其中提出了基于残差的 和基于旋转的 距离。最后，我们在第 4 节说明如何将旋转与残差之间的距离作为一种工具，用于理解功能性磁共振成像数据分析框架中受试者之间潜在的聚类。本文的分析使用 R 包 alignProMises 完成功能对齐，相关软件可通过 https://github.com/angeella/alignProMises 获取；并使用 R 包 rotoDistance 计算基于旋转的 和基于残差的距离，相关软件可通过 https://github.com/angeella/ rotoDistance 获取。

2 普罗克鲁斯特斯分析

设 $\{\boldsymbol{X}_i \in \mathbb{R}^{n\times m}\}_{i=1,\dots,N}$ 是一组待对齐的矩阵。普罗克鲁斯特斯分析采用相似变换（Gower，1975），即缩放、旋转/反射和平移，将 $\{\boldsymbol{X}_i \in \mathbb{R}^{n\times m}\}_{i=1,\dots,N}$ 映射到共同的参考空间中。

如果只分析两个矩阵，即 $N = 2$，我们可以将其中一个矩阵视为共同参考矩阵。此时应用正交 普罗克鲁斯特斯问题（OPP），其定义如下：

$$\min_{\boldsymbol{R}_i,\alpha_i,t_i} ||\alpha(X_i - 1_n t)R - X_j||_F^2 \quad \text{subject to } R_i \in \mathcal{O}(m) \tag{1}$$

其中，$\mathcal{O}(m)$是维度为 m 的正交群，$||\cdot||_F$ 是弗罗贝尼乌斯范数，$\alpha \in \mathbb{R}^+$ 是各向同性缩放，$t \in \mathbb{R}^{1\times m}$ 定义平移向量，$1_n \in \mathbb{R}^{1\times n}$ 是全 1 向量。

最优平移等价于列中心化，而 $\boldsymbol{R}$ 和 α 等于

$$\hat{R} = \boldsymbol{U}\boldsymbol{V}^\top; \quad \hat{\alpha}_{\hat{R}} = \frac{\mathrm{tr}(D)}{||\hat{R}^\top X_i^\top||_{\mathrm{F}}^2} \tag{2}$$

其中，$UDV^\top$ 是 $\boldsymbol{X}_i^\top \boldsymbol{X}_j$的奇异值分解。

如果分析的矩阵多于两个，即 $N > 2$，则必须应用广义*Pro-crustes* 分析（GPA）。在这种情况下，矩阵集合$\{\boldsymbol{X}_i \in \mathbb{R}^{n\times m}\}_{i=1,\ldots,N}$ 通过相似变换映射到共同参考矩阵 $\boldsymbol{M} \in \mathbb{R}^{n\times m}$。该共同参考矩阵可以通过多种方式定义，例如逐元素算术平均。GPA定义为

$$\min_{\boldsymbol{R}_i,\alpha_i,t_i} \sum_{i=1}^{N} ||\alpha_i(X_i - 1_n t_i)R_i - M||_F^2 \quad \text{subject to } R_i \in \mathcal{O}(m). \tag{3}$$

与 OPP 不同，GPA 对 R_i 和 α_i 没有闭式解，因此必须使用迭代算法，并在每一步更新参考矩阵（Gower，1975）。

另一种方法是 Goodall（1991）提出的扰动模型。在该模型中，公式 3 所定义的最小二乘问题被转化为一个统计模型，并假设 $\{\boldsymbol{X}_i\}_{i=1,\ \ldots,\ N}$ 是公共空间$\boldsymbol{M}$的带噪旋转。

扰动模型定义如下:

$$\boldsymbol{X}_i = \alpha_i(\boldsymbol{M} + \boldsymbol{E}_i)\boldsymbol{R}_i^\top + \mathbf{1}_n^\top t_i \tag{4}$$

其中， $\boldsymbol{E}_i$ 是服从正态矩阵分布（Gupta 和 Nagar，2018年）的随机误差矩阵 $\boldsymbol{E}_i \sim \mathcal{MN}_{nm}(0,\ \boldsymbol{\Sigma}_n,\ \boldsymbol{\Sigma}_m)$，且 $\boldsymbol{\Sigma}_n \in \mathbb{R}^{n\times n}$和$\boldsymbol{\Sigma}_m \in \mathbb{R}^{m\times m}$。相似变换由以下参数表示： $\boldsymbol{R}_i$、 α_i和 t_i，这些参数需要针对每个 $i = 1,\ \ldots,\ N$进行估计。最优相似变换 R_i 和 $^{\alpha_i \hat{R}_i}$ 是对 OPP 和 GPA 所得结果的略微修改:

$$\hat{R}_i = \{\boldsymbol{U}_i\boldsymbol{V}_i^\top\}_{i=1,\ldots,N}; \quad \hat{\alpha_i}_{\hat{R}_i} = \frac{\text{tr}(D_i)}{||\Sigma_m^{-1/2\text{^}}R_i^\top X_i^\top \Sigma_n^{-1/2}||_{\text{F}}^2} \quad \forall i \in \{1,\ldots,N\} \tag{5}$$

其中， $U_iD_iV_i^\top$是 $\boldsymbol{X}_i^\top\Sigma_n^{-1}\boldsymbol{X}_j\Sigma_m^{-1}$的奇异值分解。

Andreella 和 Finos（2022）提出了扰动模型的扩展，其中正交矩阵参数 $\boldsymbol{R}_i$ 服从 von Mises-Fisher 分布（Downs，1972年）：

$$f(\boldsymbol{R}_i) \sim C(\boldsymbol{F}, k)\exp(k\boldsymbol{F}\boldsymbol{R}_i)$$

其中， $\boldsymbol{F} \in \mathbb{R}^{m\times m}$ 是位置矩阵参数， $k \in \mathbb{R}^+$ 表示正则化参数， $C(\boldsymbol{F},\ k)$ 是归一化常数。Andreella 和 Finos（2022）发现，最大后验估计只是对 Goodall（1991）提出的扰动模型稍作修改（即不对 $\boldsymbol{R}_i$施加冯·米塞斯-费舍尔先验分布）。参数集 $\{\boldsymbol{R}_i\}_{i=1,\ \ldots,\ N}$ 和 $\{\boldsymbol{\alpha}_i\}_{i=1,\ \ldots,\ N}$ 的估计量本质上相同，只是分解的对象是 $\boldsymbol{X}_i^\top\Sigma_n^{-1}\boldsymbol{M}\Sigma_m^{-1} + k\boldsymbol{F}$ 而非 $\boldsymbol{X}_i^\top\Sigma_n^{-1}\boldsymbol{M}\Sigma_m^{-1}$。这些直接的解法得益于冯·米塞斯–费舍尔分布与矩阵正态分布的共轭性（Green 和 Mardia，2006；Andreella 和

Finos，2022）。因此，先验信息直接融入估计过程中的奇异值分解步骤。

对正交矩阵参数 $\boldsymbol{R}_i$ 施加先验分布的动机，源自这样一个假设：“解剖对齐与真实情况相差不远”。通过适当定义先验位置参数 $\boldsymbol{F} \in \mathbb{R}^{m\times m}$，便可将体素三维空间坐标的信息纳入估计过程。Andreella 和 Finos (2022) 将 $\boldsymbol{F}$ 定义为相似欧几里得距离。这样一来，组合较近体素的旋转载荷会高于组合相距较远体素的旋转载荷。此外，将 $\boldsymbol{F}$ 定义为相似欧几里得矩阵，会使 $\boldsymbol{X}_i^\top \Sigma_n^{-1}\boldsymbol{M}\Sigma_m^{-1} + k\boldsymbol{F}$ 满秩，即 $\boldsymbol{R}_i$存在唯一解。

最后，Andreella 和 Finos (2022) 针对高维数据提出了高效版本的 ProMises 模型。当$m >> n$时，问题便会出现，因为 ProMises 模型以及扰动模型都必须计算 N 个维度为$m \times m$的矩阵的奇异值分解。
Andreella 和 Finos (2022) 使用特定的半正交变换，将矩阵 $\boldsymbol{X}_i \in \mathbb{R}^{n\times m}$ 投影到低维空间$\mathbb{R}^{n\times n}$中。具体而言，若将来自 $\boldsymbol{X}_i$ 的薄奇异值分解（Bai 等，2000）的半正交变换$\boldsymbol{Q}_i$ 作为 $m \times n$，则可获得相同的数据拟合效果，同时将时间复杂度从 $\mathcal{O}(m^3)$ 降至 $\mathcal{O}(mn^2)$，并将空间复杂度从 $\mathcal{O}(m^2)$ 降至 $\mathcal{O}(mn)$。

简而言之，高效ProMises 将半正交变换$\boldsymbol{Q}_i$ 应用于 $\boldsymbol{X}_i$，然后在低维矩阵集合 $\{\boldsymbol{X}_i\boldsymbol{Q}_i \in \mathbb{R}^{n\times n}\}$上应用 ProMises 模型。高效 ProMises 模型能够对高维数据进行对齐，例如功能性磁共振成像数据，其中维度 m （即体素数量）约为 200，000。

有关 ProMises 模型及其高效版本的更多详情，请参见 Andreella和Finos (2022)。

3 基于Procrustes的距离

$Procrustes$变换方法（即 OPP、GPA、扰动模型或 ProMises 模型）会寻找正交矩阵，将其应用于原始矩阵后，可使所得矩阵之间的 Frobenius距离最小。因此，自然可以基于这一量——对齐矩阵之间的残差平方和——来定义距离。在这种情况下，我们衡量两个矩阵除旋转之外的差异。两个矩阵可能看起来截然不同，但经过旋转后却可能非常相似。基于残差的距离能够捕捉这一特征，因此仅评估扣除旋转影响后矩阵之间的距离。

我们将定义的第二种距离，基于将一个矩阵 $\boldsymbol{X}_i$ 与另一个矩阵 $\boldsymbol{X}_j$对齐所需的旋转所需的努力。这一努力以求解

ˆ Procrustes 问题 $\boldsymbol{R}_i$ 和 $\boldsymbol{I}_m$ 的正交矩阵之间的距离来衡量（即不进行任何 ˆ 旋转的矩阵）： $\boldsymbol{R}_i$ 与 $\boldsymbol{I}_m$之间的距离越大，将 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$对齐所需的努力就越大。

下面，我们给出基于残差和基于旋转的距离的正式定义：

定义1 考虑一组矩阵 $\{\hat{}\boldsymbol{X}_i \in \mathbb{R}^{n\times m}\}_{i=1,\dots,N}$，它们通过第 *2* 节介绍的某种基于Procrustes的方法进行了功能对齐，即：

$$\hat{\boldsymbol{X}}_i = \hat{\alpha}_{i\hat{\boldsymbol{R}}_i}\boldsymbol{X}_i\hat{\boldsymbol{R}}_i.$$

基于残差的距离定义如下：

$$d_{Re}(\hat{\boldsymbol{X}}_i, \hat{\boldsymbol{X}}_j) = ||\hat{\boldsymbol{X}}_i - \hat{\boldsymbol{X}}_j||_F^2. \tag{6}$$

我们可以注意到，公式6中定义的基于残差的距离与公式 3中定义的 GPA 直接相关。若考虑两个矩阵，该距离就是这对矩阵在 GPA 最小化问题中的贡献，确切地说，是优化过程中的残差。

ˆ**定义2** 考虑一组正交矩阵 $\{\boldsymbol{R}_i \in \mathcal{O}(m)\}_{i=1,\dots,N}$，它们通过第 *2* 节介绍的某种基于Procrustes的方法估计得到。基于旋转的距离定义如下：

$$d_{Ro}(\hat{\boldsymbol{R}}_i, \hat{\boldsymbol{R}}_j) = ||\hat{\boldsymbol{R}}_i - \hat{\boldsymbol{R}}_j||_F^2. \tag{7}$$

由于两种距离都基于矩阵的弗罗贝尼乌斯范数，因此这意味着 $d_{Re}(\cdot)$ 和 $d_{Ro}(\cdot)$ 可直接视为有效的度量，即距离函数 d_{Re}：$\mathbb{R}^{n\times m} \times \mathbb{R}^{n\times m} \to \mathbb{R}^{\geq 0}$ 和 d_{Ro}：$\mathcal{O}(m) \times \mathcal{O}(m) \to \mathbb{R}^{\geq 0}$。

因此，若 $d_{Re} = 0$，这两个矩阵在不考虑方向特征的情况下，在功能上相似。同样，随着 d_{Re} 增大，在不考虑方向的情况下，功能上的差异也会增大。相反，若 $d_{Ro} = 0$，则两个图像具有相同的方向，即相对于参考矩阵 $\boldsymbol{M}$，在功能上（未）对齐。同样，若 $d_{Ro} > 0$，两个矩阵在列维度方面具有不同的方向。

事实上，基于旋转的距离的定义可以大幅简化，从而简化距离本身的计算和解释。下面将对此进行形式化表述：

命题1 定义*2*中定义的基于旋转的距离可以表示为：

$$d_{Ro}(\hat{\boldsymbol{R}}_i, \hat{\boldsymbol{R}}_j) = 2m - 2tr(\hat{\boldsymbol{R}}_i^\top \hat{\boldsymbol{R}}_j)$$

且取值于 $[0, 4m]$。当ˆ $\boldsymbol{X}_i$、ˆ $\boldsymbol{X}_j \in \mathcal{O}(m)$时，也可以使用公式*6*中定义的基于残差的距离 d_{Re} 得到相同结果。

证明 1

$$d(\hat{\boldsymbol{R}}_i, \hat{\boldsymbol{R}}_j) = ||\hat{\boldsymbol{R}}_i - \hat{\boldsymbol{R}}_j||_F^2 = tr[(\hat{\boldsymbol{R}}_i - \hat{\boldsymbol{R}}_j)^\top (\hat{\boldsymbol{R}}_i - \hat{\boldsymbol{R}}_j)]$$
$$= tr(\hat{\boldsymbol{R}}_i^\top \hat{\boldsymbol{R}}_i - \hat{\boldsymbol{R}}_i^\top \hat{\boldsymbol{R}}_j - \hat{\boldsymbol{R}}_j^\top \hat{\boldsymbol{R}}_i + \hat{\boldsymbol{R}}_j^\top \hat{\boldsymbol{R}}_j) = 2m - 2tr(\hat{\boldsymbol{R}}_i^\top \hat{\boldsymbol{R}}_j).$$

考虑基于旋转的距离，两个ˆ ˆ正交矩阵 $\boldsymbol{R}_i$ 和 $\boldsymbol{R}_j$ 乘积的迹只能取 m 到 $-m$ 之间的值，因为正交矩阵的特征值位于单位圆上（即其模等于1）。如果 $tr(\hat{\boldsymbol{R}}_i^\top \hat{\boldsymbol{R}}_j) = m$，这意味着ˆ $\boldsymbol{R}_i = \hat{\boldsymbol{R}}_j$，因为正交矩阵的所有特征值均为1 的唯一情形是它为单位矩阵。同理，如果迹等于 $-m$，则意味着$\hat{\boldsymbol{R}}_i^\top \hat{\boldsymbol{R}}_j = -I_m$，即ˆ $\boldsymbol{R}_i = -\hat{\boldsymbol{R}}_j$。

随后，针对每一对对齐矩阵 $\{\boldsymbol{X}_i\}_{i=1,\ldots,N}$ 以及每一对正交矩阵 $\{\hat{\boldsymbol{R}}_i\}_{i=1,\ldots,N}$，分别计算基于残差的距离和基于旋转的距离，从而得到全局距离矩阵 $D \in \mathbb{R}^{N\times N}$。所提出的距离矩阵能够汇总来自维度较大的不同矩阵的信息，而其维度最终会更低，即为 $N \times N$ 维。这些距离矩阵在各种应用中都很实用，尤其是在处理大数据时。文献中有多种统计方法以距离矩阵为基础。不过，这些方法通常侧重于分析由单个矩阵描述的若干协变量之间的距离，或分析多个距离矩阵（例如 Carroll 和 Arabie（1998）提出的 INDSCAL 方法）。相比之下，本文提出的距离矩阵可直接概括多个大型矩阵的相似性和不相似性特征。

随后，这个矩阵 D 可用于基于不相似性的算法，例如多维尺度分析技术（Carroll 和 Arabie，1998年）、层次聚类（Murtagh 和 Contreras，2012年）和 t 分布随机邻域嵌入（t-SNE）（Van der Maaten 和 Hinton，2008年）。

4 应用

我们分析了 Smeets 等人（2013）收集的 24 名受试者被动观看食物和非食物（办公用品）图像时的数据。食物/非食物图像交替呈现给参与者（呈现食物图像 24 秒，非食物图像 24 秒），其间平均安排 12 秒的休息时段，屏幕上显示十字准线。食物刺激由一组诱人的食物组成，用于捕捉受试者观看诱人（即美味且高热量）食物图像时与自我调节相关的脑激活（Smeets等，2013年）。

该数据集遵循标准处理流程，使用大脑功能磁共振成像软件库（FSL）（Jenkinson 等人，2012）进行了预处理。使用 FLIRT（Jenkinson 和 Smith，2001）计算向标准空间图像的配准步骤，使用 MCFLIRT 进行运动校正

（Jenkinson 等人，2002），使用 BET（Jenkinson 等人，2002）去除非脑组织，并使用半高全宽为 6 mm 的高斯核进行空间平滑。最后，整个四维数据集的强度归一化通过一个单一的乘法因子计算，并应用了高通时间滤波（高斯加权最小二乘直线拟合，sigma=64.0s）。原始数据集可在 https://openneuro.org/datasets/ ds000157/versions/00001 获取，预处理后的数据集可在 R 包 rotoDistance（ https://github.com/angeella/rotoDistance）中获取。有关实验设计和数据采集的详情，请参阅 Smeets 等人（2013）。

我们分析了由 237 个体素组成的右侧距状沟，该区域参与视觉信息处理，并与参与调节食物摄入的区域相关（Smeets等，2013年）。不过，也可以分析整个大脑，而不只是某个感兴趣区域（例如右侧距状沟）。事实上，所提出的距离能够通过一个低维矩阵，概括复杂的高维数据，例如功能性磁共振成像数据；这类数据通常由维度为 N 矩阵组成，000（即 300 个时间点和 200，000 个体素）。

在预处理数据上拟合 ProMises 模型，随后使用对齐图像计算距离矩阵 $\boldsymbol{D}_{Re} \in \mathbb{R}^{24\times24}$ （即基于残差的距离），并使用相应的最优旋转矩阵计算 $\boldsymbol{D}_{Ro} \in \mathbb{R}^{24\times24}$ （即基于旋转的距离），如第 3 节所述。

这两种基于$Procrustes$的距离捕捉了不同的信息，即功能对齐前后脑激活方面的被试间差异。图1展示了每对受试者的距离 $\boldsymbol{D}_{Re}$ 和 $\boldsymbol{D}_{Ro}$ 。两者之间的相关性很低，即 ≈ 0.03；从图1可以看出， $\boldsymbol{D}_{Re}$ 和 $\boldsymbol{D}_{Ro}$ 对脑激活的被试间差异给出了截然不同的洞见。

现采用多维尺度分析（MDS）技术（Carroll 和 Arabie，1998），并将 $\boldsymbol{D}_{Ro}$ 作为距离矩阵。随后使用 $\boldsymbol{D}_{Re}$进行了第二次分析。为便于比较，我们还计算了未进行功能对齐的图像之间的欧几里得距离。我们将相应的距离矩阵记为 $D_{raw}(\boldsymbol{X}_i, \boldsymbol{X}_j) = ||X_i - X_j||^2_F$。我们使用 smacof R 包（De Leeuw 和 Mair（2009））应用多维尺度分析技术。我们决定采用尽可能灵活的样条多维尺度分析（单调样条变换）。更多详情请参见 De Leeuw 和 Mair（2009）。

此外，我们还要分析每位受试者的一些协变量，这些协变量与年龄、体质指数（BMI）及其他信息一同在表1中作了简要介绍。接下来，我们使用多维尺度分析方法计算出的拟合构型矩阵来分析这些协变量。更多详情请参见 Smeets 等人（2013年）。

首先关注距离矩阵 D_{Ro}，图2显示了考虑多个维度数 $K = \{1, \ldots, 20\}$ 时的应力值

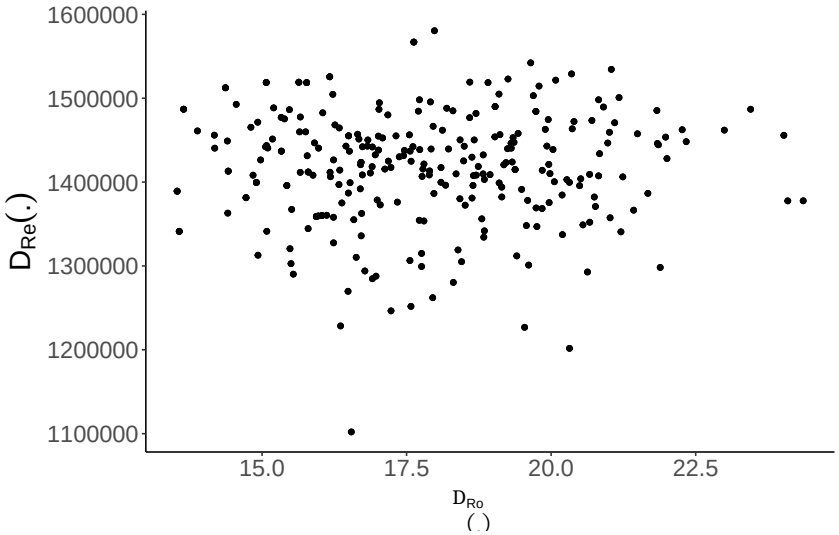


图1： $\boldsymbol{D}_{Re}$ 与 $\boldsymbol{D}_{Ro}$之间的散点图。

协变量	**描述**
饮食重要性	节食重要性（五点评分量表）
饮食成功	饮食成功的五点评分量表
实验前食欲	扫描前食欲，采用包含 3 个项目的五点评分量表（Cronbach's $\alpha = 0.84$）
实验后食欲	扫描后的食欲，采用包含3个条目的五点评分量表（Cronbach's $\alpha = 0.91$）
月经周期阶段	末次月经日期（卵泡期和排卵期、黄体期和月经期）

表1：关于（Smeets等，2013年）数据集的协变量描述

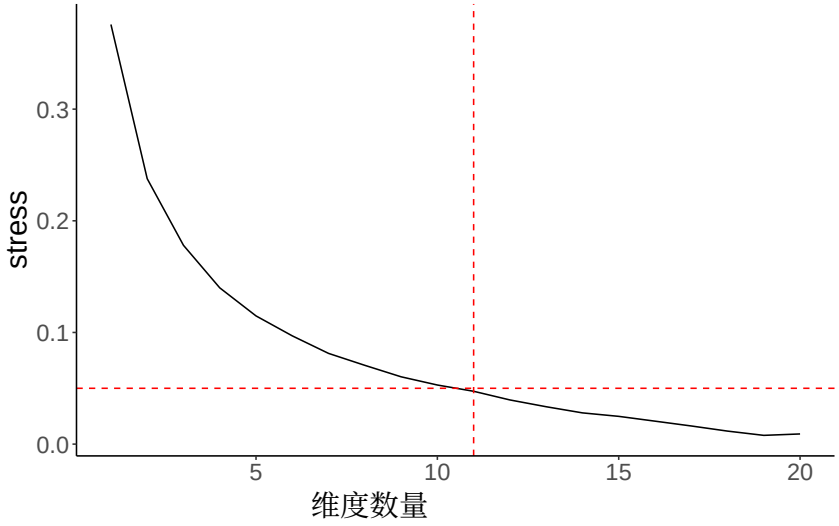


图2： D_{Ro} 分析：在多维尺度分析方法中，考察不同维度数量时的应力值。红色虚线表示应力值等于 0.05，维度数量等于 11。

多维尺度分析方法。我们评估认为， $K = 11$ 是一个合适的值，对应的应力值为 ≈ 0.05。

随后，我们进行了简单的广义线性回归，以协变量为因变量，以通过多维尺度分析拟合的 11 种配置为解释变量。我们发现，饮食成功协变量与第 6 个维度之间存在显著关系（$t = -4.171$， $p = 0.0065$），这与 Smeets 等人（2013年）的研究结果一致。报告的 p 值采用 Bonferroni 方法进行多重检验校正（Goeman 和 Solari，2014）。因此，图3展示了第 1 个和第 6 个拟合配置，以及主要协变量（饮食成功）和月经周期阶段协变量。

乍看之下，我们可以看到 y 轴表示饮食成功程度，其中负值对应较低的饮食成功程度，正值则对应较高的饮食成功程度。例如，我们还可以看到，受试者16和18具有相同的功能失配、相近的饮食成功值以及相同的月经周期阶段。

然而，如果我们改为对基于残差的 距离矩阵应用多维尺度分析，就不会发现像使用基于旋转的 距离时那样清晰的模式，如图4所示。在这种情况下，我们自动将维度数设为11（相当于应力值为0.05）。与基于旋转距离的首次分析不同，广义线性回归未显示任何显著特征。

最后，如果我们使用 D_{raw} 作为距离矩阵（即使用未经功能对齐的图像），就会出现使用 D_{Re} 时相同的情况（甚至更糟）。考虑11个维度时，应力值为0.02；广义线性回归未发现任何显著维度。图5展示了使用 D_{raw}进行多维尺度分析的结果。这两个维度

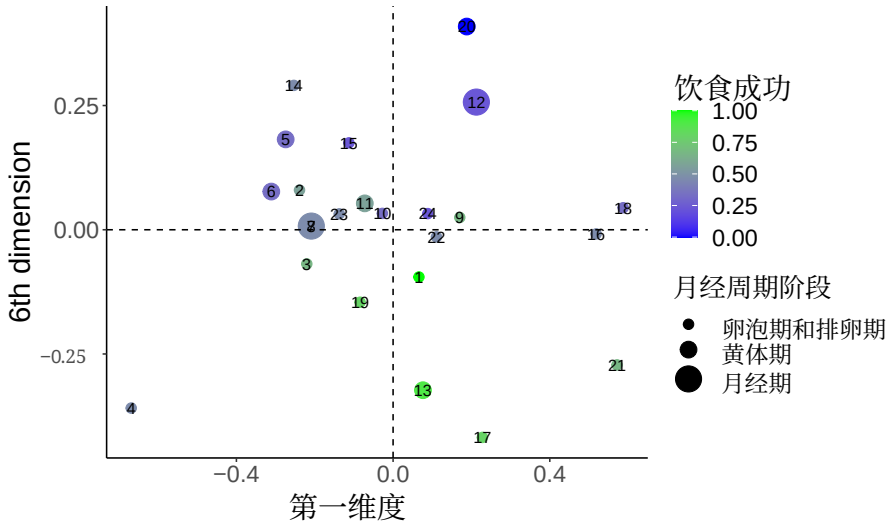


图 3： D_{Ro} 分析： x轴表示通过多维尺度分析方法计算得出的第 1 个拟合配置，而 y轴显示第 6 个拟合配置。颜色渐变表示饮食成功协变量（已缩放），点的大小表示月经周期阶段。

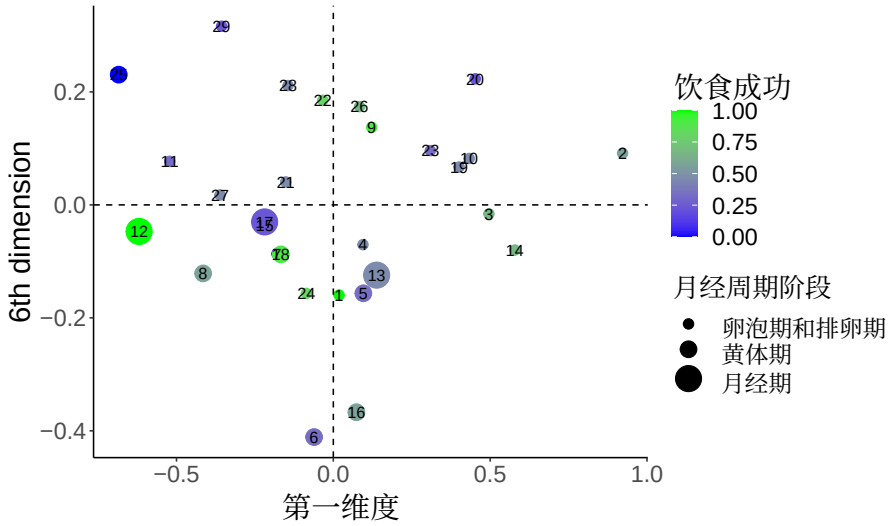


图 4： D_{Re} 分析： x轴表示通过多维尺度分析方法计算得出的第 1 个拟合配置，而 y轴显示第 6 个拟合配置。颜色渐变表示食欲后协变量（已缩放），点的大小表示月经周期阶段。

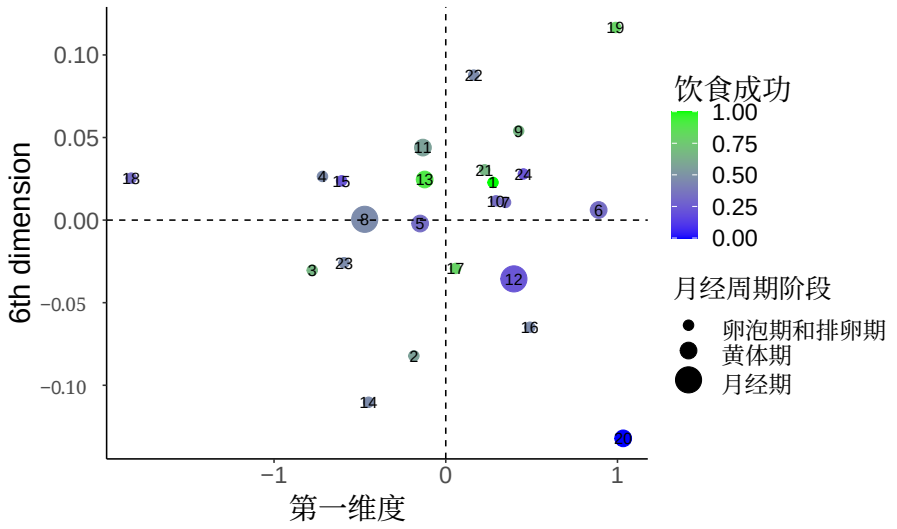


图5： D_{raw} 分析： x轴表示通过多维尺度分析方法计算的第1个拟合构型，而 y轴显示第6个拟合构型。颜色渐变表示食欲后协变量（缩放后），点的大小表示月经周期阶段。

未能捕捉所分析的受试者层面特征。

总之，基于旋转的距离 D_{Ro} 能够通过旋转所需的努力来捕捉神经反应中的功能变异性。随后，受试者的神经激活在方向上存在差异。然而，在应用功能对齐之前和之后，这些差异都会减小；也就是说，在同时分析一些受试者特异性协变量时，考虑残差基础距离 D_{Re}和原始 基础距离 D_{raw}。

5 结论

在本文中，我们提出了基于普罗克鲁斯特的方法所估计的对齐图像和正交变换的普罗克鲁斯特距离。这些距离可从两个独立的角度探索矩阵之间的不相似性。基于残差的距离通过旋转后的功能列来表达不相似性，即消除方向分量。相反，基于旋转的距离则通过矩阵各列的功能（不）对齐来描述不相似性，也就是矩阵各列的方向有多相似。当研究目标是分析以任意坐标系表示的矩阵时，该方法很有帮助。此外，当关注点是探索大数据矩阵之间的距离时，例如在功能性磁共振成像应用中，所提出的距离也具有优势。在这一框架下，每位受试者由一个庞大矩阵表示，其维度约为 $300 \times 200,000$。普罗克鲁斯特距离可在一个维度等于所分析矩阵／受试者数量的空间中探索这些矩阵。

该空间的维度等于所分析的矩阵／受试者数量。在功能性磁共振成像应用中，我们发现这些度量能够可靠地衡量个体差异。事实上，基于普罗克鲁斯特的功能对齐有助于减少地形特异性带来的混杂因素，并捕捉个体间共享的功能和解剖反应周围的变异。本文提出的距离有助于发现共享神经脑激活模式的个体群体。总之，基于普罗克鲁斯特的距离为普罗克鲁斯特方法领域增添了有价值的探索与可视化工具。

致谢

Angela Andreella 衷心感谢帕多瓦大学（意大利）提供的 BIRD2020 / SCAR ASSEGNIBIRD2020_01 资助，以及威尼斯卡福斯卡里大学（意大利）提供的 PON 2014-2020/DM 1062 资助。本稿中的部分计算分析借助帕多瓦大学 2017 年战略研究基础设施资助完成：“CAPRI：用于研究与创新的高性能计算”，http://capri. dei.unipd.it。

作者贡献

Angela Andreella：概念构思、软件、数据整理、正式分析、调查研究以及初稿撰写、审阅与编辑。**RiccardoDe Santis**：概念构思、方法学、初稿撰写、审阅与编辑。**Anna Vesely**：概念构思、方法学、初稿撰写、审阅与编辑。**Livio Finos**：概念构思、方法学、初稿撰写及监督。

利益冲突声明

作者声明不存在利益冲突。

参考文献

Andrade, J. M., Gómez-Carracedo, M. P., Krzanowski, W., 和 Kubista, M.（2004）。分析化学中的 Procrustes 旋转：教程。化学计量学与智能实验室系统，72(2):123–132。Andreella, A. 和 Finos, L.（2022）。高维数据的普罗克鲁斯特斯分析。*Psychometrika*，第 1–17 页。Andreella, A.、Finos, L. 和 Lindquist, M. A.（2022）。借助空间先验信息增强超对齐。*(即将发表) doi: 10.1002/HBM.26170*。Bai, Z.、Demmel, J.、Dongarra, J.、Ruhe, A. 和 van der Vorst, H.（2000）。代数特征值问题求解模板：实用指南。SIAM。Berge, J. M. F.（1977）。两个或多个矩阵的正交Procrustes旋转。*Psychometrika*，42(2):267–276。Blinded, N.（2022）。对功能对齐的fMRI图像进行组分析的有效推断。前言 *XX 1* 全体会议，第 1987 页。Carroll, J. D. 和 Arabie, P.（1998）。多维尺度分析。测量、判断与决策，第 179–250 页。Collins, D. L.、Neelin, P.、Peters, T. M. 和 Evans, A. C.（1994）。在标准 Talairach空间中对MR体积数据进行自动三维受试者间配准。计算机辅助断层扫描杂志，18(2):192–205。De Leeuw, J. 和 Mair, P.（2009）。使用主化方法进行多维尺度分析：R中的SMACOF。统计软件杂志，31:1–30。Deza, M.-M. 和 Deza, E.（2006）。距离词典。Elsevier。Downs, T. D.（1972）。定向统计。*Biometrika*，59(3):665–676。Dryden, I. L. 和 Mardia, K. V.（2016）。统计形状分析：*R*语言应用，第 995 卷。John Wiley & Sons。Fischl, B.、Sereno, M. I.、Tootell, R. B. 和 Dale, A. M.（1999）。皮层表面的高分辨率受试者间平均与坐标系。人脑绘图，8(4):272–284。Goeman, J. J. 和 Solari, A.（2014）。基因组学中的多重假设检验。医学统计，33(11):1946–1978。Goodall, C.（1991）。形状统计分析中的Procrustes方法。英国皇家统计学会的 *Wiley*，53(2):285–339。

Gower, J. C.（1975）。广义普罗克鲁斯特分析。*Psychometrika*, 40(1):33– 51。Gower, J. C., Dijksterhuis, G. B., 等（2004）。*Procrustes* 问题，第 30 卷。Oxford University Press on Demand。Green, B. F.（1952）。因子分析中斜交结构的正交近似。*Psychometrika*, 17(4):429–440。Green, P. J. 和 Mardia, K. V.（2006）。使用层次模型进行贝叶斯对齐及其在蛋白质生物信息学中的应用。*Biometrika*，93(2):235–254。Gupta, A. K. 和 Nagar, D. K.（2018年）。矩阵变量分布。Chapman and Hall/CRC。Hasson, U., Nir, Y., Levy, I., Fuhrmann, G. 和 Malach, R.（2004）。自然视觉过程中皮层活动的被试间同步。*Science*, 303(5664):1634–1640。Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M. 和 Ramadge, P.（2011）。人类腹侧颞叶皮层表征空间的通用高维模型。*Neuron*，72(1):404–416。Jenkinson, M., Bannister, P., Brady, M. 和 Smith, S.（2002）。改进脑图像稳健、精确线性配准及运动校正的优化方法。*Neuroimage*，17(2):825–841。Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. 和 Smith, S. M.（2012）。FSL。*Neuroimage*，62(2):782–790。Jenkinson, M. 和 Smith, S.（2001）。一种用于脑图像稳健仿射配准的全局优化方法。*Medical Image Analysis*, 5(2):143–156。McCrae, R. R., Zonderman, A. B., Costa Jr, P. T., Bond, M. H. 和 Paunonen, S. V.（1996）。评估修订版 NEO 人格量表中因子的可重复性：验证性因子分析与 Procrustes 旋转。*Journal of Personality and Social Psychology*，70(3):552。Murtagh, F. 和 Contreras, P.（2012）。层次聚类算法：综述。*Wiley Interdisciplinary Reviews: Data Mining and KnowledgeDiscovery*，2(1):86–97。Rohlf, F. J. 和 Slice, D.（1990）。对 Procrustes 方法的扩展：标志点的最优叠合。*Systematic Biology*，39(1):40–59。Saito, V. S., Fonseca-Gessner, A. A. 和 Siqueira, T.（2015）。生态学家应如何定义采样力度？Procrustes 分析在研究群落组成变异方面的潜力。*Biotropica*，47(4):399–402。

Smeets, P. A., Kroese, F. M., Evers, C. 和 de Ridder, D. T.（2013）。受诱惑还是警觉：大脑对食物诱惑的反制控制反应。行为脑研究，248:41–45。Talairach, J.（1988年）。人脑共面立体定向图谱——三维比例系统。脑成像方法。Tootell, R., Reppas, J. B., Kwong, K. K., Malach, R., Born, R. T., Brady, T. J., Rosen, B. R. 和 Belliveau, J. W.（1995）。利用磁共振成像对人类 MT 区及相关视觉皮层区域进行功能分析。《神经科学杂志》，15(4):3215–3230。Van der Maaten, L. 和 Hinton, G.（2008年）。使用 t-SNE 可视化数据。《机器学习研究杂志》，9(11)。Watson, J. D., Myers, R., Frackowiak, R. S., Hajnal, J. V., Woods, R. P., Mazziotta, J. C., Shipp, S. 和 Zeki, S.（1993年）。人脑 V5 区：正电子发射断层扫描与磁共振成像联合研究的证据。《大脑皮层》，3(2):79–94。