

Reference


nature
computational
science

ARTICLES

<https://doi.org/10.1038/s43588-021-00029-8>



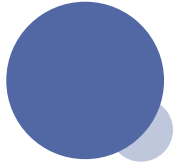
Similarity-driven multi-view embeddings from high-dimensional biomedical data

Brian B. Avants  , Nicholas J. Tustison  and James R. Stone

Diverse, high-dimensional modalities collected in large cohorts present new opportunities for the formulation and testing of integrative scientific hypotheses. Similarity-driven multi-view linear reconstruction (SiMLR) is an algorithm that exploits inter-modality relationships to transform large scientific datasets into smaller, more well-powered and interpretable low-dimensional spaces. SiMLR contributes an objective function to identify joint signal regularization based on sparse matrices representing prior within-modality relationships and an implementation that permits application to joint reduction of large data matrices. We demonstrate that SiMLR outperforms closely related methods on supervised learning problems in simulation data, a multi-omics cancer survival prediction dataset and multiple modality neuroimaging datasets. Taken together, this collection of results shows that SiMLR may be applied to joint signal estimation from disparate modalities and may yield practically useful results in a variety of application domains.

来自高维生物医学数据的相似性驱动的多视图嵌入

文献来源：2021.2.22/nature computational science/弗吉尼亚大学放射学和医学影像系



Abstract & introduction

多视图（也称为多模态或多块） (Multi-view (also known as multiple modality or multi-block)) **数据集**在生物医学科学中越来越常见。

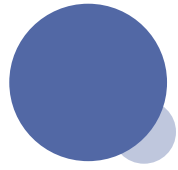
在理想化的情况下，每个视图或模态将提供底物生物学的完全唯一的测量。然而往往每一种观点都对一种复杂的现象提供了一种**部分的、而不是完全独立的观点**。在这种情况下，可以利用协变来筛选噪声测量并更好地识别有意义的信号。

预先指定的联合假设使科学家能够避免对可能的相互作用进行组合爆炸测试。尽管在足够大的、充分理解的数据集中功能强大。(PCA&ICA)

相似性驱动的多视图线性重建 (SiMLR), 是一种-针对生物医学数据提出的联合嵌入方法。它采用两种或多种模式作为输入, 允许自定义正则化模型。

SiMLR 为每个最能预测其伙伴模态 (partner modalities) 的模态输出局部最优的低维矩阵嵌入。它通过从伙伴模态派生的基组重构每个模态矩阵来实现这一点。

SiMLR是一种**利用模态间关系将大型科学数据集转换为更小、更强大且可解释的低维空间的算法。**



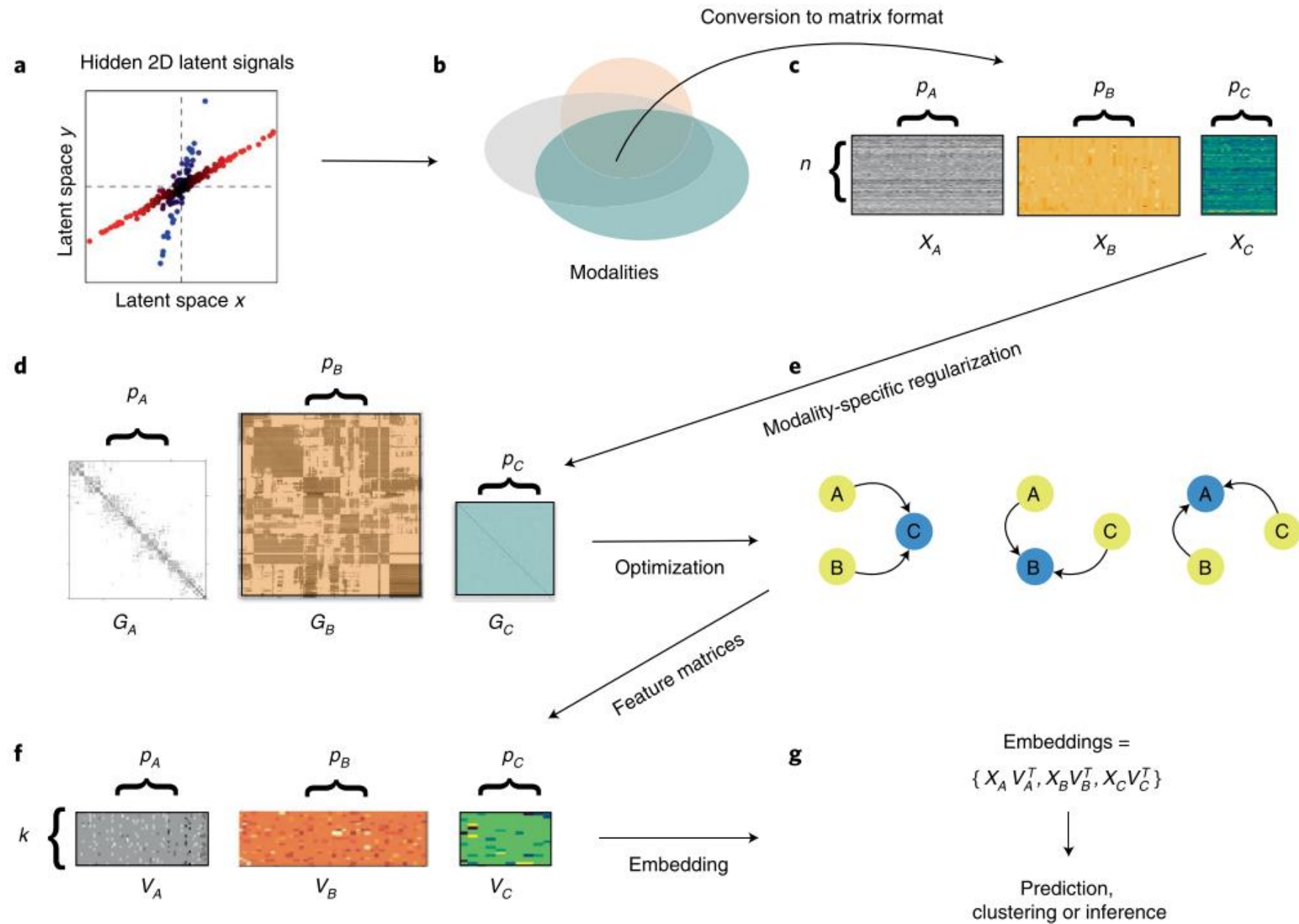
method & results

相关术语介绍

- **Multi-view:** several modalities collected in one cohort; alternatively, the same measurements taken across different studies⁶¹. We focus on the first case here.
 - 在一个队列中收集了几种模式；或者，在不同的研究中进行相同的测量。
- **Covariation:** we use the term in two contexts. As a general concept, we mean systematic changes in one modality are reflected in a predictable amount of change in other modalities. In the mathematical context, we use the definition of covariation for discrete random variables.
 - 作为一般概念，意思是一种模式的系统性变化反映在其他模式的可预测变化量中。
 - 在数学分析中，使用离散随机变量的协方差的定义。

- **Latent space or embeddings:** both terms refer to a representation of high-dimensional data that is often lower-dimensional. These are also known as components in PCA. In the context of this paper, we are ‘approximating’ the (hidden) latent space with the learned embeddings. Often, the true latent space cannot be known. We compute embeddings (or components) by multiplying feature vectors against input data matrices. Importantly, SiMLR can compute latent spaces that target either statistical independence (the ICA source separation algorithm⁴⁰) or orthogonality (the SVD algorithm). Deflation-based schemes, on the other hand, target only orthogonality.

➤ 潜在空间或嵌入：这两个术语都是指通常是低维的高维数据的表示。这些也称为PCA中的关键组件。在本文中，作者用学习的嵌入（embedding）来“逼近”（隐藏的）潜在空间。通常，无法知道真正的潜在空间。通过将特征向量与输入数据矩阵相乘来计算嵌入（或分量）。重要的是，SiMLR可以计算以统计独立性（ICA源分离算法）或正交性（SVD算法）为目标的潜在空间。





Similarity-driven multi-view linear reconstruction. 相似性驱动的多视图线性重建。

SIMLR是一个通用框架，可以以与稀疏PCA（一种类似回归的目标）或稀疏CCA（一种与协方差相关的目标）相关的形式指定。主要形式如图所示。

将应用SIMLR的数据集做了两个假设：

假设 1：真正的潜在信号是独立的，并且在收集多个测量值的生物系统中线性混合（盲源分离的标准假设）。

假设2：稀疏的正则化特征向量可以通过线性运算将假设1中估计的潜在信号与原始数据矩阵相关联。

盲源分离在维基百科的定义： 指的是从多个观测到的混合信号中分析出没有观测的原始信号。

盲信号的“盲”字强调了两点：1)原始信号并不知道；2)对于信号混合的方法也不知道。

最常用在的领域是在数字信号处理，且牵涉到对混合讯号的分析。盲信号分离最主要的目标就是将原始的信号还原出原始单一的讯号。

Data representation.

SiMLR假设输入是“干净”的数据。这些数据没有缺失值，并且以矩阵格式构造，每个模态沿着行(主题或样本)和列(表示特征)匹配。

The SiMLR objective function.

SIMLR的核心概念是，它结合了灵活的方法来测量模态之间的差异(相似性驱动)，可以将几个不同的矩阵作为输入(多视图)，并执行本质上都是线性代数的操作(线性重建)。

首先，对于给定的测量或视图或模态，将 X_i 定义为 $N \times P_i$ (主题特征)矩阵。 i 的值的范围从1到 m ， m 是模态(或视图)的数量。然后，SIMLR优化目标函数，该目标函数通过稀疏特征矩阵(V_i)和低维表示($U_{\neq i}$)寻求从其伙伴矩阵逼近每个模态：

$$\arg \min_{V_i} \sum_{i=1}^m S(X_i, f(U_{\neq i}), V_i) + \text{Regularization}(V_i),$$

Multiple regression.

多元回归解决了一个最小二乘问题，该问题将多个预测因子 ($n \times p$ 矩阵 X) 最优拟合到一个结果 (Y)。

$$\arg \min_{\beta} \|y - X\beta\|^2,$$

具有最优最小二乘解

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

文中，SiMLR 在进行矩阵线性重构过程中的 V_i 对应 X_i 模态的特征或解向量（类似于 β ）的 $p_i \times k$ 矩阵；

$$\|y - x\beta\|^2 = (y - x\beta)^T \cdot (y - x\beta)$$

由矩阵微分 $\frac{\partial x^T a}{\partial x} = \frac{\partial a^T x}{\partial x} = a$

$$\frac{\partial x^T A x}{\partial x} = Ax + A^T x$$

$$Ax + A^T x = 2Ax \quad (A \text{ 对称})$$

$$(y - x\beta)^T (y - x\beta) = (x\beta - y)^T (x\beta - y) = \beta^T X^T x \beta - y^T x \beta - \beta^T X^T y + y^T y$$

$$\frac{\partial \|y - x\beta\|^2}{\partial \beta} = 2X^T x \beta - 2X^T y$$

$$\text{令 } \frac{\partial}{\partial \beta} = 0 \Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y$$

Principal component analysis.

$$\arg \min_{U,V} \|X - UV^T\|^2 + \sum_k \lambda_k \|V_k\|_1,$$

$$U = XV \text{ and } V^T V = I,$$

对于SiMLR, 有:

$$\forall_i U_i = X_i V_i;$$

$$\arg \min_{V_i} \sum_{i=1}^m S(X_i, f(U_{\neq i}), V_i) + \text{Regularization}(V_i),$$

除 X_i 以外的模态的低维表示组成的列矩阵

根据假设1, 真实的潜在信号在整个生物系统中是独立且线性混合的, 所以有 $f()$

$$f(U_{\neq i}) = \tilde{U}_{\neq i}$$

通过对 $j \neq i$ 的集合执行盲源分离得到的，本文主要采用的是**独立成分分析**（Independent component analysis, ICA）和奇异值分解，依据不同情况可以合理选用其他分离方法。

通过对 $j \neq i: \{X_j V_j\}$ 的集合执行盲源分离得到 $(U_{\neq i})$ embedding，以低维嵌入再去逼近潜在空间。

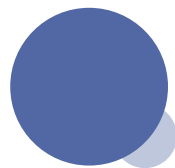
Similarity options.

相似项

根据假设2，稀疏的、正则化的特征向量可以通过线性运算将假设1中的估计的潜在信号与原始数据矩阵相关联，所以这里定义相似项s，测量来自其他模态的 X_i 的近似的质量的函数，

$$S(X_i, \tilde{U}_{\neq i}, V_i) = \| X_i - \tilde{U}_{\neq i} V_i^T \|^2.$$

SIMLR试图在最小误差意义上直接从其他N-1个模态的基表示来重构每个矩阵 X_i 。



Results

Table 1 | Summary of experimental results

| Study | RGCCA | SGCCA | SiMLR-CCA-ICA | SiMLR-CCA-SVD | SiMLR-Reg-ICA | SiMLR-Reg-SVD | Metric |
|--------------------|-----------------|-----------------|-----------------|-----------------------------------|-----------------------------------|-----------------|------------------|
| Signal sensitivity | 0.35 ± 0.18 | 0.45 ± 0.17 | 0.5 ± 0.15 | 0.51 ± 0.14 | 0.49 ± 0.13 | 0.49 ± 0.14 | <i>R</i> squared |
| Noise sensitivity | 0.09 | 0.16 | 0.09 | 0.06 | 0.07 | 0.1 | <i>R</i> squared |
| Multi-omic | N/A | 0.56 ± 0.12 | 0.56 ± 0.13 | 0.56 ± 0.14 | 0.64 ± 0.08 | 0.64 ± 0.11 | Concordance |
| Brain age | N/A | 2 ± 1.5 | 1.6 ± 1.2 | 1.4 ± 1.2 | 1.6 ± 1.3 | 1.7 ± 1.2 | MAE |
| PING anxiety | N/A | 1 component | N/A | 3 components | 3 components | N/A | Inferential |
| PING depression | N/A | 0 components | N/A | 1 component (trend) | 1 component (trend) | N/A | Inferential |



总结

SiMLR 提取的信号与相关方法的信号不同。该特征可能与该方法的核心有关：高维嵌入向量纯粹由模态内数据构建，而低维基则来自用户选择的源分离算法确定的跨模态表示。

如果选择 SVD 源分离方法，则表示将是正交的；如果选择 ICA，它们将在统计上独立，其中独立性是通过测量非高斯性来定义的（FastICA 的原则之一是“非高斯性就是独立性”）。

这种方法仅在表现出某种程度的跨模态协变的数据集中有效，这些协变可以有意义地解码为多个“真实”源信号。