arXiv:2304.04692v1 [stat.ME] 10 Apr 2023

面向多视图数据整合与预测的可扩展随机核方法

预印本

SandraE. Safo*明尼苏达大学双城校区生物统计学部，明尼阿波利斯，明尼苏达州 55455
ssafo@umn.edu

Han Lu明尼苏达大学双城校区生物统计学部，明尼阿波利斯，明尼苏达州 55455
lu000054@umn.edu

2023年4月11日

摘要

我们开发了可扩展的随机核方法，用于联合关联来自多个来源的数据，并同时预测结果变量，或将一个单元分类为两个或更多类别中的一个。所提出的方法对多视图数据中的非线性关系进行建模，同时预测临床结局，并能够识别对各视图之间关系贡献最大的变量或变量组。我们利用随机傅里叶基可以近似平移不变核函数这一思想，构建每个视图的非线性映射，并利用这些映射和结局变量学习与视图无关的低维表示。通过模拟研究，我们展示了所提出的方法在多视图数据整合方面优于其他几种线性和非线性方法。当将所提出的方法应用于与COVID-19 相关的基因表达、代谢组学、蛋白质组学和脂质组学数据时，我们识别出了若干与COVID-19状态和严重程度相关的分子特征。真实数据应用和小样本量模拟的结果表明，所提出的方法可能适用于小样本量问题。

可用性： 我们的算法采用 PyTorch 实现，并通过 R 提供接口，网址为：https://github.com/lasandrall/RandMVLearn。

关键词 多高维 多视图学习 · 数据整合 · 非线性 · 核 · 随机傅里叶特征 · 数据

1 引言

许多生物医学研究会针对同一组个体生成多种类型的数据（例如基因组学、蛋白质组学），其典型目标是通过这些独特但互补的数据，理解复杂疾病的机制。对于许多复杂疾病，已有成熟的治疗和风险分层方法，但这些方法往往针对的是症状，而非潜在机制。要理解复杂疾病的病理生物学，需要采用一种超越对每种数据类型分别进行分析的分析方法。

1.1 COVID-19 分子研究

关于COVID-19 的病理生物学，仍有许多内容有待了解。研究表明，患有严重 COVID-19和未患有严重COVID-19 的患者具有不同的遗传、病理和临床特征 [Severe-Covid-19-GWAS-Group,2020,Overmyer 等人, 2021]。这凸显了使用多种分子数据来更好地了解该疾病严重程度的必要性。我们的工作受到 Overmyer等人分析的 COVID-19 相关数据的启发 [2021]。血液样本

*C通讯作者：Sandra Safo，www.sandraesafo .com

这些样本取自128名患者，他们患有类似于COVID-19的中度至重度呼吸问题，并于2020年4月6日至2020年5月1日期间入住纽约州奥尔巴尼医学中心。其中，102人和26人分别检测出COVID-19阳性和阴性。对血液样本进行了代谢组学、RNA测序（RNA测序）、蛋白质组学和脂质组学定量分析。在Overmyer等人 [2021], 的研究中，使用包括套索回归在内的机器学习方法，将分子与疾病严重程度相关联，并确定COVID-19严重程度的关键决定因素。主要分析侧重于分别研究每种分子数据。在次要分析中进行了成对相关性分析，以确定成对分子数据之间的关联。

我们采用整体性方法，同时整合结果变量与多种分子数据，从而利用所有分子数据中的信息，同时纳入先验生物学信息。我们的目标是识别可能影响 COVID-19 严重程度和状态的分子特征与通路，最终进一步揭示 COVID-19 严重程度的发病机制。Overmyer 等人 [2021] 使用两种方法衡量疾病严重程度：i) 世界卫生组织（WHO）0–8 级疾病特异性量表，其中 8 表示死亡；ii) 满分为 45 分的评分，表示无住院天数（HFD-45）。值得注意的是，HFD-45 值为 0 意味着患者在 45 天后仍在住院，或患者已经死亡。建议采用 HFD-45 指标，因为该指标并非 COVID-19 [特异性指标（Overmyer 等人， 2021]）。

1.2 现有方法

人们提出了许多线性方法和非线性方法，用于整合来自多个来源的数据。有些方法会学习每个视图的低维表示，使各视图之间的相关性最大化（例如，典型相关分析[CCA [Hotelling， 1936]], Deep CCA [Andrew等人，2013a], 核CCA [Lopez-Paz等人， 2014]）。其他方法则学习所有视图的视图无关低维表示（例如，广义典型相关分析 [Horst，1961，Kettenring，1971], 深度广义典型相关分析 [Deep GCCA [Benton等人， 2017]）。还有一些方法会学习视图中共同的低维结构和视图依赖的低维结构（例如，JIVE [Lock等人， 2013]）。随后，这些低维表示会与结果变量关联，用于下游任务；这些任务可以是有监督的，也可以是无监督的。我们将这类方法称为两步方法。

也有一些方法被提出，用于同时关联多个视图并预测结果变量（例如，SIDA [Safo 等人， 2021], BIP [Chekouo 和 Safo， 2020], sJIVE [Palzer等人， 2022], Deep IDA [Wang 和 Safo， 2021] 以及 MOMA[Moon 和 Lee， 2022]）。我们将这些方法称为一步法。与两步法不同，这些方法将关联视图和预测结果变量的问题结合起来。通过这种方式，结果变量可用于指导数据中低维表示的学习，从而可能得到临床意义更强的发现。本文将开发用于联合关联来自多个来源的数据并预测结果变量的方法。

现有大多数用于联合关联多视图并预测结果变量的方法，都着重于学习视图之间以及视图与结果变量之间的线性关系。然而，多视图与结果变量之间的内在关系过于复杂，无法仅通过线性方法来理解。另一方面，现有的有限非线性方法要么缺乏提供可解释结果的能力，要么仅适用于分类结果，要么不利用先验生物学信息，或者在处理大规模数据时计算成本过高 [Hu et al., 2019, Wang and Safo，2021]。近期，有研究提出了一种用于多视图学习的数据整合与分类方法（MOMA），该方法采用注意力机制以提高可解释性 [Moon 和 Lee，2022]。MOMA 为每个视图创建一个模块（例如基因集），并利用注意力机制识别与分类任务相关的模块和变量。值得注意的是，MOMA 中模块的检测由数据驱动。上述方法可用于分类结果，但不适用于连续结果。这些方法不支持使用先验生物学知识，例如变量组信息。此外，Deep IDA 中用于变量排序的重采样方法在处理大型数据时往往计算成本高昂。

1.3 我们的方法

我们为整合分析带来了四项主要贡献。首先，我们开发了一种可扩展的核方法，用于学习多视图数据中的视图无关非线性低维表示，以最大化视图间的关联并预测结果变量。为降低计算开销，我们使用傅里叶随机特征近似每个视图输入数据的核函数计算 [Rahimi 和 Recht, 2008, Lopez-Paz等人，2014]。该方法有助于揭示各视图中的非线性模式，并可扩展至较大的训练样本量。其次，我们利用结果变量来指导对相关变量的学习，以构建视图无关表示。第三，为提高可解释性，我们通过随机化非线性映射，学习驱动数据中共享低维非线性结构的变量。我们将该方法扩展到变量的先验生物学信息（即变量组）可用的情形，从而有望获得更具可解释性的发现。据我们所知，这是首批基于非线性方法的

整合分析与预测方法之一。第四，我们为所提出的方法提供了高效的 PyTorch 实现，并将其与 R 接口，以扩大这些方法的应用范围。真实数据应用和小样本量模拟的结果表明，所提出的方法可用于解决小样本量问题。

本文其余部分安排如下。第 2 节介绍我们提出的方法。第 3 节介绍针对所提方法的优化流程和算法。第 4 节通过大量模拟评估该方法的预测准确性和变量选择准确性。第 5 节将该方法应用于用于研究的问题数据，以识别与 COVID-19 严重程度和状态相关的分子特征。最后，我们在第 6 节进行简要讨论。

2 方法

对于可用数据，我们采用以下记号。对于一个临床结果变量，我们将受试者 $i = 1, \ldots, n$ 的该结果记为 y_i。对于所有受试者，我们用$\mathbf{y}$ 表示结果变量。该变量可以是连续型（例如 COVID-19 严重程度），也可以是二元型（例如 COVID-19 状态）。如果受试者 i 有多个（或 q）连续型临床结果变量，我们将其收集为一个长度为 q 的向量 $\mathbf{y}_i$。对于所有受试者，我们将这些向量收集为一个矩阵 $\mathbf{Y} \in \Re^{n\times q}$。假设从 $d = 1, \ldots, D, D \geq 2$ 个不同来源获得了分子（或组学）和/或表型数据，并且每个视图都排列为一个 $n \times p^d$ 矩阵 $\mathbf{X}^d$，其中上标 d 表示第 d 个来源。例如，对于同一组 n 个个体，矩阵 $\mathbf{X}^1$ 包含代谢组学数据，$\mathbf{X}^2$ 包含基因表达水平，$\mathbf{X}^3$ 包含蛋白质水平，而$\mathbf{X}^4$ 包含表型数据。表型数据可能包含混合数据类型：连续型和分类变量。我们使用指示矩阵表示分类变量 [Gifi, 1990]。我们的目标有三项。首先，我们希望通过联合或共享的低维非线性嵌入，对这些不同数据类型之间复杂的非线性关系进行建模。其次，我们希望识别可能对视图之间这些非线性关系有所贡献的单个变量或变量组。第三，我们希望利用临床结果变量来引导共享低维非线性表示的提取。对视图之间的关系进行联合建模并预测临床结果变量，可使共享的非线性表示以及提取出的变量或变量组具有预测能力，并提高可解释性。我们强调，这三项目标将同时实现。

2.1 关联多视图数据

许多现有的多视图数据关联方法假设存在一个共用（与视图无关）的低维嵌入，例如 $\mathrm{G} \in \Re^{n\times r}$，它驱动各视图之间的关系。每个视图都表示为该共享嵌入的线性函数，即 $\mathrm{X}^d \approx \mathrm{G}\mathrm{B}^{d\mathrm{T}}$。这里，$\mathrm{B}^d$ 是一个 p^d 乘以 r 的载荷矩阵，用于视图 d，其每一行对应于特定变量的 r 个系数。通常，$\mathrm{B}^{d\mathrm{T}}\mathrm{B}^d = \mathrm{I}_{r\times r}$。因此，可以写成 $\mathrm{X}^d\mathrm{B}^d \approx \mathrm{G}$。令 $\mathrm{g}_i \in \Re^r$、$i = 1$、.....、n 表示 G 中的第 i行。上述近似假设存在一个低维空间$\Re^r$，使得受试者 i 在该空间中表示为 $\mathrm{g}_i \in \Re^r$。对于每个 $d = 1$、.....、D，B^d 将受试者 i 的低维表示 g_i 映射为该受试者的观测值 x_i^d。这些映射被限制为线性映射。我们使用非线性函数重新构建这些映射，以捕捉各视图中的复杂非线性关系。

我们假设，对于受试者 i，共享低维表示g_i 是 x_i^d的非线性函数，即$\mathrm{g}_i \approx \mathrm{f}^d(\mathrm{x}_i^d)$ $d = 1, .., D, \ i = 1, ..., n$，以及 $\mathrm{f}^d = (f_1^d, \ldots, f_r^d)$。我们的任务之一是学习 p^d–维输入空间 $\Re^{p^d}$ 与 r-维输出空间 $\Re^r$之间未知的泛函关系。我们假设函数f^d 属于向量值函数的再生核希尔伯特空间（RKHS）。令 $\mathcal{H}(\mathrm{K}^d)$ 为函数f^d : $\Re^{p^d} \to \Re^r$ 的向量值再生核希尔伯特空间，该空间由半正定再生核矩阵 K^d定义。因此，K^d 接收两个输入x^d、$\tilde{}\mathrm{x}^d \in \Re^{p^d}$，并生成一个矩阵 $\Re^{r\times r}$。对于函数f^d: $\Re^{p^d} \to \Re^r$ 的向量值再生核希尔伯特空间，我们有：对于每个$\mathrm{a}^d \in \Re^r$和 $\mathrm{x} \in \Re^{p^d}$，在 $\tilde{}\mathrm{x} \in \Re^{p^d}$处计算得到的核与 a^d 的乘积属于 $\mathcal{H}$，即$\mathrm{K}^d(\mathrm{x}\ \tilde{}\mathrm{x})\mathrm{a} \in \mathcal{H}$。关于向量值再生核希尔伯特空间的综述，请参阅Alvarez 等人 [2011] 及其中的参考文献。根据 K^d的再生性质，对于每个$\mathrm{a}^d \in \Re^r$ 和$\mathrm{x}^d \in \Re^{p^d}$，$\langle \mathrm{f}^d, \mathrm{K}^d(\cdot, \mathrm{x}^d)\mathrm{a}^d\rangle_{\mathbf{K}} = \mathrm{f}^{d\mathrm{T}}\mathrm{a}^d$，其中 $\langle\cdot\rangle_{\mathbf{K}}$ 是 $\mathcal{H}$ 中的内积。

对于固定的$\mathbf{G}$，我们遵循向量值输出的核学习标准理论（即$\mathbf{g}_i \in \Re^r$），定义一种用于关联多视图的正则化方法（例如 Alvarez 等人 [2011], Baldassarre 等人 [2012]）。具体而言，我们最小化正则化经验误差:

$$(\widehat{\mathbf{f}}^1, \cdots, \widehat{\mathbf{f}}^D) = \min_{\mathbf{f}^1\in\mathcal{H}(\mathbf{K}^1),\ldots,\mathbf{f}^d\in\mathcal{H}(\mathbf{K}^D)} \sum_{d=1}^{D}\sum_{j=1}^{r}\frac{1}{2n}\sum_{i=1}^{n}(g_{i,j} - f_j^d(\mathbf{x}_i^d))^2 + \frac{\lambda^d}{2}\sum_{d=1}^{D}\|\mathbf{f}^d\|^2_{\mathcal{H}(\mathbf{K}^d)}. \tag{1}$$

其中，$g_{i,\ j}$ 是$\mathbf{G}$的第 ij个元素，$\mathbf{x}_i^d$是$\mathbf{X}^{n\times p^d}$ 的第 i行，而 $\|\mathbf{f}^d\|_{\mathcal{H}(\mathbf{K^d})}$ 是由再生核希尔伯特空间（RKHS）中的内积所诱导的范数。方程（1）中的第一项是在预测 $f_j^d(\mathbf{x}_i^d)$而非$g_{i,\ j}$时的经验损失； $\lambda^d > 0$ 是一个正则化参数，用于控制模型拟合度与复杂度之间的权衡。最小化（1）等价于最小化 D 个独立方程：

$$\widehat{\mathbf{f}}^d = \min_{\mathbf{f}^d\in\mathcal{H}(\mathbf{K}^d)} \sum_{j=1}^{r}\frac{1}{2n}\sum_{i=1}^{n}(g_{i,j}-f_j^d(\mathbf{x}_i^d))^2 + \frac{\lambda^d}{2}\|\mathbf{f}^d\|^2_{\mathcal{H}(\mathbf{K}^d)}, d=1,\ldots,D. \tag{2}$$

根据 RKHS 的表示定理 [Alvarez 等人，2011，Kimeldorf 和 Wahba， 1970], 解 $\widehat{\ }\mathbf{f}^d$ 在任意点$\tilde{\ }\mathbf{x}^d \in \Re^{p^d}$ 都可以表示为核矩阵$\mathbf{K}^d$ 在 n 数据点上的有限线性组合：

$$\widehat{\mathbf{f}}^d(\widetilde{\mathbf{x}}^d) = \sum_{i=1}^{n}\mathbf{K}^d(\mathbf{x}_i^d,\widetilde{\mathbf{x}}^d)\boldsymbol{\alpha}_i^d,\ \ \boldsymbol{\alpha}_i^d\in\Re^r, \tag{3}$$

其中我们注意到 $\mathbf{K}^d(\mathbf{x}_i^d,\tilde{\ }\mathbf{x}^d)$ 是一个 $r\times r$ 矩阵，作用于 r长向量 $\boldsymbol{\alpha}_i^d$。参数 $\boldsymbol{\alpha}^d, d=1,..,D$ 对于所有受试者的组合满足线性系统

$$\widetilde{\boldsymbol{\alpha}}^d = (\mathbf{K}^d(\mathbf{X}^d,\mathbf{X}^d)+\lambda^d\mathbf{I}_{nr})^{-1}\widetilde{\mathbf{g}}, \tag{4}$$

其中 $\widetilde{\boldsymbol{\alpha}}^d$ 是长度为nr 的向量。值得注意的是， $\tilde{\ }\mathbf{g}$ 也是一个长度为nr 的向量，它通过连接$\mathbf{G}$中的各列得到；$\mathbf{K}^d(\mathbf{X}^d,\mathbf{X}^d)$是一个由大小为 $n\times n$的块组成的 $nr\times nr$ 块矩阵。

注意，为了获得 $\widetilde{\boldsymbol{\alpha}}^d$，我们必须求逆 $nr\times nr$ 矩阵 $\mathrm{K}^d(\mathrm{X}^d,\mathrm{X}^d)$，对于较大的样本而言，这在计算上可能代价高昂。为便于计算，我们采取如下做法。我们假设 G 输出矩阵中的 r 分量彼此独立，因此不相关。该假设是合理的，并使我们能够获得 G 中每个分量的唯一信息。我们还假设核 K^d 是可分离的。也就是说，K^d 可以写成两个核函数 $\mathcal{K}^d\otimes\mathcal{B}$的克罗内克积。其中， $\mathcal{B}$ 是一个 $r\times r$ 对称半正定矩阵，用于捕捉 G 中各分量之间的依赖关系，而 $\mathcal{K}^d$ 是输入空间的一个 $n\times n$ 核矩阵。因此，核矩阵 K^d 是仅关于输入空间的函数（即 $\mathcal{K}^d$）与输出空间中变量之间的相互作用（即 $\mathcal{B}$）的乘积；在本例中，是 G 中 r 分量之间的相互作用。假设各分量相互独立，我们可以将方程（3）中的 $\mathrm{K}^d(\mathrm{x}_i^d,\tilde{\ }\mathrm{x}^d)$写为 $\mathcal{K}^d(\mathrm{x}_i^d,\tilde{\ }\mathrm{x}^d)\mathrm{I}$ [Alvarez 等人，其中 $\mathcal{K}^d(\mathrm{x}_i^d,\tilde{\ }\mathrm{x}^d)$是标量核，而 $\mathcal{B}=\mathrm{I}$ 是一个 $r\times r$ 单位矩阵。因此，核矩阵 $\mathrm{K}^d(\mathrm{X}^d,\mathrm{X}^d)$变为块对角矩阵 [Alvarez 等人,2011]。由于我们使用相同的输入数据来估计第 j 个和第 j' 个分量，因此所有块都相等。因此，对 $nr\times nr$ 矩阵求逆可简化为对 $n\times n$ 矩阵 $\mathcal{K}^d(\mathrm{X},\mathrm{X})$ 求逆。

借助这些概念，方程（2）中的泛函可以写为 $\sum_{j=1}^r\|f_j^d\|^2_{\mathcal{H}(\mathcal{K}^d)}$，而最小化问题（2）可归结为对 r 独立泛函的最小化：

$$(\widehat{f}_1^d,\ldots,\widehat{f}_r^d) = \min_{f_1^d\in\mathcal{H}(\mathcal{K}^d),\cdots,f_r^d\in\mathcal{H}(\mathcal{K}^d)} \sum_{j=1}^{r}\frac{1}{2n}\sum_{i=1}^{n}(g_{i,j}-f_j^d(\mathbf{x}_i^d))^2+\frac{\lambda^d}{2}\sum_{j=1}^{r}\|f_j^d\|^2_{\mathcal{H}(\mathcal{K}^d)}, d=1,\ldots,D. \tag{5}$$

相应地，问题（5）的公式（3）和（4）变为

$$\widehat{f}_j^{\ d}(\widetilde{\mathbf{x}}^d) = \sum_{i=1}^{n}\mathcal{K}^d(\mathbf{x}_i^d,\widetilde{\mathbf{x}}^d)\alpha_{i_j}^d,\ \ \widetilde{\boldsymbol{\alpha}}_j^d = (\mathcal{K}^d(\mathbf{X}^d,\mathbf{X}^d)+\lambda^d\mathbf{I}_n)^{-1}\widetilde{\mathbf{g}}_j \quad \text{for } j=1,\ldots,r, \tag{6}$$

其中， $\widetilde{\boldsymbol{\alpha}}_j^d$ 和 $\tilde{\ }\mathbf{g}_j$ 都是长度为 n的向量，而 $\tilde{\ }\mathbf{g}_j$ 是 $\mathbf{G}$中的第 j列。这里， $\mathcal{K}^d(\mathbf{X}^d,\mathbf{X}^d)$ 是一个 $n\times n$ 矩阵，其第 ij个元素为 $\mathcal{K}^d(\mathbf{x}_i^d,\mathbf{x}_j^d)$。

2.1.1 随机傅里叶变换

对于较大的 n，求逆 $n\times n$ 矩阵 $\mathcal{K}^d(\mathrm{X}^d,\mathrm{X}^d)$ 仍然可能代价高昂。为进一步减少计算量，我们遵循 Rahimi 和 Recht [2008]的思路，使用随机傅里叶变换构造数据的低维近似。具体而言，我们通过随机特征映射将数据映射到低维欧几里得内积空间中，$\mathrm{z}:\ \Re^{p^d}\to\Re^M$，使得该空间中两个变换后点的内积能够很好地近似相应的核函数计算： $\mathcal{K}^d(\mathrm{x}^d,\tilde{\ }\mathrm{x}^d)\approx\langle\,\mathrm{z}(\mathrm{x}^d),\mathrm{z}(\tilde{\ }\mathrm{x}^d)\rangle_{\Re^M}=\mathrm{z}(\mathrm{x}^d)^{\mathrm{T}}\mathrm{z}(\tilde{\ }\mathrm{x}^d)$ [Rahimi 和 Recht， 2008]。使用非线性特征 $\mathrm{z}(\mathrm{x}^d)$，方程 (5) 中最小化问题的解（由方程 (6) 给出）可近似为：

$$\widehat{f}_j^d(\widetilde{\mathbf{x}}^d)\approx\langle\,\mathbf{z}(\mathbf{x}^d),\boldsymbol{\beta}_j^d\rangle_{\Re^M},\ \ (\mathbf{Z}^{d^{\mathrm{T}}}\mathbf{Z}^d+\lambda^d\mathbf{I}_M)\boldsymbol{\beta}_j^d=\mathbf{Z}^{d^{\mathrm{T}}}\mathbf{g}_j,\forall j, \tag{7}$$

其中$\mathbf{Z}^d$是一个 $n \times M$矩阵，而 $\boldsymbol{\beta}_j^d$是一个长度为M的向量。令 $\mathbf{A}^d = (\boldsymbol{\beta}_1^d, \ldots, \boldsymbol{\beta}_r^d) \in \Re^{M \times r}$ 表示视图 d的所有 r系数 $\boldsymbol{\beta}_j^d, j = 1, \ldots, r$的集合。然后，$\mathbf{A} = (\mathbf{Z}^{d^{\mathrm{T}}}\mathbf{Z}^d + \lambda^d \mathbf{I}_M)\mathbf{Z}^{d^{\mathrm{T}}}\mathbf{G}$。该解要求我们求一个$M \times M$矩阵的逆，而不是 $n \times n$，因为 $M \ll n$，所以计算效率较高。

为了构造近似效果良好的特征映射，作者 Rahimi 和 Recht [2008] 运用了 Bochner 定理 [Rudin,2017], 该定理指出，平移不变核矩阵 $\mathcal{K}(\mathrm{X}, \mathrm{X}) \in \Re^{n \times n}$ （此处省略上标 d）的每个元素都可以近似表示为：$\mathcal{K}(\mathrm{x}_l - \mathrm{x}_j) = \int p(\mathrm{w}) e^{i\mathrm{w}^{\mathrm{T}}(\mathrm{x}_l - \mathrm{x}_j)} d(\mathrm{w}) \approx \frac{1}{M}\sum_{m=1}^{M} \langle \sqrt{2}\cos(\mathrm{w}_m^{\mathrm{T}}\mathrm{x}_l + b_m)\ \sqrt{2}\cos(\mathrm{w}_m^{\mathrm{T}}\mathrm{x}_j + b_m) \rangle = \langle \mathrm{z}(\mathrm{x}_l), \mathrm{z}(\mathrm{x}_j) \rangle$. 这里，$\mathrm{z}(\mathrm{x}_l) = [\sqrt{2}\cos(\mathrm{w}_1^{\mathrm{T}}\mathrm{x}_l + b_1) \cdots, \sqrt{2}\cos(\mathrm{w}_M^{\mathrm{T}}\mathrm{x}_l + b_M)] \in \Re^M$, $p(\mathrm{w})$ 是平移不变核 $\mathcal{K}(\cdot)$ 的逆傅里叶变换，$d(\mathrm{w})$ 是 w 的概率测度；$w_1, \ldots, w_M$从 $p(\mathrm{w})$ 中采样得到，且 $b_m \sim Unif(0, 2\pi)$。本质上，为了得到 $\mathrm{z}(\mathrm{x}^d)$ 的每个分量，数据 x^d被投影到从 $p(\mathrm{w})$ 中抽取的随机方向 w^d 上，然后旋转随机角度 b^d，并将所得标量输入正弦函数。对于高斯核 $\mathcal{K}(\mathrm{x}_l, \mathrm{x}_j) = \exp(-\frac{\|\mathrm{x}_l - \mathrm{x}_j\|_2^2}{2\nu^2})$，其逆傅里叶变换 $p(\mathrm{w})$ 是一个高斯分布，即 $\mathrm{w}_m^d \sim N(0, \frac{1}{\nu^2}\mathrm{I})$ [Yang et al., 2012]。

—

2.1.2 多视图数据非线性关联的变量选择与优化问题

我们的一个主要目标是识别$\mathbf{X}^d$中的重要变量，这些变量构成视图的低维结构，并影响结果变量的变化。这里的任务是确保$\mathbf{w}^d, d = 1, \ldots, D$的所有 M 个随机样本具有相同的稀疏结构。我们采用 Gregorová 等人 [2018] 提出的重新参数化方法，将$\mathbf{w}^d$改写为$\mathbf{w}^d = \epsilon^d \cdot \boldsymbol{\gamma}^d$，其中 ϵ^d从 $p(\mathbf{w})$中抽取，$\cdot$为逐元素乘积。因此，我们使用 $\boldsymbol{\gamma}^d$对每个特征进行缩放，并考虑随机特征映射：

$$\mathbf{z}(\mathbf{x}^d) = [\sqrt{2}\cos(\epsilon_1^{d\mathrm{T}}(\boldsymbol{\gamma}^d \cdot \mathbf{x}^d) + b_1^d), \sqrt{2}\cos(\epsilon_2^{d\mathrm{T}}(\boldsymbol{\gamma}^d \cdot \mathbf{x}^d) + b_2^d), \ldots, \sqrt{2}\cos(\epsilon_M^{d\mathrm{T}}(\boldsymbol{\gamma}^d \cdot \mathbf{x}^d) + b_M^d)].$$

因此，视图 d 中$\mathbf{w}^d$ （从而也包括学习到的模型）的稀疏性，是通过 $\boldsymbol{\gamma}^d$的稀疏性实现的。与特征缩放方法类似，我们提出一种多步交替过程，在$\mathbf{G}$固定的情况下学习模型参数$\mathbf{A}$ 和 γ^d。对于固定的γ^d 和$\mathbf{G}$，我们为每个视图的所有输入数据点生成随机特征，并求解线性问题（7），以获得 $M \times r$ 矩阵$\mathbf{A}$。固定该矩阵后，我们求解特征缩放的 p^d维向量 γ^d。由于我们不知道$\mathbf{G}$，因此需要对其进行估计。我们建立以下优化问题，以学习共同的低维非线性表示$\mathbf{G}$、系数矩阵$\mathbf{A}^d$ 以及稀疏权重 γ^d:

$$\min_{\mathbf{G}, \boldsymbol{\gamma}^1, \ldots \boldsymbol{\gamma}^D, \mathbf{A}^1, \cdots, \mathbf{A}^D} \left\{ \frac{1}{2n}\sum_{d=1}^{D} \|\mathbf{G} - \mathbf{Z}^d\mathbf{A}^d\|_F^2 + \frac{\lambda^d}{2}\sum_{d=1}^{D} \|\mathbf{A}^d\|_F^2 + \sum_{d=1}^{D} \mathcal{P}(\boldsymbol{\gamma}^d) \right\} \quad \text{subject to } \mathbf{G}^{\mathrm{T}}\mathbf{G} = \mathbf{I}_r, \tag{8}$$

其中显然 $\mathbf{Z}^d$ 取决于 $\boldsymbol{\gamma}^d$。我们考虑每个视图中稀疏性的两种选择：个体变量稀疏性和组稀疏性。对于个体变量稀疏性（即允许个体变量的系数为零），我们限制 $\boldsymbol{\gamma}^d$ 位于概率单纯形（Δ）中，即 $\boldsymbol{\gamma}^d \in \Delta$。如果 $0 \leq \gamma_s^d \leq 1, s = 1, \ldots, p^d$，并且 $\sum_{s=1}^{p^d} \gamma_s^d = 1$，则 $\boldsymbol{\gamma}^d$ 位于概率单纯形中。对于组稀疏性，我们考虑非重叠稀疏组惩罚：$\mathcal{P}(\boldsymbol{\gamma}^d) = \eta\rho\|\boldsymbol{\gamma}^d\|_1 + (1-\eta)\rho\sum_{l=1}^{G^d} \|\sqrt{p_l^d}\gamma_l^d\|_2$，其中 $\|\cdot\|_1$ 是 l_1 范数，$\|\cdot\|_2$ 是 l_2 范数，G^d 是视图 d 中已知的先验非重叠组数量，p_l^d是组 l中的变量数量, $l = 1, \ldots, G^d$，$\eta \in [0, 1]$ 结合了套索回归和组套索回归惩罚，而 $\rho > 0$ 是一个稀疏性参数，用于控制视图 d的稀疏程度，在 η固定时。该惩罚既可以选择组，也可以选择组内的个体变量，这些变量驱动视图中的底层共同低维结构以及结果变量中的变化。对于固定的 ρ，参数 η 平衡了整体组的变量选择（即至少有一个非零系数的组的数量）和组内变量选择（即非零组内非零系数的数量）。η 的较小取值会促进组选择（即选择更多非零组）以及组内的个体变量选择（即组内往往有更多变量具有非零系数），而 η 的较大取值会抑制组选择，并促进组内的稀疏性（即组内有更多零系数）。

2.2 监督式非线性关联研究

在本节中，我们将研究置于同时关联来自多个视图的数据并预测结果变量的方法背景下。我们的主要目标之一是利用结果变量来指导共同低维非线性表示的构建，以及检测影响各视图潜在结构的变量。为此，我们假设结果变量 $\mathbf{y}$，或对于多个连续结果（或多类别问题）为 $\mathbf{Y}$，通过共同因子$\mathbf{G}$依赖于第 d个视图$\mathbf{X}^d$。这使得整合分析能够与

同时预测临床结局。具体而言，通过最小化损失函数 $\mathcal{L}(\mathbf{Y}, \mathbf{G}\ \boldsymbol{\Theta}\ \text{or}\ \boldsymbol{\theta})$，使$\mathbf{G}$ 与结果变量相关，其中 $\boldsymbol{\theta}$ 或 $\boldsymbol{\Theta}$ 分别表示共享因子对$\mathbf{y}$ 或$\mathbf{Y}$的影响。我们构建以下优化问题，用于刻画多视图的非线性关联、预测一个或多个结果变量，并筛选相关的单个变量或变量组：

$$\min_{\mathbf{G},\boldsymbol{\Theta},\ldots,\boldsymbol{\gamma}^1,\ldots\boldsymbol{\gamma}^D,\mathbf{A}^1,\cdots,\mathbf{A}^D} \left\{ \mathcal{L}(\mathbf{Y},\mathbf{G},\boldsymbol{\Theta}) + \frac{1}{2n}\sum_{d=1}^{D}\|\mathbf{G}-\mathbf{Z}^d\mathbf{A}^d\|_F^2 + \frac{\lambda^d}{2}\sum_{d=1}^{D}\|\mathbf{A}^d\|_F^2 + \sum_{d=1}^{D}\mathcal{P}(\boldsymbol{\gamma}^d) \right\} \quad \text{使得} \mathbf{G}^{\mathrm{T}}\mathbf{G} = \mathbf{I}_r. \tag{9}$$

对于连续型结果，$\mathcal{L}(\mathbf{y},\mathbf{G}\ \boldsymbol{\theta}) = \frac{1}{2n}\|\mathbf{y}-\mathbf{G}\boldsymbol{\theta}\|_F^2$且$\|\cdot\|_F$是弗罗贝尼乌斯范数。我们假设结果变量数据已中心化，因此不需要截距项。对于多个连续结果，我们令 $\mathcal{L}(\mathbf{Y}$ $\frac{1}{2n}\|$ $-$ $\boldsymbol{\Theta}\|_F^2$ $\boldsymbol{\Theta}\in\Re^{r\times q}$, $=$ $\mathbf{G}\ \boldsymbol{\Theta})\ \mathbf{Y}\,\mathbf{G}$，其中。

对于类别数为 $K \geq 2$的分类结局，我们考虑线性判别分析（LDA）问题。具体而言，我们采用 Hastie 等人 [1994] 提出的最优评分方法，该方法给出了 LDA 分类规则。最优评分方法通过使用一系列评分将分类变量转换为连续型变量，从而将分类问题转换为回归问题。令 W 为一个 $n \times K$ 指示矩阵：如果样本 i 属于类别 K，则其第 i行第 K列的元素为 $\mathbf{1}^{\sim}$，否则为 0。令 B 为最优评分的 $K \times K-1$ 矩阵，并令 $\bar{\ }\mathrm{Y} = \mathrm{WB} \in \Re^{n\times(K-1)}$ 为转换后的多类别响应。最优评分方法最小化目标函数 $\mathcal{L}(\bar{\ }\mathrm{Y}, \mathrm{G}\ \boldsymbol{\Theta}) = \frac{1}{2n}\|\bar{\ }\mathrm{Y} - \mathrm{G}\boldsymbol{\Theta}\|_F^2$，其约束条件为 $\bar{\ }\mathrm{Y}^{\mathrm{T}}\bar{\ }\mathrm{Y} = \mathrm{B}^{\mathrm{T}}\mathrm{W}^{\mathrm{T}}\mathrm{WB} = n\mathrm{I}_{K-1}$ and $\bar{\ }\mathrm{Y}^{\mathrm{T}}\mathrm{W}\mathbf{1} = \mathrm{B}^{\mathrm{T}}\mathrm{W}^{\mathrm{T}}\mathrm{W}\mathbf{1} = \mathbf{0}$，其中$\mathbf{1} \in \Re^K$ 为全一矩阵向量，$\mathbf{0} \in \Re^K$ 为零矩阵向量。这里，$\boldsymbol{\Theta}$ 是一个$r \times K-1$ 系数矩阵。为避免任何歧义，我们将明确说明 $\boldsymbol{\Theta}$ 是分类结局的系数矩阵，还是多个连续型结果的系数矩阵。令 $\mathrm{B}_l \in \Re^K, l = 1,\ldots,K-1$ 为 B 中的第 l列，令 n_k 为类别 k 的样本数，且 $s_k = \sum_{i=1}^{k} n_{i\circ}$ 然后，B_l 可根据样本量 n_k [Gaynanova, 2020] 定义为：$\mathrm{B}_l = \left(\left\{(nn_{l+1})^{1/2}(s_l s_{l+1})^{-1/2}\right\}_l, -(ns_l)^{1/2}(n_{l+1}s_{l+1})^{-1/2}, \mathbf{0}_{K-1-l}\right)^{\mathrm{T}}$。

2.3 预测结果

本节介绍我们如何针对给定的目标数据 $\mathbf{X}_{target}^d$预测结果变量。利用训练数据，我们求解方程 9 中的优化问题，并使用估计值 $\widehat{\boldsymbol{\gamma}}^d$, $\widehat{\ }\mathbf{A}^d, d = 1,\ldots,D$ 和 $\widehat{\boldsymbol{\Theta}}$（或 $\widehat{\boldsymbol{\theta}}$），以及目标数据来预测结果变量。首先，我们为受试者 i 构造目标非线性特征，具体如下

$$\mathbf{z}(\mathbf{x}_{target}^d)_{target} = [\sqrt{2}\cos(\epsilon_1^{d\mathrm{T}}(\widehat{\boldsymbol{\gamma}}^d \cdot \mathbf{x}_{target}^d) + b_1^d), \ldots, \sqrt{2}\cos(\epsilon_M^{d\mathrm{T}}(\widehat{\boldsymbol{\gamma}}^d \cdot \mathbf{x}^d) + b_M^d)],$$

其中，$\widehat{\boldsymbol{\gamma}}^d$ 是视图 d 的学习所得稀疏模式，$\mathbf{x}_{target}^d$ 是受试者 i的长度为 p^d 的目标数据。请注意，我们使用与训练数据相同的 ϵ_j^d和 $b_j^d, j = 1,\ldots,M$样本。令 $\mathrm{Z}_{target} \in \Re^{n_{target}\times M}$为一个矩阵，其行是 $\mathrm{z}(\mathrm{x}_{target}^d)$。给定 Z^{target}，我们求解以下优化问题，以获得目标共享低维表示：$\widehat{\ }\mathrm{G}^{target} = \min_{\mathbf{G}:\mathbf{G}^{\mathrm{T}}\mathbf{G}=\mathrm{I}_r} \frac{1}{2n}\sum_{d=1}^{D}\|\mathrm{G} - \mathrm{Z}_{target}^{d}\widehat{\ }\mathrm{A}^d\|_F^2$。该问题$\widehat{\ }$ — $\widehat{\ }$ $\widehat{\ }$的解很容易求得。现在，给定估计值 G^{target}，对于单个连续型结果，我们预测连续型结果 y^{pred} 为 $\mathrm{y}^{pred} = \mathrm{G}_{target}\widehat{\boldsymbol{\theta}}$ $\widehat{\ }$；对于多个连续结果，则预测 $\mathrm{Y}^{pred} = \mathrm{G}_{target}\boldsymbol{\Theta}$。对于分类结局，$\widehat{\ }$ $\widehat{\ }$ $\widehat{\ }\widehat{\ }$我们获得目标得分 $\mathrm{U}^{target} = \mathrm{G}_{target}\boldsymbol{\Theta}$ 和训练得分 $\mathrm{U} = \mathrm{G}\boldsymbol{\Theta}$。然后，我们使用最近质心算法，将每位受试者分配到目标得分与训练得分中的类别平均值之差最小的类别。

3 算法

(9)的优化通过对 $\mathbf{A}^d$、$\mathbf{G}$、$\boldsymbol{\Theta}$ 或 $\boldsymbol{\theta}$以及 $\boldsymbol{\gamma}^d$ 进行交替最小化来完成。有关我们的初始化、算法和超参数选择的更多详细信息，请参阅补充信息。我们按如下方式估计每个参数。

估计视图特异性载荷，$\mathbf{A}^d$：在迭代 $t+1$时，为估计 $\mathbf{A}^d$，我们将$\mathbf{G}$ 和 $\boldsymbol{\gamma}^d$ 固定为其在迭代 t、$t \geq 0$ 时的值，并求解最小化问题：

$\min_{\mathbf{A}^1,\ \cdots,\ \mathbf{A}^D}\left\{\frac{1}{2n}\sum_{d=1}^{D}\|\mathbf{G} - \mathbf{Z}^d\mathbf{A}^d\|_F^2 + \frac{\lambda^d}{2}\sum_{d=1}^{D}\|\mathbf{A}^d\|_F^2\right\}$。令 $\bar{\ }\mathbf{Z}^d = \left(\frac{1}{\sqrt{n}}\mathbf{Z}^d, \sqrt{\lambda^d}\mathbf{I}_{D\times D}\right)_{(n+D)\times D}$。另令$\bar{\ }\mathbf{G} = \left(\frac{1}{\sqrt{n}}\mathbf{G}\ \mathbf{0}_{D\times r}\right)_{(n+D)\times r}$。该最小化问题可写为：

$\min_{\mathbf{A}^1,\cdots,\mathbf{A}^D}\left\{\frac{1}{2}\sum_{d=1}^{D}\|\tilde{\mathbf{G}}-\tilde{\mathbf{Z}}^d\mathbf{A}^d\|_F^2.\right\}$。这归结为对 D 个相互独立的最小二乘优化问题进行最小化，针对$\mathbf{A}^d, d=1,\ldots,D$，且每个问题的闭式解为 $\hat{\mathbf{A}}^d=(\tilde{\mathbf{Z}}^{d\mathrm{T}}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^{d\mathrm{T}}\tilde{\mathbf{G}}$。

估计共享低维表示 G：当 $\mathbf{A}^d$ 在迭代 $t+1$ 中取其值固定不变，并且 $\boldsymbol{\Theta}$（或 $\boldsymbol{\theta}$）在迭代 t 中取其值固定不变时，我们在迭代 $t+1$ 中求解 G：$\min_{\mathbf{G}:\mathbf{G}^{\mathrm{T}}\mathbf{G}=\mathbf{I}_r}\left\{\mathcal{L}(\mathbf{Y},\mathbf{G},\boldsymbol{\Theta})+\frac{1}{2n}\sum_{d=1}^{D}\|\mathbf{G}-\mathbf{Z}^d\mathbf{A}^d\|_F^2\right\}$。G 的解可以写成闭式形式。具体而言，令 $\tilde{\mathbf{Y}}=\left(\mathbf{Y}\text{或}\bar{\mathbf{Y}},\mathbf{Z}^1\mathbf{A}^{1(t+1)},\cdots,\mathbf{Z}^D\mathbf{A}^{D(t+1)}\right)_{n\times(Dr+q)}$。对于单个连续型结果，$q=1$。需要注意的是，Y 是连续型响应结果变量，而 Y 是转换后的多类别响应。对于 K-类别分类问题，$q=K-1$。令 $\boldsymbol{\Theta}=(\boldsymbol{\Theta},\mathbf{I}_r,\cdots,\mathbf{I}_r)_{r\times(Dr+q)}$。于是，该问题等价于 $\min_{\mathbf{G}:\mathbf{G}^{\mathrm{T}}\mathbf{G}=\mathbf{I}_r}\frac{1}{2n}\|\tilde{\mathbf{Y}}-\mathbf{G}\tilde{\boldsymbol{\Theta}}\|_F^2$，这是一个 Procrustes优化问题。但最小化上述问题等价于求解以下问题：$\max_{\mathbf{G}:\mathbf{G}^{\mathrm{T}}\mathbf{G}=\mathbf{I}_r}tr\left(\frac{1}{n}\tilde{\mathbf{Y}}\tilde{\boldsymbol{\Theta}}^{\mathrm{T}}\mathbf{G}\right)$，其中 $tr(\cdot)$ 是迹函数。上述问题的解为 $\mathbf{UV}^{\mathrm{T}}$，其中 U 和 V 是谱分解 $\frac{1}{n}\tilde{\mathbf{Y}}\tilde{\boldsymbol{\Theta}}^{\mathrm{T}}=\mathbf{UDV}^{\mathrm{T}}$所得的左、右奇异向量。

估计 $\boldsymbol{\Theta}$ 或 $\boldsymbol{\theta}$：对于$\mathbf{G}$ 固定为其在第 $t+1$ 次迭代时的值，我们使用以下优化问题求解 $\boldsymbol{\Theta}$（用于多重响应或多类别问题）或单个结果变量的 $\boldsymbol{\theta}$：$\min_{\boldsymbol{\Theta}\text{或}\boldsymbol{\theta}}\mathcal{L}(\mathbf{Y}[\text{或}\mathbf{y}\text{或}\mathbf{Y}],\mathbf{G},\boldsymbol{\Theta})$。这是一个最小二乘问题，对于连续型结果变量，其解为 $\hat{\boldsymbol{\Theta}}=(\mathbf{G}^{\mathrm{T}}\mathbf{G})^{-1}\mathbf{Y}$；对于分类结局，其解为 $\hat{\boldsymbol{\Theta}}=(\mathbf{G}^{\mathrm{T}}\mathbf{G})^{-1}\bar{\mathbf{Y}}$。

估计 $\boldsymbol{\gamma}^d$：回顾一下，$\mathbf{Z}^d$ 取决于 $\boldsymbol{\gamma}^d$。对于固定为当前值的$\mathbf{G}$ 和$\mathbf{A}^d$，我们针对视图 d求解以下问题：$\min_{\boldsymbol{\gamma}^d}\frac{1}{2n}\|\mathbf{G}-\mathbf{Z}^d\mathbf{A}^d\|_F^2+\mathcal{P}(\gamma^d)$。如第 2.1.2 节所述，我们针对 $\mathcal{P}(\gamma^d)$考虑两种变量选择方案：(i) 个体变量选择，其中我们限制 $\boldsymbol{\gamma}^d$ 位于概率单纯形 (Δ)中，即 $\boldsymbol{\gamma}^d\in\Delta$；以及 (ii) 组变量选择，其中我们使用非重叠稀疏组惩罚：$\mathcal{P}(\boldsymbol{\gamma}^d)=\eta\rho\|\boldsymbol{\gamma}^d\|_1+(1-\eta)\rho\sum_{l=1}^{G^d}\|\sqrt{p_l^d}\gamma_l^d\|_2$。对于任一方案，我们都使用加速梯度下降来求解优化问题。具体而言，我们实现了 [Beck和Teboulle，2009] 中提出的带回溯的快速迭代收缩阈值算法（FISTA）。

4 模拟

我们开展模拟研究，以评估所提出方法的实证表现。我们的主要目标是评估该方法在预测测试结果变量以及根据学习得到的低维表示和稀疏性参数选择变量方面的准确性。我们根据结果变量的类型考虑两种情景。每种情景都包含 $D=2$ 个视图，并模拟视图之间及视图内部存在非线性关系的数据。在第一种情景中，使用低维表示生成取值连续的结果变量（补充信息）。在第二种情景中，生成的数据包含两个类别。在所有情景中，我们生成 20 组蒙特卡洛训练集和测试集。我们使用以下标准评估所提出的方法和现有方法：i）测试集预测/分类准确率；ii）特征选择。对于预测，我们评估所提出方法基于学习得到的共享低维表示，正确预测结果变量的能力（对于连续结果，使用均方误差 [MSE] ），或正确分类的能力（对于二元结果，使用分类准确率）。对于特征选择，我们评估这些方法选择真实信号而非噪声变量的能力。

4.1 二元结果

我们考虑三种不同的设置，每种设置的样本数和变量数各不相同。在每种设置中，20 个变量具有非线性关系（图 1），并被视为信号变量。我们生成的视图中，每个视图都包含两个类别。

- 按如下方式生成类别一的数据：
  –生成 $\boldsymbol{\theta}_1$ 作为一个由 $n_1/2$ 个在 0.6和 2.5 之间均匀间隔的点组成的向量–构造向量 $\mathbf{s}=[(\boldsymbol{\theta}_1-\mathbf{1})^2;(\boldsymbol{\theta}_1+0.1\mathbf{1})^2-2(\boldsymbol{\theta}_1-\mathbf{1})^2]\in\Re^{n_1\times 1}$– 设置 $\tilde{\mathbf{X}}_{11}=[\tilde{\boldsymbol{\theta}}_1;\ \mathbf{s}\mathbf{1}_{p_1-1}^{\mathrm{T}}]\in\Re^{n_1\times p^1}$，$\tilde{\boldsymbol{\theta}}_1=[\boldsymbol{\theta}_1;\ \boldsymbol{\theta}_1]\in\Re^{n_1\times 1}$– 生成 $\mathbf{X}_{11}=\tilde{\mathbf{X}}_{11}\cdot\mathbf{W}+\sigma_{11}\mathbf{E}_{11}$，其中 $(\cdot)$ 是逐元素乘法，$\mathbf{W}\in\Re^{n_1\times p^1}=[\mathbf{1}_{20},\mathbf{0}_{p^1-20}]$ 是一个由全一矩阵和零矩阵组成的矩阵，1是全一矩阵，0 是零矩阵，$\sigma_{11}=0.1$，且 $\mathbf{E}_{11}\sim N(0,1)$。

- 按如下方式生成第二类的数据：
  –生成 $\boldsymbol{\theta}_2$，作为由 0.96 和 1.67 之间 $n_2/2$ 个均匀间隔点组成的向量–构造向量 $\mathbf{s} = [(\boldsymbol{\theta}_2 - \mathbf{1})^2 + 0.25;\ (\boldsymbol{\theta}_2 + 0.1\mathbf{1})^2 - 3.5(\boldsymbol{\theta}_2 - \mathbf{1})^2 + 0.25] \in \Re^{n_2 \times 1}$– 设置 $\tilde{\mathbf{X}}_{12} = [\tilde{\boldsymbol{\theta}}_2;\ \mathbf{s}\mathbf{1}^{\mathrm{T}}_{p^1-1}] \in \Re^{n_2 \times p^1}$，$\tilde{\boldsymbol{\theta}}_2 = [\boldsymbol{\theta}_2;\ \boldsymbol{\theta}_2] \in \Re^{n_2 \times 1}$– 生成$\mathbf{X}_{12} = \tilde{\mathbf{X}}_{12} \cdot \mathbf{W} + \sigma_{12}\mathbf{E}_{12}$，其中(·) 表示逐元素乘法，$\mathbf{W} \in \Re^{n_2 \times p^1} = [\mathbf{1}_{20}, \mathbf{0}_{p^1-20}]$ 是由全一矩阵和零矩阵组成的矩阵，1是全一矩阵，0 是零矩阵，$\sigma_{12} = 0.1$，且 $\mathbf{E}_{12} \sim N(0, 1)$。

- 将两个类别的数据连接起来，形成视图 1的数据，即 $\mathbf{X}^1 = [\mathbf{X}_{11}; \mathbf{X}_{12}] \in \Re^{n \times p^1}$。
- 生成第二视图：$\mathbf{X}^2 = 5\mathbf{X}^1 + \sigma_2\mathbf{E}_2$，其中 $\mathbf{E}_2 \sim N(0, 1)$，且 $\sigma_2 = 0.2$。
- 通过这种方式生成 $\mathbf{X}_1$ 和 $\mathbf{X}_2$，我们假设前20个变量具有非线性关系，并能够区分两个类别。图1展示了视图1 中信号变量（左图）、信号变量和噪声变量（左起第二幅图）、噪声变量（左起第三幅图）之间关系的示意图，以及数据的图像图。

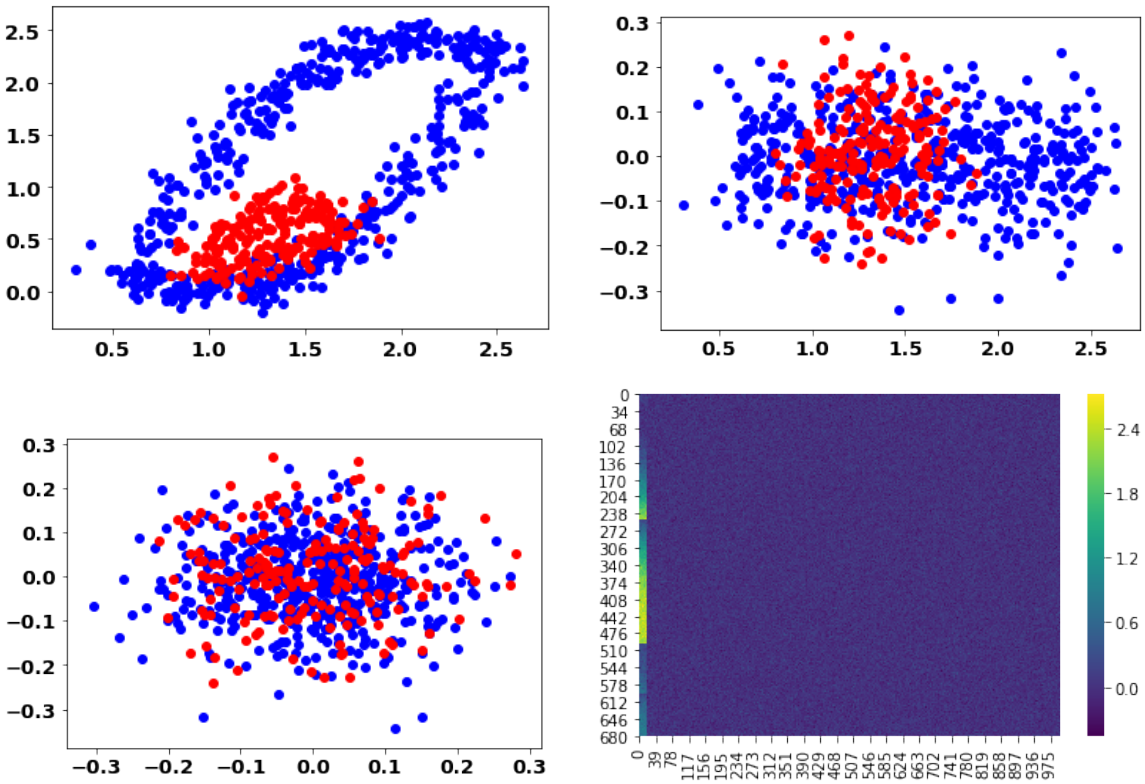


图1：左上：视图 1中信号变量 1和 2 之间的关系；右上：视图 1中信号变量1和噪声变量21之间的关系；左下：噪声变量之间的关系；右下：视图 1的图像图。前二十个变量是信号变量，其余变量是噪声。视图 2 的图与视图 1 类似。

4.2 比较方法

我们将所提出的方法RandMVLearn 和RandMVLearnG（用于选择组变量）与用于关联多视图数据的线性和非线性方法进行比较：稀疏典型相关分析 [sCCA][Safo等人， 2018], 典型变量回归（CVR） [Luo等人， 2016], 针对多视图数据的稀疏整合判别分析（SIDA） [Safo等人， 2021], BIP [Chekouo和 Safo， 2023], 深度典型相关分析（Deep CCA） [Andrew 等， 2013b] 以及 MOMA [Moon 和 Lee， 2022]。由于 sCCA 和 Deep CCA 主要用于

在关联视图时，我们使用这些方法得到的得分构建线性或非线性回归模型，以探究这些方法学习到的低维表示的预测和分类性能。为考察适当整合来自多个来源的数据所带来的优势，将这些整合分析方法与在堆叠视图上针对连续型结果的多层感知机（MLP）和针对二元结果的支持向量机（SVM） [Ben-Hur 等人， 2008] –的朴素实现进行比较。我们将 DeepCCA 与教师-学生框架（TS） [Mirzaei 等人， 2020] 结合，用于变量选择。

我们比较了所提出的两种变量选择方案——个体变量选择和组变量选择的表现。对于组变量选择，变量分为两组：一组是信号变量，另一组是噪声变量。我们固定稀疏组套索惩罚中的 $\eta = 0.5$ 。对于每个视图，我们考虑 ρ 在 (0， ρ_{max}) 范围内的 5 个取值。选择 ρ_{max} 以避免 $\gamma^d_{SGLasso}$ 中出现平凡解。我们对每个视图的 ρ 可能组合进行网格搜索和随机搜索，并通过三折交叉验证选出最优的 ρ 组合。

4.3 结果

我们首先比较二元结果变量下的结果。图 2 展示了所提出的方法和比较方法在预测和变量选择方面的表现，这些结果汇总了各视图，并对 20 次重复实验取了平均值。RandMVLearnG（包括网格搜索和随机搜索）的变量选择表现优于 RandMVLearn，凸显了在可获得组信息时利用组信息的优势。与 CVR 和 MOMA 相比，RandMVLearnG 的真阳性更多、假阳性更少；其变量选择表现则与 sCCA 相当。我们注意到，在所有设置下，所提出的方法，尤其是 RandMVLearnG，分类错误率均低于其他所有方法。对于连续型设置（见《补充信息》中的图 2），RandMVLearn 的均方误差与组变量选择方法（即 RandMVLearnG）相当。此外，我们提出的方法的均方误差与所比较的方法相当（设置 1、2 和 4）或更低（设置 3）。RandMVLearnG 在所有设置下的变量选择表现均与 sCCA 相当，且优于 CVR。就均方误差而言，除设置 3 外，sCCA 后接线性回归在所有设置下均得到更低的均方误差。使用随机搜索调优超参数的 RandMVLearnG，其估计值与使用网格搜索调优超参数的 RandMVLearnG 相当。不过，RandMVLearnG（随机搜索）比 RandMVLearnG（网格搜索）更快。此外，与其他方法相比，RandMVLearn 的计算效率更高。

总体而言，对于二元型和连续型结果变量，无论是否使用先验组信息，预测和变量选择结果均表明，我们的方法能够检测出数据中的真实信号并省略噪声变量。即使在样本量小于变量数量的情况下，所提出的方法也能提供具有竞争力或更优的预测估计值。

5 COVID-19 数据分析

我们应用所提出的方法，整合了动机研究中的蛋白质组学、RNA测序、脂质组学和代谢组学数据。我们的目标是对分子数据与结果变量数据中的非线性关联进行建模，并识别有助于COVID-19严重程度（由HFD-45定义）和状态的分子特征及通路。

5.1 数据预处理及所提方法与竞争方法的应用

在128名患者中，120名同时拥有组学数据和临床数据。我们的主要结果变量是 HFD-45（连续型结果）和 COVID-19 状态（二元结果）。初始数据集包含18，212个基因、517种蛋白质、111项代谢组学特征和3，357项脂质组学特征。我们使用了 Lipman 等人 [2022] 提供的预处理数据。更多详情请参阅 Lipman 等人 [2022]。我们通过两种方式研究所提出的方法。首先，我们在预处理数据上应用不含组信息的方法：脂质组学的 $\mathbf{X}^1 \in \Re^{120\times 1,\ 015}$ 、代谢组学的 $\mathbf{X}^2 \in \Re^{120\times 72}$ 、RNA测序的$\mathbf{X}^3 \in \Re^{120\times 5,\ 800}$ ，以及蛋白质组学的$\mathbf{X}^4 \in \Re^{120\times 264}$ 。在补充信息中，我们假设 RNA测序数据具有组信息，并研究所提出方法的稀疏组实现。对于这两种实现，我们在保持 COVID-19 状态比例与完整数据相近的前提下，将数据随机划分为50个训练集（$n = 71$）和测试集（$n = 49$）。我们使用训练数据拟合模型，并使用测试数据评估错误率——分别针对连续型结果和二元结果计算均方误差和误分类率。我们将随机特征数设为45，使其小于训练集和测试集的样本量。我们采用所提出的简易方法确定共享低维表示的数量，并使用中位数启发式确定核参数。

表 1：预测估计值与变量选择。Deep GCCA + MLP 指先进行 Deep 广义典型相关分析，再使用多层感知机（MLP）；MLP 指在堆叠视图上应用多层感知机；L：脂质组学；M：代谢组学；R：RNA测序；P：蛋白质组学；MSE 指均方误差

方法	预测（标准误）	所选变量数量 (L,M,R,P)
连续型结果变量 (HFD-45)	平均均方误差	
RandMVLearn	0.8872 (0.02)	144.98/1015, 47.8/72, 215.98/5800, 109.66/264
Deep GCCA +多层感知机	0.9446 (0.22)	-
BIP	0.8376 (0.01)	320.92/1015, 14.94/72, 4122.54/5800, 5.28/264
二元结果变量（COVID 状态）	平均值分类误差（%）	
RandMVLearn	9.76（0.74）	170.26/1015, 48.47/72, 129.98/5800, 114.9/264
SIDA	6.86 (3.61)	10.14/1015, 11.8/72, 61.84/5800, 29.52/264
Deep GCCA +支持向量机	18.08 (1.23)	-
MLP	10.82 (4.40)	-

5.1.1 无组信息时的预测估计和所选分子

我们讨论没有先验组信息时的结果。表2列出了 COVID-19 状态的平均测试误分类率，以及 HFD-45 的平均测试均方误差

二元结果：***COVID-19*** 状态：SIDA 的平均测试误差率最低（6.86%），其次是所提出的方法（9.76%），而 Deep GCCA 的平均测试误差率最高（18.08%）。总体而言，与 SIDA 相比，所提出的方法选择的变量更多。所提出的方法平均选出的基因、蛋白质、代谢组学特征和脂质组学特征数量分别为 129.98、170.56、114.9 和 48.72。我们进一步考察了在 50 次重复中被选中超过 25 次（$> 50\%$）的变量。在这些变量中，我们识别出 12 个脂质组学特征、55 个代谢组学特征和 62 个蛋白质。没有基因符合这一标准。为了解蛋白质的功能分类（受篇幅限制，我们聚焦于蛋白质），我们使用 Ingenuity Pathway Analysis（IPA）软件进行了功能富集分析。与 COVID-19 状态相关的蛋白质列表中显著富集的通路见补充信息的表2。许多富集通路与免疫学有关。例如，B细胞发育、IL-15 信号传导、IL-12 信号传导和巨噬细胞中的 IL-12 生成。这些通路调节促成疾病发展的过程或机制。分子 IGHV1-18、IGHV1-24、IGHV1-3、IGHV2-70、IGKV1-17、IGKV2-30、IGLV3-19、IGLV4-69 似乎是这些免疫学通路共有的分子。与免疫功能无关的通路包括 p70S6K信号通路、LXR/RXR 激活通路、动脉粥样硬化信号通路，以及 THOP1 在阿尔茨海默病中的神经保护作用。LXR/RXR 激活通路在脂质代谢、炎症和胆固醇向胆汁酸分解代谢的调节中发挥关键作用。动脉粥样硬化信号通路值得关注，因为动脉粥样硬化研究发现，动脉粥样硬化与 COVID-19 [之间存在相似之处和差异2021]。IPA 中的重叠经典通路（图3）用于通过参与这些通路的共有分子，直观呈现通路间共享的生物学特征。我们的蛋白质列表中过度代表的疾病和生物学功能包括（B-H 校正后的 p 值 ≤ 0.01）：心血管疾病（例如动脉闭塞、动脉粥样硬化）；机体损伤与异常（例如进行性神经系统疾病、低级别前列腺癌）；炎症反应（例如皮肤红斑狼疮）和免疫疾病（例如霍奇金淋巴瘤）。文献证据表明，受某些此类疾病影响的个体往往会出现严重的 COVID-19 结局。

连续型结果：***HFD-45***：当结果变量为无住院天数时，与非线性方法 DeepGCCA相比，所提出的方法具有最低的均方误差。BIP的均方误差最低。所提出的方法选定的基因、蛋白质、代谢组学和脂质组学特征的平均数量分别为 215.98、109.66、47.8 和 144.98。在这些变量中，有 57个代谢组学特征和 54 个蛋白质被选定超过 25 次。没有基因或脂质组学特征达到这一标准。与COVID-19 状态类似，我们使用 IPA 确定蛋白质列表中显著富集的通路和疾病。许多被确定为与 COVID-19状态相关的通路，也被确定为与COVID-19 严重程度相关（补充信息中的表3）。在我们的蛋白质列表中，过度代表的疾病和生物学功能包括（B-H校正后的p值 ≤ 0.01）：炎症反应、代谢性疾病（例如阿尔茨海默病、葡萄糖代谢紊乱）、神经系统疾病（例如阿尔茨海默病、进行性神经系统疾病）以及机体损伤与异常。许多疾病在我们关于COVID-19 严重程度和状态的蛋白质列表中富集，这一事实表明，COVID-19 会扰乱许多生物系统，因此有必要研究该疾病的后遗症效应，以更好地了解其机制并开发有效的治疗方法。

6 讨论

我们开发了可扩展的基于核的非线性方法，用于联合整合来自多个来源的数据并预测临床结局。我们的框架假设所有视图都具有共享的或视图无关的低维表示，并且可以根据视图的非线性函数对其进行估计。我们使用核方法对非线性函数建模，并将这些函数限制在再生核希尔伯特空间中。基于随机傅里叶基可以近似平移不变核函数这一思想，我们构建了每个视图的非线性映射，并利用这些非线性映射和结局变量来学习视图无关的低维表示。因此，所提出的方法能够揭示视图中的非线性模式，并可扩展到较大的训练样本量。在估计这些视图无关的低维表示时，会使用结局变量，从而赋予其可解释性。通过随机化非线性映射，我们能够学习可能驱动视图中潜在共享低维结构的变量。我们的方法允许整合有关变量的先验生物学信息，并提高研究结果的可解释性，使其有别于其他用于多视图学习和预测的非线性方法。我们使用 Python 3（具体而言是 PyTorch）开发了一种用户友好的算法，并将其与 R 接口，以扩大算法的适用范围。通过模拟研究，我们证明，所提出的方法优于其他几种用于多视图数据整合的线性和非线性方法。将所提出的方法应用于与 COVID-19 相关的基因表达、代谢组学、蛋白质组学和脂质组学数据时，我们识别出若干与重症疾病和 COVID-19 状态相关的分子特征。在我们筛选出的、与 COVID-19 严重程度和状态相关的候选特征中，富集的通路包括与炎症和免疫功能相关的通路。

我们提出的方法存在一些局限性。首先，这些方法假设变量组之间不重叠。如果变量组存在重叠，则可能可以重构出互不重叠的变量组。为此，可以考虑每个变量组中的变量间连接，然后将某个变量保留在连接较多的变量组中。这一做法假设：与在另一个变量组中仅具有较少连接的情况相比，在某个变量组中具有较多连接的变量更有可能在该变量组中发挥关键作用。其次，所提出的方法仅对视图之间的共享低维表示进行建模。未来的工作可以考虑同时对共享低维表示和依赖于视图的低维表示进行建模。

总之，我们开发了可扩展的随机化核方法，用于联合学习多视图数据中的非线性关系并预测临床结局。所提出的方法能够识别有望对视图之间的非线性关联及结果变量的变异有所贡献的变量或变量组。尽管存在上述局限性，我们仍认为模拟和真实数据应用的结果令人鼓舞，并相信所提出的方法将推动未来的应用。

经费支持与致谢

本研究获得美国国立卫生研究院下属美国国家普通医学科学研究所资助，资助编号为 1R35GM142695。本文内容由作者全权负责，不代表美国国立卫生研究院的官方观点。

数据可用性声明

所用数据来自 Overmyer 等人 [2021]。我们提供了一个与 R 接口的 Python 代码包，*RandMVLearn*，以便于使用我们的方法。其源代码及 README 文件将发布于：https: //github.com/lasandrall/RandMVLearn。

参考文献

重症 COVID-19 GWAS 研究组。伴呼吸衰竭的重症 COVID-19 全基因组关联研究。《新英格兰医学杂志》，383(16):1522–1534，2020。

Katherine A. Overmyer、Evgenia Shishkova、Ian J. Miller、Joseph Balnis、Matthew N. Bernstein、Trenton M. Peters-Clarke、Jesse G. Meyer、Qiuwen Quan、Laura K. Muehlbauer、Edna A. Trujillo、Yuchen He、Amit Chopra、Hau C. Chieng、Anupama Tiwari、Marc A. Judson、Brett Paulson、Dain R. Brademan、Yunyun Zhu、Lia R. Serrano、Vanessa Linke、Lisa A. Drake、Alejandro P. Adam、Bradford S. Schwartz、Harold A. Singer、Scott Swanson、Deane F. Mosher、Ron Stewart、Joshua J. Coon 和 Ariel Jaitovich。大规模 COVID-19 严重程度多组学分析。《Cell Systems》，12(1):23–40.e7，2022/08/23 2021。doi:10.1016/j.cels.2020.10.003。URL https://doi.org/10.1016/j.cels. 2020.10.003。

H. Hotelling。两组变量之间的关系。《Biometrika》，第 312–377 页，1936 年。Galen Andrew、Raman Arora、Jeff Bilmes 和 Karen Livescu。深度典型相关分析。载于《国际机器学习会议》，第 1247–1255 页。PMLR，2013a。David Lopez-Paz、Suvrit Sra、Alex Smola、Zoubin Ghahramani 和 Bernhard Schölkopf。随机化非线性成分分析。载于《国际机器学习会议》，第 1359–1367 页。PMLR，2014 年。Paul Horst。广义典型相关及其在实验数据中的应用。第 14 号。《Journal of Clinical Psychology》，1961 年。Jon R Kettenring。多组变量的典型分析。《Biometrika》，58(3):433–451，1971 年。Adrian Benton、Huda Khayrallah、Biman Gujral、Dee Ann Reisinger、Sheng Zhang 和 Raman Arora。深度广义典型相关分析。arXiv 预印本 arXiv:1702.02519，2017 年。Eric F Lock、Katherine A Hoadley、James Stephen Marron 和 Andrew B Nobel。联合与个体变异解释（JIVE）：多种数据类型整合分析方法。《The Annals of Applied Statistics》，7(1):523，2013 年。Sandra E. Safo、Eun Jeong Min 和 Lillian Haine。面向多视图结构化数据的稀疏线性判别分析。《Biometrics》，n/a(n/a)，2021 年。doi:https://doi.org/10.1111/biom.13458。URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13458。Thierry Chekouo 和 Sandra E Safo。贝叶斯整合分析与预测及其在动脉粥样硬化性心血管疾病中的应用。arXiv 预印本 arXiv:2005.11586，2020 年。Elise F. Palzer、Christine H. Wendt、Russell P. Bowler、Craig P. Hersh、Sandra E. Safo 和 Eric F. Lock。sJIVE：有监督的联合与个体变异解释。《Computational Statistics and Data Analysis》，175:107547，2022 年。ISSN 0167-9473。doi:https://doi.org/10.1016/j.csda.2022.107547。URL https://www.sciencedirect.com/ science/article/pii/S016794732200127X。Jiuzhou Wang 和 Sandra E Safo。Deep IDA：一种用于多视图数据整合判别分析并进行特征排序的深度学习方法——应用于 COVID-19 严重程度。ArXiv，2021 年。Sehwan Moon 和 Hyunju Lee。MOMA：用于多组学数据解释与分类的多任务注意力学习算法。《Bioinformatics》，38(8):2287–2296，2022 年 2 月。ISSN 1367-4803。doi:10.1093/bioinformatics/btac080。URL https://doi.org/10.1093/bioinformatics/btac080。Peng Hu、Dezhong Peng、Yongsheng Sang 和 Yong Xiang。多视图线性判别分析网络。《IEEE Transactions on Image Processing》，28(11):5352–5365，2019 年。Ali Rahimi 和 Benjamin Recht。用于大规模核机器的随机特征。载于《神经信息处理系统进展》，第 1177–1184 页，2008 年。Albert Gifi。非线性多元分析。Wiley，1990 年。Mauricio A Alvarez、Lorenzo Rosasco 和 Neil D Lawrence。向量值函数的核：综述。arXiv 预印本 arXiv:1106.6251，2011 年。Luca Baldassarre、Lorenzo Rosasco、Annalisa Barla 和 Alessandro Verri。通过谱滤波进行多输出学习。《Machine Learning》，87(3):259–301，2012 年。George S Kimeldorf 和 Grace Wahba。随机过程的贝叶斯估计与样条平滑之间的对应关系。《The Annals of Mathematical Statistics》，41(2):495–502，1970 年。Walter Rudin。群上的傅里叶分析。Courier Dover Publications，2017 年。Tianbao Yang、Yu-Feng Li、Mehrdad Mahdavi、Rong Jin 和 Zhi-Hua Zhou。Nyström 方法与随机傅里叶特征：理论与实证比较。《神经信息处理系统进展》，25，2012 年。Magda Gregorová、Jason Ramapuram、Alexandros Kalousis 和 Stéphane Marchand-Maillet。通过核随机特征进行大规模非线性变量选择。载于《机器学习与知识发现联合欧洲会议》，第 177–192 页。Springer，2018 年。Trevor Hastie、Robert Tibshirani 和 Andreas Buja。通过最优评分进行灵活判别分析。《Journal of the American Statistical Association》，89(428):1255–1270，1994 年。doi:10.1080/01621459.1994.10476866。URL https://doi.org/10.1080/01621459.1994.10476866。Irina Gaynanova。稀疏多类别惩罚最优评分的预测与估计一致性。《Bernoulli》，26 (1):286–322，2020 年。Amir Beck 和 Marc Teboulle。用于线性逆问题的快速迭代收缩阈值算法。《SIAM Journal on Imaging Sciences》，2(1):183–202，2009 年。

Sandra E Safo、Jeongyoun Ahn、Yongho Jeon 和 Sungkyu Jung。用于甲基化与基因表达数据整合分析的稀疏广义特征值问题及其在典型相关分析中的应用。Biometrics，74(4):1362–1371，2018。Chongliang Luo、Jin Liu、Dipak K. Dey 和 Kun Chen。典型变量回归。Biostatistics，17(3):468–483，02 2016。ISSN 1465-4644。doi:10.1093/biostatistics/kxw001。URL https://doi.org/10.1093/biostatistics/ kxw001。Thierry Chekouo 和 Sandra E Safo。贝叶斯整合分析与预测及其在动脉粥样硬化性心血管疾病中的应用。Biostatistics，24(1):124–139，2023。Galen Andrew、Raman Arora、Jeff Bilmes 和 Karen Livescu。深度典型相关分析。Journal of Machine Learning Research: Workshop and Conference Proceedings，2013b。Asa Ben-Hur、Cheng Soon Ong、Sören Sonnenburg、Bernhard Schölkopf 和 Gunnar Rätsch。用于计算生物学的支持向量机与核函数。PLoS computational biology，4(10):e1000173，2008。Ali Mirzaei、Vahid Pourahmadi、Mehran Soltani 和 Hamid Sheikhzadeh。利用教师-学生网络进行深度特征选择。Neurocomputing，383:396–408，2020。Danika Lipman、Sandra E. Safo 和 Thierry Chekouo。多组学分析揭示了与 COVID-19 及 COVID-19 严重程度相关的富集通路。PLOS ONE，17(4):1–30，04 2022。doi:10.1371/journal.pone.0267047。URL https://doi.org/10.1371/journal.pone.0267047。Marios Sagris、Panagiotis Theofilis、Alexios S. Antonopoulos、Costas Tsioufis、Evangelos Oikonomou、Charalambos Antoniades、Filippo Crea、Juan Carlos Kaski 和 Dimitris Tousoulis。COVID-19 与动脉粥样硬化中的炎症机制：当前的药物治疗视角。International Journal of Molecular Sciences，22(12)，2021。URL https://www.mdpi.com/1422-0067/22/12/6607。Lei Yuan、Jun Liu 和 Jieping Ye。重叠组套索的高效方法。Advances in neural information processing systems，24，2011。Suvrit Sra。快速投影到混合范数球及其应用。Data Mining and Knowledge Discovery，25(2): 358–377，2012。Weiran Wang 和 Miguel A Carreira-Perpinán。投影到概率单纯形：一种高效且证明简洁的算法及其应用。arXiv预印本 arXiv:1309.1541，2013。Damien Garreau、Wittawat Jitkrittum 和 Motonobu Kanagawa。中位数启发式的大样本分析。arXiv预印本 arXiv:1707.07269，2017。James Bergstra 和 Yoshua Bengio。用于超参数优化的随机搜索。Journal of Machine Learning Research，13(Feb):281–305，2012。Noah Simon、Jerome Friedman、Trevor Hastie 和 Robert Tibshirani。一种稀疏组套索。Journal of computational and graphical statistics，22(2):231–245，2013。Aravind Subramanian、Pablo Tamayo、Vamsi K Mootha、Sayan Mukherjee、Benjamin L Ebert、Michael A Gillette、Amanda Paulovich、Scott L Pomeroy、Todd R Golub、Eric S Lander 等。基因集富集分析：一种用于解读全基因组表达谱的知识驱动方法。Proceedings of the National Academy of Sciences，102(43):15545–15550，2005。

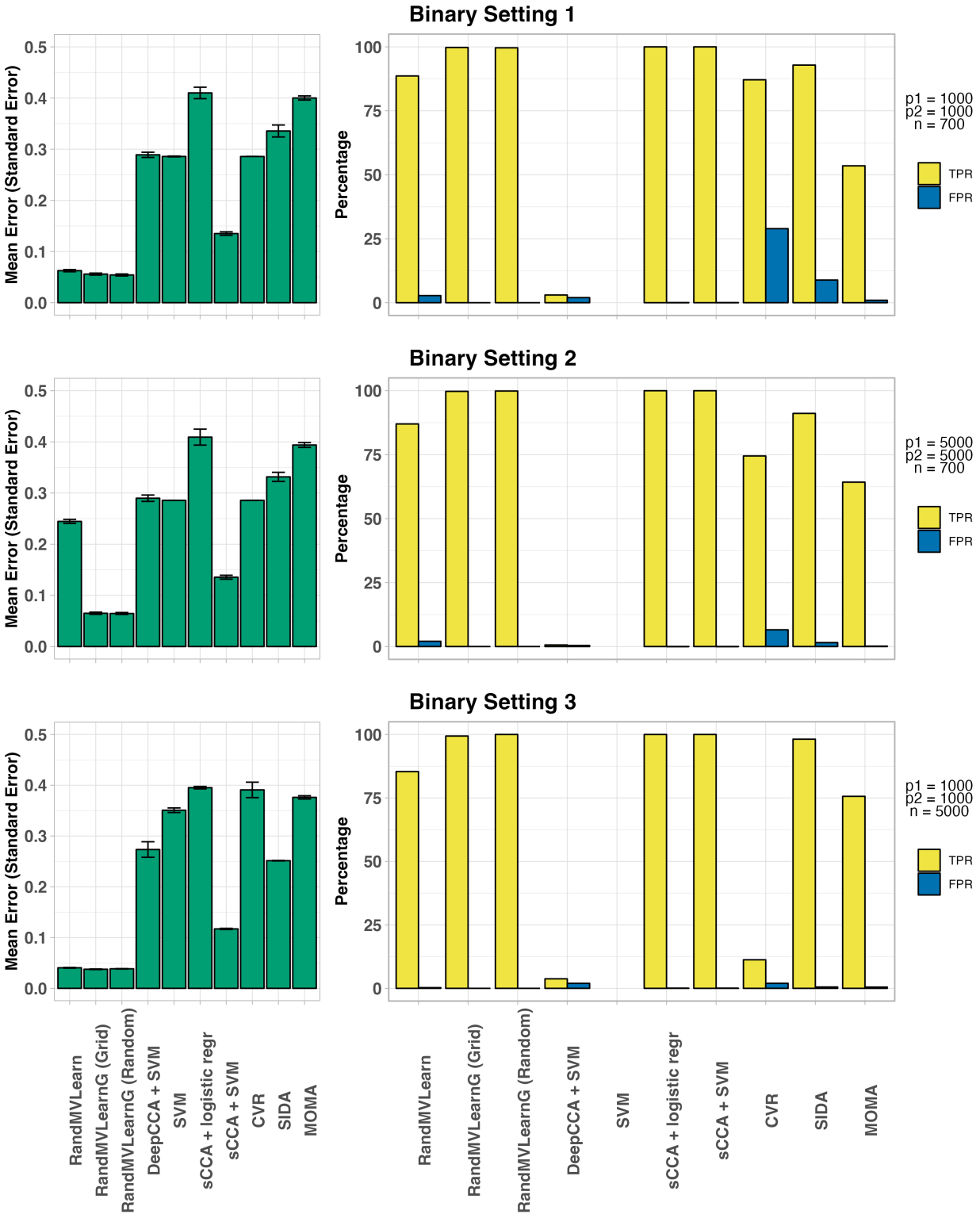


图 2：二元型设置的模拟结果。随机特征数量 M 设为 300。 (n_1, n_2) =(500，200)，(500，200)，(3000，2000)分别对应设置 1、2 和 3。成分数量固定为 $K = 5$。所提出的方法 RandMVLearn、RandMVLearnG 在几乎所有设置下的错误率都较低，真阳性率（TPR）较高或相当，假阳性率（FPR）较低或相当。

.

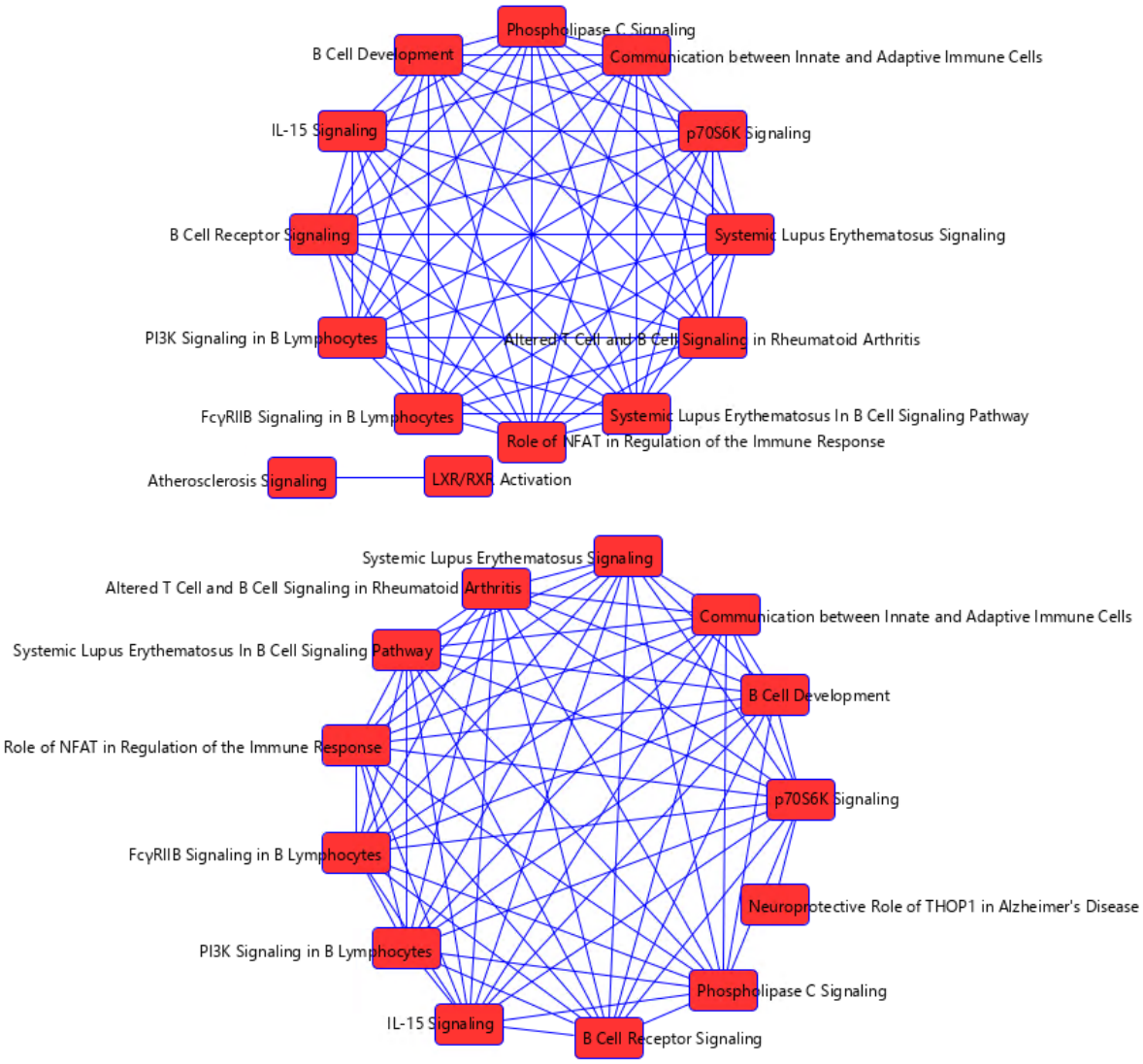


图 3：IPA中的重叠经典通路用于可视化我们蛋白质列表中参与前 15 个显著（p 值 < 0.05）富集通路的共同分子。上：COVID-19 状态的重叠通路。下：COVID-19严重程度（HFD-45）的重叠通路。尽管 COVID-19 严重程度和状态之间有许多重叠通路，但仍有少数独特通路。

补充材料

关于算法的更多内容

组变量选择

组变量选择的优化问题是:

$$\min_{\boldsymbol{\gamma}^d} f(\boldsymbol{\gamma}^d) = \mathcal{L}(\boldsymbol{\gamma}^d) + \mathcal{P}_{\eta_2}^{\eta_1}(\boldsymbol{\gamma}^d) \tag{10}$$

其中 $\mathcal{L}(\boldsymbol{\gamma}^d) = \frac{1}{2n}\|\mathbf{G} - \mathbf{Z}^d\mathbf{A}^d\|_F^2$ 且 $\mathcal{P}_{\eta_2}^{\eta_1}(\boldsymbol{\gamma}^d) = \eta_1\|\boldsymbol{\gamma}^d\|_1 + \eta_2\sum_{l=1}^{G^d}\|\sqrt{p_l^d}\boldsymbol{\gamma}_l^d\|_2$、$\eta_1 = \eta\rho$ 和 $\eta_2 = (1-\eta)\rho$。遵循 Beck 和 Teboulle [2009], 的思路，我们构造二次函数 $Q_L(\boldsymbol{\gamma}^d, \widetilde{\boldsymbol{\gamma}}^d) = \mathcal{L}(\boldsymbol{\gamma}^d) + \langle\boldsymbol{\gamma}^d - \widetilde{\boldsymbol{\gamma}}^d, \nabla\mathcal{L}(\widetilde{\boldsymbol{\gamma}}^d\rangle + \mathcal{P}_{\eta_2}^{\eta_1}(\boldsymbol{\gamma}^d) + \frac{L}{2}\|\boldsymbol{\gamma}^d - \widetilde{\boldsymbol{\gamma}}^d\|_2^2$，其中 $L > 0$ 经过适当选择，并用它在点 $\widetilde{\boldsymbol{\gamma}}^d$ 处近似 $f(\boldsymbol{\gamma}^d)$。步长 L 的选取满足 $f(\boldsymbol{\gamma}^d) \leq Q_L(\boldsymbol{\gamma}^d, \widetilde{\boldsymbol{\gamma}}^d)$。FISTA 实现中的一个关键步骤是计算惩罚项 $\mathcal{P}_{\eta_2}^{\eta_1}(\boldsymbol{\gamma}^d)$ 的近端算子（或映射）。我们采用 Yuan 等人 [2011] 和 Sra [2012] 关于近端算子的思路。为完整起见，我们在下文对其进行说明。对于给定的 $\widehat{\boldsymbol{\gamma}}^d$，稀疏组套索惩罚 $\mathcal{P}_{\eta_2}^{\eta_1}(\boldsymbol{\gamma}^d)$ 的近端算子定义如下:

$$\pi_{\eta_2}^{\eta_1}(\widehat{\boldsymbol{\gamma}}^d) = \arg\min_{\boldsymbol{\gamma}^d} \frac{1}{2}\|\widehat{\boldsymbol{\gamma}}^d - \boldsymbol{\gamma}^d\|_2^2 + \mathcal{P}_{\eta_2}^{\eta_1}(\boldsymbol{\gamma}^d). \tag{11}$$

令 $\widehat{\boldsymbol{\gamma}}^d = \widetilde{\boldsymbol{\gamma}} - \frac{1}{L}\nabla\mathcal{L}(\widetilde{\boldsymbol{\gamma}}^d)$，其中 $\widetilde{\boldsymbol{\gamma}}$ 是前两次迭代解的仿射组合。那么，解 $\boldsymbol{\gamma}^d = \arg\min_{\widehat{\boldsymbol{\gamma}}} Q_L(\boldsymbol{\gamma}^d, \widetilde{\boldsymbol{\gamma}}^d)$ 可使 $\pi_{\eta_2}^{\eta_1}(\widehat{\boldsymbol{\gamma}}^d))$ 最小化 [Yuan 等人， 2011]。稀疏组套索回归的近端算子可通过求解套索回归的近端算子，并将其用于组套索回归的近端算子中得到（Yuan 等人， [2011]）。套索回归的近端算子可归结为求解问题 (11)，其中 $\eta_1 > 0$，且 $\eta_2 = 0$，这对应于令 $\eta = 1$ 对于 $\rho > 0$；即 $\arg\min_{\boldsymbol{\gamma}^d}\frac{1}{2}\|\widehat{\boldsymbol{\gamma}}^d - \boldsymbol{\gamma}^d\|_2^2 + \rho\|\boldsymbol{\gamma}^d\|_1$。该问题的解为: $\gamma_{Lasso}^d = \text{sgn}(\hat{\gamma}^d) \cdot \max(|\hat{\gamma}^d| - \rho, 0)$。根据 Yuan 等人 [2011], 中的定理 1，优化问题 $\pi_{\eta_2}^{\eta_1}(\widehat{\boldsymbol{\gamma}}^d)$ 的解为 $= \pi_{\eta_2}^0(\boldsymbol{\gamma}_{Lasso}^d)$。因此，对于稀疏组套索回归的解，我们求解近端算子问题:

$$\pi_{\eta_2}^0(\boldsymbol{\gamma}_{Lasso}^d) = \arg\min_{\boldsymbol{\gamma}^d} \frac{1}{2}\|\boldsymbol{\gamma}^d - \boldsymbol{\gamma}_{Lasso}^d\|_2^2 + \eta_2\sum_{l=1}^{G^d}\sqrt{p_l^d}\|\boldsymbol{\gamma}_l^d\|_2. \tag{12}$$

根据 Yuan 等人的引理2， [2011], 如果第 l 个组满足 $\|\boldsymbol{\gamma}_{Lasso_l}^d\|_2 \leq \eta_2\sqrt{p_l^d}$，则第 l 个组的解为零，即 $\widehat{\boldsymbol{\gamma}}_l^d = \mathbf{0}$。我们遍历所有组，从而得到零组。对于剩余的组，即非零组，我们求解优化问题（12）。由于各组相互独立，我们分别对每个组优化（12）: $\pi_{\eta_2}^0(\boldsymbol{\gamma}_{Lasso_l}^d) = \arg\min_{\boldsymbol{\gamma}_l^d}\frac{1}{2}\|\boldsymbol{\gamma}_l^d - \boldsymbol{\gamma}_{Lasso_l}^d\|_2^2 + \eta_2\sqrt{p_l^d}\|\boldsymbol{\gamma}_l^d\|_2$，该迭代中第 l 个组的解为 $\widehat{\boldsymbol{\gamma}}_{SGLasso_l}^d = \max(1 - \eta_2\sqrt{p_l^d}\|\boldsymbol{\gamma}_{Lasso_l}^d\|_2^{-1}, 0)\boldsymbol{\gamma}_{Lasso_l}^d$ Sra [2012]。

总而言之，在迭代 $t+1$ 中，我们获得损失函数 $\mathcal{L}(\boldsymbol{\gamma}^d)$ 关于 γ^d 的梯度（即 $\nabla\mathcal{L}(\boldsymbol{\gamma}^d)$）。我们计算梯度下降步: $\widetilde{\boldsymbol{\gamma}}^d = \boldsymbol{\gamma}^d - (1/L)\nabla\mathcal{L}(\boldsymbol{\gamma}^d)$，其中 L 是步长。我们求解方程（11）中给出的近端算子问题 $\pi_{\eta_2}^{\eta_1}(\widetilde{\boldsymbol{\gamma}})$；也就是说，我们得到套索回归解 $\boldsymbol{\gamma}_{Lasso}^d$。对于第 l 组，我们检查条件 $\|\boldsymbol{\gamma}_{Lasso_l}^d\|_2 \leq \eta_2\sqrt{p_l^d}$ 是否满足。如果满足，则将第 l 组设为零，即该组的组套索回归解为零，$\widehat{\boldsymbol{\gamma}}_{SGLasso_l}^d = \mathbf{0}$。否则，我们得到解 $\widehat{\boldsymbol{\gamma}}_{SGLasso_l}^d$。我们依次遍历所有 $l = 1$、.....、G 组。我们根据 Armijo-Goldstein 规则 $f(\widehat{\boldsymbol{\gamma}}_{SGLasso}^i) \leq Q_L(\widehat{\boldsymbol{\gamma}}_{SGLasso}^i, \widetilde{\boldsymbol{\gamma}}^i)$ 更新步长 L。我们利用前两次迭代更新 $\widetilde{\boldsymbol{\gamma}}^i$，并重复该过程，直到收敛或达到某个最大迭代次数。我们在算法2中总结了优化过程。

个体变量选择

我们考虑 以下无约束优化问题，用于学习 $\boldsymbol{\gamma}^d$ 以进行个体变量选择 :

$$\min_{\boldsymbol{\gamma}^d} f(\boldsymbol{\gamma}^d) = \mathcal{L}(\boldsymbol{\gamma}^d) + \mathbf{1}^{\mathrm{T}}\boldsymbol{\gamma}^d \tag{13}$$

其中 **1** 是一个长度为 p^d 的全一矩阵向量，且 $0 \leq \boldsymbol{\gamma}^d \leq 1$。我们采用带回溯的加速投影梯度下降算法。在梯度下降算法（算法 3）的投影步中，我们使用 Wang 和 Carreira-Perpinán [2013]提出的概率单纯形算法。

超参数选择

我们在公式(9)中提出的优化包含几个值得进一步讨论的超参数： λ^d 平衡模型拟合度与复杂度；潜在成分的数量 r; 高斯核的核参数 ν；随机特征的规模 M； η 以及 ρ，用于稀疏组变量选择。我们发现 λ^d 不敏感，因此对所有视图均将其设为 1。关于随机特征的规模，在样本量范围为 500到5, 000 的模拟中，将其设为 $M = 300$ 可获得具有竞争力或更优的预测和/或变量选择性能。在样本量为 120的真实数据应用中，当我们将其设为 $M = 45$时，获得了具有竞争力的预测估计。基于我们的模拟和真实数据应用，对于大于1,000 的样本量，我们建议将其设为 $M = 300$；对于小于 1, 000 的样本量，则设为 $M \approx \frac{n}{2}$。不过，我们鼓励探索其他随机特征规模。至于高斯核参数（或带宽） ν，如果 $\nu \to 0$，Gram 矩阵将变为单位矩阵；如果 $\nu \to \infty$，Gram 矩阵中的所有元素都趋近于 1。在这两种情况下，我们都会丢失有关数据的所有相关信息。我们使用中位数启发式[Garreau等人， 2017] 来为每个视图选择 ν。我们按如下方式选择潜在成分的数量。对于每个视图，我们从Gram矩阵 $\mathcal{K}^d$(X, X)中获得特征值 $\lambda_1 > \lambda_2 \cdots > \lambda_r \cdots > \lambda_n$（在样本量较大的情况下，可能只能获得少量特征值）。然后，对于成分 $r > 2$，我们比较其与前一个成分的特征值变化 $r-1$ （即 $\frac{\lambda_r - \lambda_{r-1}}{\lambda_r}$）。我们选择使变化比例小于某个阈值的 r，例如对该视图取 0.1。潜在成分的总数 r 随后设为各视图特定成分数量中的最小值。我们知道，选择成分数量有更复杂的方法，但由于这种方法计算负担较低，且能够得到相当或更好的结果，因此我们更倾向于采用这种简单方法。稀疏组套索回归的最优超参数 η 和 ρ 可以从由 η 和 ρ构成的二维网格中搜索，但这可能计算成本高昂。因此，我们固定混合参数η，并针对 ρ 的不同取值获得解，因为 ρ 控制稀疏程度。如果许多视图具有先验组信息， ρ 的超参数搜索空间可能很大。为减少计算量，我们获得 ρ 的所有组合，并从中随机选择一些组合进行搜索，参见 [Bergstra 和 Bengio， 2012]。

算法的初始化

对于个体变量选择，我们在概率单纯形上均匀初始化 γ^d。对于组选择，我们进一步根据组权重对 γ^d 中的各组进行加权。我们从平移不变高斯核的逆傅里叶变换中随机采样 ϵ_j^d、 $j = 1$、.....、 M。我们将 $\mathbf{G}$ 初始化为一个 $n \times r$ 矩阵，使得 $\mathbf{G}^{\mathrm{T}}\mathbf{G} = \mathbf{I}_r$。我们将 $\boldsymbol{\Theta}$ 初始化为 $r \times K - 1$（用于分类结果）的随机矩阵，或 $r \times q$（用于多个连续结果）的零矩阵。我们将 $\boldsymbol{\theta}$ 初始化为

单个连续型结果对应的长度为 r 的零向量。我们将 $\mathbf{A}^d$ 初始化为一个 $M \times r$ 随机矩阵，并进行归一化，使得 $\|\mathbf{A}^d\|_F = 1$。

输入：训练数据（$\mathbf{X}^1, \ldots, \mathbf{X}^D, \mathbf{y}$）；调优参数 λ^d, M(随机特征数量), $d = 1, \ldots, D$；对于组变量选择，ρ 和 η；T，最大迭代次数

输出:估计值 $\mathbf{G}$, $\boldsymbol{\theta}$(或 $\boldsymbol{\Theta}$), $\mathbf{A}^d$, $\boldsymbol{\gamma}^d, d-1, \ldots, D$

初始化：$\mathbf{G}$，$\boldsymbol{\theta}$（或 $\boldsymbol{\Theta}$），$\boldsymbol{\gamma}^d$，$d = 1,$ ……，D 在单纯形 Δ 上均匀分布；对于组变量选择，获得 $\boldsymbol{\gamma}_l^d / \sqrt{p_l^d}$；$\overline{\ }\epsilon_j^d \sim p$（$\mathbf{w}$），$b_j^d \sim U[0, 2\pi]$；重新缩放：$\omega_j^d = \boldsymbol{\gamma}^d \cdot \epsilon_j^d$，$\forall j = 1,$ ……，M；$d = 1,$ ……，D；构造随机特征$\mathbf{Z}^d$

目标函数：$\mathcal{L}(\mathbf{Y}, \mathbf{G}\ \boldsymbol{\Theta}) + \frac{1}{2n} \sum_{d=1}^{D} \|\mathbf{G} - \mathbf{Z}^d \mathbf{A}^d\|_F^2 + \frac{\lambda^d}{2} \sum_{d=1}^{D} \|\mathbf{A}^d\|_F^2$

▷ $\mathbf{Z}^d$ 取决于 $\boldsymbol{\gamma}^d$

重复

  开始

    步骤 1：求解 $\boldsymbol{\gamma}^d$

    对于 $d = 1,$ ……，D **执行**

      $\boldsymbol{\gamma}^d \leftarrow \min_{\boldsymbol{\gamma}^d} \frac{1}{2n} \|\mathbf{G} - \mathbf{Z}^d \mathbf{A}^d\|_F^2 + \mathcal{P}(\boldsymbol{\gamma}^d)$ ▷ $\mathbf{Z}^d$ 是 γ^d的函数。 调用算法2或3。

      缩放：$\omega_j^d = \boldsymbol{\gamma}^d \cdot \epsilon_j^d, \forall j = 1, \ldots, M$

      构建随机特征$\mathbf{Z}^d$

    end

  end

  开始

    步骤 2：求解 $\mathbf{A}^d$、$\mathbf{G}$、$\boldsymbol{\theta}$（或 $\boldsymbol{\Theta}$）

    对于 $d = 1, \ldots, D$，**执行**

      $\mathbf{A}^d \leftarrow$ 最小$_{\mathbf{A}^d} \frac{1}{2n} \|\mathbf{G} - \mathbf{Z}^d \mathbf{A}^d\|_F^2 + \frac{\lambda^d}{2} \|\mathbf{A}^d\|_F^2$

    end

    求解 $\mathbf{G} \leftarrow \min_{\mathbf{G}:\mathbf{G}^\top \mathbf{G} = \mathbf{I}_r} \left\{ \mathcal{L}(\mathbf{Y}, \mathbf{G}\ \boldsymbol{\Theta}) + \frac{1}{2n} \sum_{d=1}^{D} \|\mathbf{G} - \mathbf{Z}^d \mathbf{A}^d\|_F^2 \right\}$ ▷ 最小二乘解

    ▷普鲁克斯问题

    求解 $\boldsymbol{\Theta} \leftarrow \min_{\boldsymbol{\Theta}} \mathcal{L}(\mathbf{Y}, \mathbf{G}\ \boldsymbol{\Theta})$

    ▷闭式解

  end

直到收敛或达到最大迭代次数 T

算法 1： 用于多视图学习的可扩展随机核方法算法

输入: $\mathbf{G}$,$\mathbf{A}^d$, 随机特征$\mathbf{Z}^d$, $\boldsymbol{\gamma}^d$, T (最大迭代次数); G^d (视图 d 的组数); ρ, η

输出: 估计值 $\boldsymbol{\gamma}^d$

目标函数: $f(\boldsymbol{\gamma}^d) = \frac{1}{2n}\|\mathbf{G} - \mathbf{Z}^d\mathbf{A}^d\|_F^2 + \mathcal{P}_{\eta_2}^{\eta_1}(\boldsymbol{\gamma}^d)$

▷ $\mathbf{Z}^d$ 取决于 $\boldsymbol{\gamma}^d$。 为便于记号，我们省略 $\boldsymbol{\gamma}^d$ 中的上标 d ，并且只写入迭代次数)

初始化（步骤 0）: $L > 0, L^0 = L, \beta^{-1} = 0, \beta^0 = 0, \boldsymbol{\gamma}^0 = \boldsymbol{\gamma}^d$

重复

第 $t(t \geq 1)$

集合 $\alpha^t = \frac{\beta^{t-2}-1}{\beta^{t-1}}$; $\widetilde{\boldsymbol{\gamma}}^t = \boldsymbol{\gamma}^t + \alpha^t\,(\boldsymbol{\gamma}^t - \boldsymbol{\gamma}^{t-1})$

求最小的 $L = 2^s L^{t-1}$、 $s = 0$、1、.....，使得 $f(\widehat{\boldsymbol{\gamma}}^t{}_{SGLasso}) \leq Q_L(\widehat{\boldsymbol{\gamma}}^t_{SGLasso}\widetilde{\boldsymbol{\gamma}}^t)$，其中 $\widehat{\boldsymbol{\gamma}}^t_{SGLasso}$ is 如下获得。

求损失的梯度并构造下降步: $\widehat{\boldsymbol{\gamma}}^t = \widetilde{\boldsymbol{\gamma}}^t - \frac{1}{L}\nabla\mathcal{L}(\widetilde{\boldsymbol{\gamma}}^t)$ –

▷ $\mathcal{L}(\boldsymbol{\gamma}^t) = \frac{1}{2n}\|\mathbf{G} - \mathbf{Z}^d\mathbf{A}^d\|_F^2$. $\mathbf{Z}^d$ 取决于 $\boldsymbol{\gamma}$

求套索回归解: $\widehat{\boldsymbol{\gamma}}^t_L{}_{asso} = \mathrm{sgn}(\widehat{\boldsymbol{\gamma}}^t) \cdot \max(|\widehat{\boldsymbol{\gamma}}^t| - \rho,\ 0)$

对于 $l = 1, \ldots,$ G^d **执行**

如果 $\|\widehat{\boldsymbol{\gamma}}^t_{Lasso_l}\|_2 \leq \eta_2\sqrt{p_l^d}$，那么

$\widehat{\boldsymbol{\gamma}}^i_{SGLasso_l} = \mathbf{0}$

否则，$\widehat{\boldsymbol{\gamma}}^t_{SGLasso_l} = \max(1 - \eta_2\sqrt{p_l^d}\|\widehat{\boldsymbol{\gamma}}^t_{Lasso_l}\|_2^{-1}, 0)\widehat{\boldsymbol{\gamma}}^t_{Lasso_l}$

结束循环

Set: $L^t = L$

$\boldsymbol{\gamma}^t = \widehat{\boldsymbol{\gamma}}^t_{SGLasso}$

$\beta^{t+1} = \frac{1+\sqrt{1+4(\beta^t)^2}}{2}$

直至收敛或达到最大迭代次数 T

设置 $\gamma^d = \boldsymbol{\gamma}^t$ （视图的稀疏组套索回归解 d）

算法2: 视图 d的快速稀疏组套索回归

输入: $\mathbf{G}$, $\mathbf{A}^d$, 随机特征$\mathbf{Z}^d$, $\boldsymbol{\gamma}^d$, T(最大迭代次数)
输出: 估计值 $\boldsymbol{\gamma}^d$
目标函数: $f(\boldsymbol{\gamma}^d) = \frac{1}{2n}\|\mathbf{G} - \mathbf{Z}^d\mathbf{A}^d\|_F^2 + \mathbf{1}^\intercal\boldsymbol{\gamma}^d$
▷ $\mathbf{Z}^d$ 取决于 $\boldsymbol{\gamma}^d$. 为便于表示，我们省略 $\boldsymbol{\gamma}^d$ 中的上标 d ，仅写出迭代次数)
初始化（步骤 0）： $L > 0,\ \ L^0 = L,\ \ \beta^{-1} = 0,\ \ \beta^0 = 0,\ \ \boldsymbol{\gamma}^0 = \boldsymbol{\gamma}^d$
重复
  步骤 t （$t \geq 1$）
  设定 $\alpha^t = \frac{\beta^{t-2}-1}{\beta^{t-1}}$; $\widetilde{\boldsymbol{\gamma}}^t = \boldsymbol{\gamma}^t + \alpha^t(\boldsymbol{\gamma}^t - \boldsymbol{\gamma}^{t-1})$
  找出最小的 $L = 2^s L^{t-1}$, $s = 0, \mathbf{1}, \ldots,$ 使得 $f(\widehat{\boldsymbol{\gamma}}_I^t$ $_{nd}) \leq Q_L(\widehat{\boldsymbol{\gamma}}_{I_{nd}}^t \widetilde{\boldsymbol{\gamma}}^t)$, 其中 $\widehat{\boldsymbol{\gamma}}_{I_{nd}}^t$ 得到为如下。
  求损失的梯度并构造下降步： $\widehat{\boldsymbol{\gamma}}^t = \widetilde{\boldsymbol{\gamma}}^t - \frac{1}{L}\nabla\mathcal{L}(\widetilde{\boldsymbol{\gamma}}^t)$ –
  ▷ $\mathcal{L}(\boldsymbol{\gamma}^t) = \frac{1}{2n}\|\mathbf{G} - \mathbf{Z}^d\mathbf{A}^d\|_F^2$. $\mathbf{Z}^d$ 取决于 $\boldsymbol{\gamma}$
  投影步： $\boldsymbol{\gamma}^t$ $_{Ind} =$ 最小值$_{\boldsymbol{\gamma}^d \in \Delta_2^1} \|\boldsymbol{\gamma}^t - \widehat{\boldsymbol{\gamma}}^t\|_2^2$
  ▷ 使用Wang 和 Carreira-Perpinán 中的算法 [2013]
  Set: $L^t = L$
  $\boldsymbol{\gamma}^t = \widehat{\boldsymbol{\gamma}}_{Ind}^t$
  $\beta^{t+1} = \frac{1+\sqrt{1+4(\beta^t)^2}}{2}$
直到 收敛或达到最大迭代次数 T
设置 $\gamma^d = \boldsymbol{\gamma}^t$ (视图 d的个体变量稀疏性)

算法 3： 视图 d的快速个体变量选择

更多关于模拟的内容

连续型结果变量

我们考虑四种不同的设定，它们的样本数和变量数各不相同。在每种设定中，20 个变量具有非线性关系（图 4），并被视为信号变量。我们按如下方式生成两个视图 $\mathbf{X}^1$和 $\mathbf{X}^2$, 以及结果变量$\mathbf{y}$ 的数据。

- 生成 $\boldsymbol{\theta}$，作为一个由 $n/2$ 个在 0.6和2.5 之间均匀分布的点组成的向量
- 构造向量 $\mathbf{s} = [(\boldsymbol{\theta} - \mathbf{1})^2;\ (\boldsymbol{\theta} + 0.1\mathbf{1})^2 - 2(\boldsymbol{\theta} - \mathbf{1})^2] \in \Re^{n\times 1}$
- 集合 $\tilde{\mathbf{X}}_1 = [\tilde{\boldsymbol{\theta}}; \mathbf{s}\mathbf{1}_{\mathrm{T}}{}_{p^1-1}] \in \Re^{n\times p^1}, \tilde{\boldsymbol{\theta}} = [\boldsymbol{\theta};\ \boldsymbol{\theta}] \in \Re^{n\times 1}$
- 生成 $\mathbf{X}^1 = \tilde{\mathbf{X}}_1 \cdot \mathbf{W} + \sigma_1\mathbf{E}^1$ 其中 $(\cdot)$ 是逐元素乘法，$\mathbf{W} \in \Re^{n\times p^1} = [\mathbf{1}_{20}, \mathbf{0}_{p^1-20}]$ 是一个由全一矩阵和零矩阵组成的矩阵，$\mathbf{1}$ 是全一矩阵，$\mathbf{0}$ 是零矩阵，$\sigma_1 = 0.1$，且 $\mathbf{E}^1 \sim N$服从$(\mathbf{0}, \mathbf{1})$分布。通过这种方式生成$\mathbf{X}^1$，我们假设前20个变量具有非线性关系且具有信息性。
- 生成第二视图为：$\mathbf{X}^2 = 5\mathbf{X}^1 + \sigma_2\mathbf{E}2$，其中 $\mathbf{E}^2 \sim N(\mathbf{0}, \mathbf{1})$，且 $\sigma_2 = 0.2$。
- 生成连续型结果变量 $\mathbf{y}$，表示为 $\mathbf{y} = 5\mathbf{G}\boldsymbol{\theta} + \sigma_y\mathbf{e}y$，其中 $\mathbf{G} \in \Re^{n\times 3}$ 中的列是拼接数据 $[\mathbf{X}_1$、$\mathbf{X}_2]$ $\boldsymbol{\theta} \sim U\ 0,1\ \in \Re^{2\times 1}\ \sigma_y = 0.3$ $_y \sim N\ 0,1$ ( )、和 e( ) 的奇异值分解所得的前三个左奇异向量。

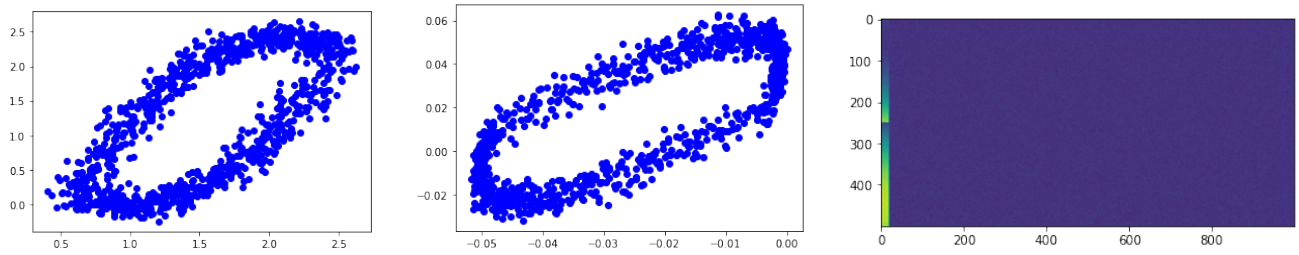


图 4：上方面板（左）：视图 1 中信号变量 1 和 2 之间的关系；（中）：视图 1 和视图 2 之间前两个共同低维表示的图（即 G）。（右）：视图 1 的图像图。前 20 个变量是信号，其余变量是噪声。视图 2 的图与视图 1 类似。

连续型结果变量

我们将所提出的方法 RandMVLearn 和RandMVLearnG（用于组变量的选择）与用于关联来自多个视图的数据的线性和非线性方法进行比较。对于线性方法，我们考虑以下方法：稀疏典型相关分析 [sCCA] [Safo 等人，2018], 典型变量回归（CVR）Luo 等人 [2016] 以及用于多视图数据的稀疏整合判别分析（SIDA）Safo 等人 [2021]。对于非线性方法，我们将我们的方法与以下方法进行比较：深度典型相关分析（Deep CCA）[Andrew等, 2013b] 和MOMA Moon和 Lee [2022]。SIDA 是一种一步法，用于同时评估多个视图之间的关联，并将受试者分类为两个或更多类别中的一个。由于sCCA和 Deep CCA 主要用于关联多个视图，我们在线性或非线性回归模型中使用这些方法得到的得分，以研究通过这些方法学习到的低维表示的预测和分类性能。为此，我们将两个视图中的典型变量进行拼接。为考察适当整合来自多个来源的数据所带来的优势，我们将这些整合分析方法与在堆叠视图上对多层感知机（MLP）进行朴素实现（用于连续型结果）以及支持向量机（SVM） [Ben-Hur 等人， 2008]（用于二元结果）进行比较。我们使用作者在GitHub 上提供的相应*SelpCCA* 和 SIDA R软件包执行 sCCA 和 SIDA。我们将所提出方法中的随机特征数量设为 300。我们使用中位数启发式为每个视图选择高斯核的带宽。我们使用作者提供的 PyTorch 代码执行 Deep CCA。为了对变量进行排序，我们将 DeepCCA 与教师-学生框架（TS） [Mirzaei 等人， 2020]相结合。我们将 TS 特征选择方法与所提出的方法进行比较。我们使用作者提供的 Python 代码执行 MOMA。

我们比较所提出的变量选择选项——个体变量选择和组变量选择的性能。对于组变量选择，有两组变量：一组信号变量和一组噪声变量。我们在稀疏组套索惩罚中固定 $\eta = 0.5$。对于每个视图，我们考虑范围 (0, ρ_{max})内的 ρ 个值。选择 ρ_{max} 以防止 $\gamma^d_{SGLasso}$中出现平凡解。我们对来自每个视图的 ρ 的合理组合进行网格搜索和随机搜索，并通过三折交叉验证选择最优的 ρ 组合。

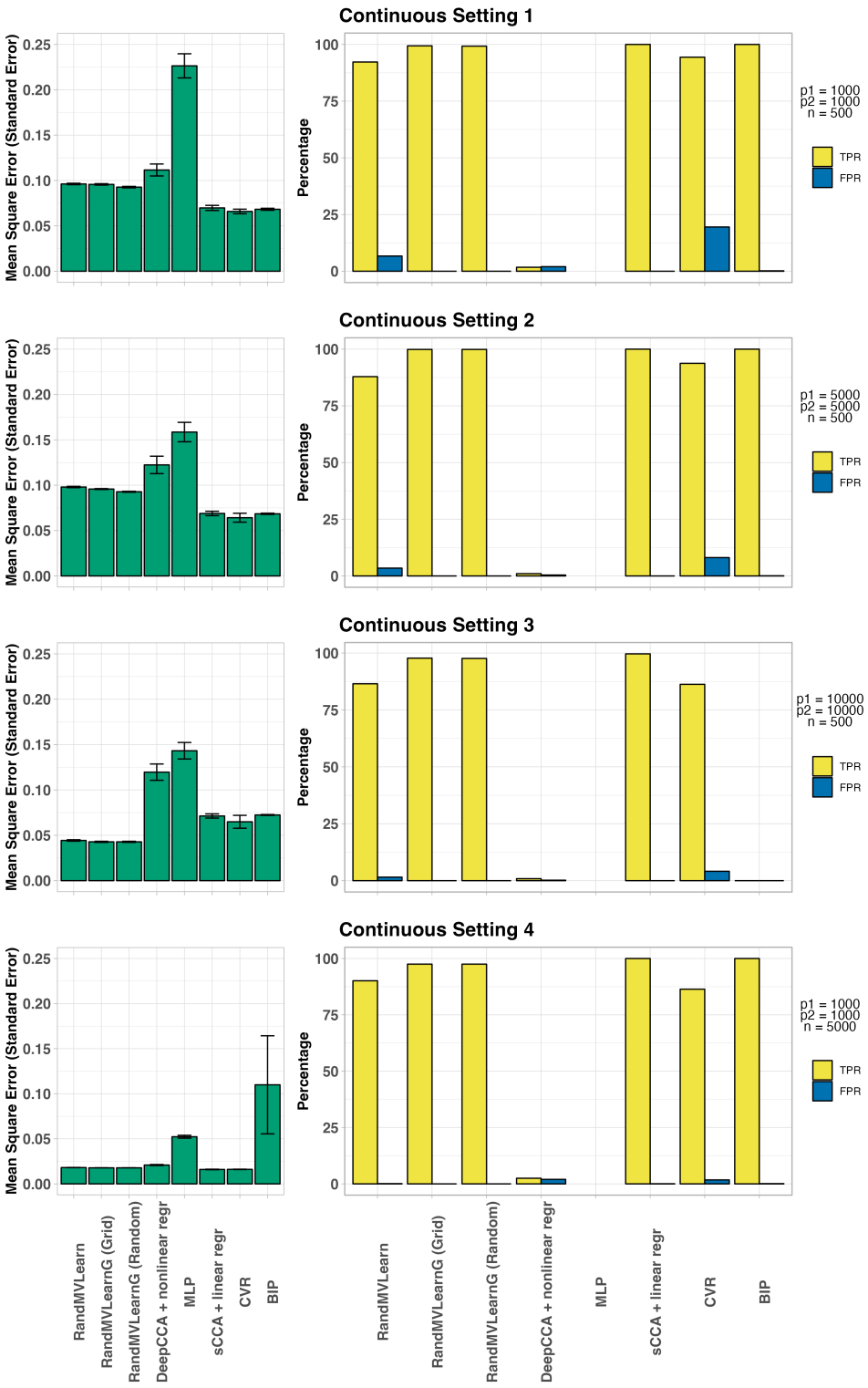


图 5：连续型设置下的模拟结果。随机特征数量 M 设置为 300。组件数量固定为 $K = 5$。所提出的方法 RandMVLearn、RandMVLearnG 在几乎所有设置下都具有更低的错误率、更高或相当的真阳性率（TPR），以及更低或相当的假阳性率（FPR）

.

表2：预测估计和变量选择。L：脂质组学；M：代谢组学；R：RNA测序；P：蛋白质组学；RandMVLearn 是所提出的不含组信息的方法。MSE 指均方误差。

方法	平均均方误差（标准误）	所选变量的平均数量 (L,M,R,P)
连续型（先验）		
RandMVLearnG	1.0029 (0.0183)	76.08/1015，33.9/72，63.0/5716，55.37/264
RandMVLearn	0.8888 (0.0212)	142.48/1015，47.48/72，207.86/5716，110.60/264
MPP 阈值为 0.9 的 BIPnet	0.8796 (0.0165)	179.20/1015, 7.52/72, 2932.82/5716, 5.96/264

图 2展示了所提出方法和比较方法的预测与变量选择性能。我们结合两个视图的变量选择性能，并报告 20 次重复实验的平均值。采用个体变量选择的方法（RandMVLearn）所得的均方误差与组变量选择方法（即 RandMVLearnG ）的均方误差相当。然而，组稀疏性方法的变量选择性能优于个体方法，突出了在组信息可用时使用组信息的优势。与 CVR（一种同时对视图之间的线性关系进行建模并预测结果变量的方法）相比，我们提出的方法具有更好的变量选择性能和具有竞争力的误差率。特别是，当变量数量较大时，我们提出的方法所得的均方误差更低；当样本数量较大时，均方误差与其相当。在所有设置下，RandMVLearnG的变量选择性能都与sCCA 相当，并且优于 MOMA。就均方误差而言，在除设置3之外的所有设置中，先使用 sCCA 再进行线性回归所得的均方误差都更低。

更多真实数据分析

预测与变量选择

假设 RNA测序数据存在组信息，我们采用所提出方法的稀疏组实现。为此，我们参考 Simon 等人 [2013] 的思路，利用细胞遗传学位置数据将基因分组成“基因集”（Subramanian 等人 [2005]的 C1 集）。在 5,800 个基因中，有 5,716 个属于 C1 基因集，我们在分析中剔除了其余基因。C1 集包含 299 个基因集，分别对应每条人类染色体及每个细胞遗传学带。我们拥有其中 291 个基因集的数据。我们选择 C1 集，是因为其中各组没有重叠。对于其余视图，我们采用个体变量选择方法。组变量选择应用中 RNA测序数据的分析数据为$\mathbf{X}^3 \in \Re^{120\times 5,\ 716}$。数据被随机划分为 50 个训练集（$n = 71$）和测试集（$n = 49$），同时确保 COVID-19 状态的比例与完整数据相近。我们使用训练数据拟合模型，并使用测试数据评估错误率——对于连续型和二元型结果，分别计算均方误差和误分类率。我们将随机特征数设为 45，使其小于训练集和测试集的样本量。我们采用所提出的简易方法确定共享低维表示的数量，并使用中位数启发式确定核参数。我们固定$\eta = 0.5$。对于 RNA测序数据的稀疏组实现，我们考虑了 ρ 的 8 个取值，范围为 (0， ρ_{max})。ρ_{max} 的选择旨在避免 $\gamma^d_{SGLasso}$中的平凡解。最优 ρ 通过三折交叉验证确定。

我们将 RandMVLearn 的预测估计值与 BIPNet [Chekouo and Safo, 2023], 这一种可同时进行数据整合和预测、并允许纳入组信息的贝叶斯方法进行了比较。BIPNet 学习视图与结果变量之间的线性关系。对于 BIPNet，我们将用于选择变量的边际后验概率设为 0.9。从表 1 可以看出，RandMVLearn 的平均预测估计值与 BIPNet 相当。RandMVLearnG 得到的预测估计值较不理想。

蛋白质列表中富集的通路、疾病和功能（无组信息）

表 3：T COVID-19 状态下蛋白质列表中排名前 15 的 IPA 显著富集通路（p 值 < 0.05） s

通路名称	-log(p 值)	分子
p70S6K信号通路	11.8	AGT,IGHV1-24,IGHV3-20,IGHV3-21,IGHV3-33, IGHV3-53 ,IGHV3-74 ,IGKV2-30 ,IGLV3-21 ,IGLV3-27 ,IGLV4-69 ,YWHAE
B细胞发育	11.2	HLA-A,IGHV1-24,IGHV3-20,IGHV3-21,IGHV3-33, IGHV3-53 ,IGHV3-74 ,IGKV2-30 ,IGLV3-21 ,IGLV3-27 ,IGLV4-69
IL-15 信号传导	9.4	IGHV1-24,IGHV3-20,IGHV3-21,IGHV3-33, IGHV3-53 ,IGHV3-74 ,IGKV2-30 ,IGLV3-21 ,IGLV3-27 ,IGLV4-69
B淋巴细胞中的 FcγRIIB 信号传导	9.36	IGHV1-24,IGHV3-20,IGHV3-21,IGHV3-33,IGHV3-53, IGHV3-74 ,IGKV2-30 ,IGLV3-21 ,IGLV3-27 ,IGLV4-69
B淋巴细胞中的PI3K信号传导	8.93	IGHV1-24,IGHV3-20,IGHV3-21,IGHV3-33,IGHV3-53, IGHV3-74 ,IGKV2-30 ,IGLV3-21 ,IGLV3-27 ,IGLV4-69
B细胞受体信号传导	8.63	IGHV1-24,IGHV3-20,IGHV3-21,IGHV3-33,IGHV3-53,
先天性免疫细胞与适应性免疫细胞之间的通信	8.2	HLA-A,IGHV1-24,IGHV3-20,IGHV3-21,IGHV3-33,IGHV3-53, IGHV3-74 ,IGKV2-30 ,IGLV3-21 ,IGLV3-27 ,IGLV4-69
类风湿关节炎中的T细胞和B细胞信号传导	8.17	HLA-A,IGHV1-24,IGHV3-20,IGHV3-21,IGHV3-33,IGHV3-53, IGHV3-74 ,IGKV2-30 ,IGLV3-21 ,IGLV3-27 ,IGLV4-69
系统性红斑狼疮信号传导	8.08	HLA-A,IGHV1-24,IGHV3-20,IGHV3-21,IGHV3-33,IGHV3-53, IGHV3-74 ,IGKV2-30 ,IGLV3-21 ,IGLV3-27 ,IGLV4-69
活化T细胞核因子在免疫应答调控中的作用	7.73	HLA-A,IGHV1-24,IGHV3-20,IGHV3-21,IGHV3-33,IGHV3-53, IGHV3-74 ,IGKV2-30 ,IGLV3-21 ,IGLV3-27 ,IGLV4-69
系统性红斑狼疮中的B细胞信号传导通路	7.6	IGHV1-24,IGHV3-20,IGHV3-21,IGHV3-33,IGHV3-53, IGHV3-74 ,IGKV2-30 ,IGLV3-21 ,IGLV3-27 ,IGLV4-69
磷脂酶 C 信号传导	6.31	IGHV1-24,IGHV3-20,IGHV3-21,IGHV3-33,IGHV3-53,IGHV3-74, IGKV2-30 ,IGLV3-21 ,IGLV3-27 ,IGLV4-69
LXR/RXR 激活	4.82	AGT,APOA4 ,LPA,S100A8
动脉粥样硬化信号传导	4.69	APOA4,ICAM1,LPA,S100A8
IL-17A 在银屑病中的作用	3.88	S100A8 ,S100A9

表4：至 p 15 IPA中COVID-19 重症蛋白质列表的显著（p-value < 0.05）富集通路 ity

通路名称	-log(p 值)	分子
p70S6K信号通路	10.4	AGT,IGHV1-18,IGHV1-24,IGHV1-3,IGHV2-70,IGKV1-17, IGKV2-30 ,IGLV3-19 ,IGLV4-69 ,YWHAE
B细胞发育	8.1	IGHV1-18,IGHV1-24,IGHV1-3,IGHV2-70,IGKV1-17, IGKV2-30 ,IGLV3-19 ,IGLV4-69
IL-15 信号传导	7.84	IGHV1-18,IGHV1-24,IGHV1-3,IGHV2-70,IGKV1-17, IGKV2-30 ,IGLV3-19 ,IGLV4-69
B淋巴细胞中的FcγRIIB信号传导	7.81	IGHV1-18,IGHV1-24,IGHV1-3,IGHV2-70, IGKV1-17 ,IGKV2-30 ,IGLV3-19 ,IGLV4-69
B淋巴细胞中的PI3K信号传导	7.46	IGHV1-18,IGHV1-24,IGHV1-3,IGHV2-70,IGKV1-17, IGKV2-30 ,IGLV3-19 ,IGLV4-69
B细胞受体信号传导	7.22	IGHV1-18,IGHV1-24,IGHV1-3,IGHV2-70,IGKV1-17, IGKV2-30 ,IGLV3-19 ,IGLV4-69
LXR/RXR 激活	7.06	AGT,APOC4 ,LBP,LPA,S100A8
系统性红斑狼疮中的B细胞信号传导通路	6.78	IGHV1-18, IGHV1-24,IGHV1-3, IGHV2-70 ,IGKV1-17 ,IGKV2-30 ,IGLV3-19 ,IGLV4-69
先天性免疫细胞与适应性免疫细胞之间的通信	5.95	IGHV1-18,IGHV1-24,IGHV1-3,IGHV2-70, IGKV1-17 ,IGKV2-30 ,IGLV3-19 ,IGLV4-69
类风湿关节炎中T细胞和B细胞信号传导异常	5.92	IGHV1-18,IGHV1-24,IGHV1-3,IGHV2-70, IGKV1-17 ,IGKV2-30 ,IGLV3-19 ,IGLV4-69
活化T细胞核因子在免疫应答调控中的作用	5.6	IGHV1-18,IGHV1-24,IGHV1-3,IGHV2-70,IGKV1-17, IGKV2-30 ,IGLV3-19 ,IGLV4-69
系统性红斑狼疮信号传导	5.51	IGHV1-18,IGHV1-24,IGHV1-3,IGHV2-70,IGKV1-17, IGKV2-30 ,IGLV3-19 ,IGLV4-69
磷脂酶 C 信号传导	5.35	IGHV1-18,IGHV1-24,IGHV1-3,IGHV2-70,IGKV1-17, IGKV2-30 ,IGLV3-19 ,IGLV4-69
IL-17A 在银屑病中的作用	4.13	S100A8 ,S100A9
THOP1在阿尔茨海默病中的神经保护作用	3.72	AGT,SERPINA3,YWHAE