Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Microbiology—Article

扩大多元回归方法在跨组学研究中的使用范围

胡小茜[a,b]，马越[a,c]，许亚昆[a,c]，Peiyao Zhao[a,d]，王军[a,c,*]

[a] CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China
[b] Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA
[c] University of Chinese Academy of Sciences, Beijing 100049, China
[d] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, USA

ARTICLE INFO

Article history:
Received 8 October 2019
Revised 14 March 2020
Accepted 25 May 2020
Available online 19 May 2021

关键词
多元回归方法
降秩回归
稀疏性
降维
变量选择

摘要

近年来科技的进步和发展使得高维数据急剧增加，研究人员对合适且有效的多元回归方法的需求也随之增长。许多传统的多元分析方法如主成分分析等已广泛应用于投资分析、图像识别和群体遗传结构分析等研究领域。然而，这些常见的方法存在其局限性，即忽略了响应之间的相关性和变量选择效率低的问题。因此，本文引入了降秩回归方法及其扩展形式——稀疏降秩回归和行稀疏的子空间辅助回归，这些方法有望满足上述需求，从而提高回归模型的可解释性。我们通过开展仿真研究来评估它们的效果，并将它们与其他几种变量选择方法进行比较。对于不同的应用场景，我们也提供了基于预测能力和变量选择精度的选择建议。最后，为了证明这些方法在微生物组研究领域的实用价值，我们将所选择的方法应用于实际种群水平的微生物组数据，结果验证了我们方法的有效性。该方法的扩展形式为未来的组学研究特别是多元回归研究提供了有价值的指导，并为微生物组学及其相关研究领域的新发现奠定了基础。

1. 引言

生物学和医学早已进入大数据时代，测序、蛋白质分析、代谢分析等分析方法和诊断技术的发展则加速了这一时代的到来。例如，相较于国际上对人类基因组计划[1]多年的努力和工作，近10年来下一代测序技术可以对个体基因组和宏基因组进行测序，甚至在单个实验室也是如此。微生物组的研究也得益于测序技术加速发展[2]，它揭示了微生物群落在人类健康和疾病[3]等领域的重要性。这些发展产生了前所未有的海量高维数据，从而促进了人们对多元回归分析的研究兴趣[4–5]。多元回归旨在对一系列响应和一系列特征之间的关系进行建模，而普通回归通常描述的是一对一的关系[6–7]。响应变量（或因变量）是研究者希望能够解释的实验结果，预测变量（或自变量）是可能引起响应变化的受控输入。例如，在基因组学研究中，此类回归中的响应变量可能是人的性状，其特征可能是遗传基因或环境因素。因此，多元回归可以应用于我们日常生活的各个方面。例如，多元回归在经济学中被广泛应用于研究影响股票收益的因素[8]。其在生物学领域也很常见[9]。例如，将其应用于临床试验，以帮助研

* Corresponding author.
E-mail address: junwang@im.ac.cn (J. Wang)

英文原文：Engineering 2021, 7(12): 1725–1731
引用本文：Xiaoxi Hu, Yue Ma, Yakun Xu, Peiyao Zhao, Jun Wang. Expanding the Scope of Multivariate Regression Approaches in Cross-Omics Research. *Engineering*, https://doi.org/10.1016/j.eng.2020.05.028.

究者解释药物成分与农药效应之间的关系[10]。近年来，在基因组学研究中，包括宏基因组学研究，以及与代谢组学、蛋白质组学等相结合的研究中，多元回归在理解重要性状的关联和潜在因果关系方面发挥了重要作用[11]。

人们在运用多元方法应对具体挑战方面进行了各种尝试。主成分分析法（PCA）是最古老、最著名的基于特征向量的多变量分析技术之一[12]。当变量个数较多时，它被广泛用于利用正交变换找到描述方差最大的变量的线性组合。通过将数据投影到低维空间可以显示其主导梯度，PCA 可以揭示数据的内部结构[13–14]。在实际应用中，主成分回归法（PCR）是一种利用 PCA 估计回归系数矩阵的线性回归模型。典范对应分析（CCA）是另一种常用的方法，它通过降维来解释两组变量之间的关系[15]。它旨在找到一个能够描述预测变量和响应变量之间最大相关性的线性组合。另一种常用于寻找两个矩阵之间关系的方法是偏最小二乘（PLS）回归法[16]。通过将响应变量和预测变量投影到新的空间中构建线性回归模型，从而总结出协方差结构。虽然上述方法在研究中得到广泛的应用，但其存在三个主要的统计学问题。第一个问题是传统方法往往忽略了观测数据和响应变量之间可能存在的相互关系。第二个问题是，有些方法不允许变量选择，但这在预测变量数量较大的探索性实验中是必不可少的。第三，一些真实数据库往往变量总数大并且样本量小，导致出现不可靠的解决方案[17–18]。

基于这些考虑，我们分析了一类新的方法[降秩回归（RRR）及其扩展]，这类方法通过考虑响应变量之间的关联来提高回归模型的可解释性[19–20]。RRR 是一种类似于 PCA 的数据缩减方法，它创建新的变量来总结原始数据中的大量信息[21]。尤其是，它通过定义一组预测变量的线性组合，以最好地解释响应变量中的总方差，并且具有简单、计算效率高、预测性能好等许多可取的优点。当预测变量数量较多时，如何选择重要变量是研究者感兴趣的另一个问题。尽管一些传统的方法仍然可用——如随机森林，它通过构建大量的决策树来进行预测——但它们更适用于响应变量单一的情况。因此，一些基于 RRR 的方法采用了群体选择方法的思想，包括群最小绝对值收敛和选择算子（群 lasso）方法等，其用于在确定变量间的群体结构时用于选择变量。在营养流行病学和遗传学方面，许多研究使用了 RRR 方法；但对不同基于 RRR 方法的性质以及其对真实数据（尤其是以宏基因组为中心数据）的适用性的全面分析还有待进行[22–25]。在此，我们使用不同维度的仿真数据，比较了基于 RRR 的各种方法与其他性质相似的多元回归方法的性能，考察了它们在不同场景下的优势和局限性，并将其应用于大规模公共宏基因组数据集。

2. 方法

2.1. 本研究中的测试方法

在本研究中，我们以基本的多元线性回归方法为出发点，然后与一些基于 RRR 和其他多元回归的方法进行比较。每一种方法的定义和原理解释如下。

2.1.1. 多元线性回归

多元线性回归模型由多个预测变量 $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p$ 和多个响应变量 $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_q$ 组成。每个响应变量由预测因子的线性回归来表示，即

$$\mathbf{Y}_j = \sum_{k=1}^{p} \mathbf{X}_k c_{kj} + \epsilon_j, \; j = 1, 2, \ldots, q \tag{1}$$

式中，c_{kj} 是与 $\mathbf{X}_k$ 和 $\mathbf{Y}_j$ 相关的回归系数；ϵ_j 是均值为零的误差项。

另外，这个公式可以用 n 个观测值改写如下：

$$\mathbf{Y} = \mathbf{XC} + \mathbf{E} \tag{2}$$

式中，$\mathbf{X}$ 为 $n \times p$ 的预测矩阵；$\mathbf{Y}$ 为 $n \times q$ 的响应矩阵；$\mathbf{C}$ 为 $p \times q$ 的回归系数矩阵；$\mathbf{E}$ 为 $n \times q$ 的误差矩阵。

我们基于最小二乘准则估计出系数矩阵 $\mathbf{C}$，即

$$\min_{\mathbf{C}} \; \| \mathbf{Y} - \mathbf{XC} \|^2 \tag{3}$$

式中，$\|\cdot\|$ 表示弗罗贝尼乌斯范数。

利用普通最小二乘法（OLS）得到 $\mathbf{C}$, $\hat{\mathbf{C}}$ 的估计值，计算公式如下：

$$\hat{\mathbf{C}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{4}$$

2.1.2. 降秩回归

然而，OLS 方法忽略了响应变量之间可能存在的相互关系，而只对每个响应变量单独进行估计，因此它只能提供一个粗略的估计。因此，我们引入了 RRR 方法，它限制了系数矩阵 $\mathbf{C}$ 的秩。我们假设 $\mathbf{C}$ 的秩较低，$r = \mathrm{rank}(\mathbf{C}) \leqslant \min(p,q)$，$\mathbf{C}$ 可以表示为两个秩为 r 的矩阵乘积，$\mathbf{C} = \mathbf{BA}^{\mathrm{T}}$，其中，$\mathbf{B}$ 和 $\mathbf{A}$ 的维数分别是 $p \times r$ 和 $q \times r$。多元回归模型（2）可以改写为：

$$\mathbf{Y} = \mathbf{XBA}^{\mathrm{T}} + \mathbf{E} \tag{5}$$

此外，XB 有 $n \times r$ 维数，表示 $\mathbf{X}$ 的一组 r 个线性组合，可以解释为驱动 $\mathbf{Y}$ 变化的潜在因素，因此，RRR 有助于降低预测变量的维度并且提高计算效率。

我们可以重写优化函数（3），即

$$\min_{\mathbf{A},\,\mathbf{B}} \left\| \mathbf{Y}-\mathbf{XBA}^{\mathrm{T}} \right\|^2 \tag{6}$$

$\hat{\mathbf{A}}$和$\hat{\mathbf{B}}$解的集合如下形式：

$$\hat{\mathbf{A}}=\mathbf{V} \tag{7}$$

$$\hat{\mathbf{B}}=\boldsymbol{\Sigma}_{\mathrm{XX}}^{-1}\boldsymbol{\Sigma}_{\mathrm{XY}}\mathbf{V} \tag{8}$$

式中，$\boldsymbol{\Sigma}_{\mathrm{XX}}=(1/n)\mathbf{X}^{\mathrm{T}}\mathbf{X}$, $\boldsymbol{\Sigma}_{\mathrm{XY}}=(1/n)\mathbf{X}^{\mathrm{T}}\mathbf{Y}$, $\boldsymbol{\Sigma}_{\mathrm{YX}}=(1/n)\mathbf{Y}^{\mathrm{T}}\mathbf{X}$，$\mathbf{V}$表示$\boldsymbol{\Sigma}_{\mathrm{YX}}\boldsymbol{\Sigma}_{\mathrm{XX}}^{-1}\boldsymbol{\Sigma}_{\mathrm{XY}}$所对应特征值的特征向量[26]。

2.1.3. 稀疏降秩回归

稀疏降秩回归（SRRR）是一种广泛的RRR方法，该方法不仅关注降维，而且关注变量选择[26–27]。它通过在最小二乘法估计中加入惩罚项来增加系数矩阵的稀疏性，从而具有独特的性质。与RRR利用所有的预测变量来构建潜在因子相比，SRRR可以从大量变量中选择有用变量，并通过引入群lasso惩罚来排除冗余变量[28]。因此，优化公式（6）可以改写如下：

$$\min_{\mathbf{A},\,\mathbf{B}} \left\| \mathbf{Y}-\mathbf{XBA}^{\mathrm{T}} \right\|^2+\sum_{j=1}^{p}\lambda_i\left\| \mathbf{B}^i \right\| \ s.\mathrm{t.}\ \mathbf{A}^{\mathrm{T}}\mathbf{A}=\mathbf{I} \tag{9}$$

式中，λ_i是惩罚参数。该约束$\mathbf{A}^{\mathrm{T}}\mathbf{A}=\mathbf{I}$用于满足可辨识条件，其中，$\mathbf{I}$表示单位矩阵。此外，如果将$\left\| \mathbf{B}^i \right\|$设为零，则整个第$i$行的矩阵$\mathbf{B}$将为零，第$i$个预测变量将不再活跃。

利用次梯度法或变分法可以求解该优化问题，但通过交叉验证（CV）定义p惩罚参数（λ_i）可能会消耗时间。为了减少调优参数的数量，通常使用如下两种策略[26]。

（1）**群lasso惩罚**：设置所有λ_i值为λ。

（2）**自适应加权lasso惩罚**：根据原始数据结构$\lambda_i=1/\left\| \tilde{\mathbf{C}}^i \right\|^{\gamma}\cdot\lambda$计算每一个$\lambda_i$，这里$\tilde{\mathbf{C}}$是$\mathbf{C}$的根-$n$一致估计，$\gamma$是正整数[29]。

2.1.4. 具有行稀疏性的子空间辅助回归

具有行稀疏性的子空间辅助回归（SARRS）也侧重于解决系数矩阵中低秩性和稀疏性的问题[30]。这种新的估计方法可以用于自适应稀疏降秩多元回归，并达到降维和变量选择的目的。此外，与前面讨论的SRRR相比，当预测变量个数超过样本容量时，SARRS能够改善其性能。

在利用群体稀疏性优化回归的过程中，可以使用两个惩罚函数。

（1）**群lasso惩罚**：$\rho(\mathbf{B};\lambda)=\lambda\left\| \mathbf{B} \right\|$，其中，$\lambda$为惩罚参数，$\mathbf{B}$为待优化的参数矩阵。

（2）**群极大极小凹惩罚（MCP）**：$\rho(\mathbf{B};\lambda)=\lambda\cdot\int_0^{\left\| \mathbf{B} \right\|}\left(1-t/\gamma^{\lambda}\right)_+\mathrm{d}t$，其中，$\gamma$是大于1的正整数，$(1-t/\gamma^{\lambda})_+$表示其正部分，即$\left(1-t/\gamma^{\lambda}\right)_+=\left(1-t/\gamma^{\lambda}\right)\cdot 1_{\left\{\left(1-t/\gamma^{\lambda}\right)\geq 0\right\}}$[31]。

2.1.5. 稀疏偏最小二乘回归

稀疏偏最小二乘回归（SPLS）方法是在PLS的基础上，进一步增强预测空间多维方向上的稀疏性，从而实现变量选择[17]。它首先选取与响应具有较强相关性的预测变量，然后再加入具有较强偏相关性的预测变量。与SRRR相比，SPLS采用了不同的降秩结构，并且没有直接聚焦于响应变量的预测，从而形成了其预测方面的弱点。

2.1.6. 识别主预测变量的正则化多元回归

识别主预测变量的正则化多元回归（REmMap）方法与上述方法不同，它不仅假设只有部分预测变量与响应相关，而且这些预测变量可能只影响部分响应[32]。这一做法是合理的，因为在实际情况中，研究者往往比其他人更关注具体的响应。REmMap能够拟合高维和小样本规模的多元回归模型，并且能够在系数矩阵中同时引入总体稀疏性和群体稀疏性来检测主预测因子。

2.1.7. 总结

表1总结了本文所讨论方法的特点。

2.2. 基于仿真数据的方法性能测试

2.2.1. 仿真设置

为了解释和比较SRRR、SARRS、SPLS、REmMap以及一些传统方法（PCR、群lasso和随机森林）的性能，我们首先引入一个仿真研究，利用上述方法生成数据并进行分析。我们采用与Chen和Huang [26]类似的仿真设置。仿真研究的核心思想是分析一些与响应变量相关的预测变量和一些不相关的预测变量。然后，我们使用这些方法来检验其中哪种方法能够最准确地判定其关系以及取得良好的预测性能。

我们用多元线性方程$\mathbf{Y}=\mathbf{XBA}^{\mathrm{T}}+\mathbf{E}$来生成数据。在该模型中，$n\times p$的设计矩阵$\mathbf{X}$服从多元正态分布$N\left(0,\,\Sigma_{\mathbf{X}}\right)$，其中，$\Sigma_{\mathbf{X}}$的对角元素为1，非对角元素为$\rho_x$。矩阵$\mathbf{B}$和矩阵$\mathbf{A}$构成了模型的系数矩阵。在$p\times r$的矩阵$\mathbf{B}$中，第一行$p_0$服从$N(0,1)$，其余$p-p_0$行为零。$q\times r$的矩阵$\mathbf{A}$由$N(0,1)$生成。矩阵$\mathbf{E}$是由$N\left(0,\,\sigma^2\Sigma_{\mathrm{E}}\right)$定义的随机噪声矩阵，其中，$\sigma^2$是噪声的大小，$\Sigma_{\mathrm{E}}$表示对角线元素为1和非对角线元素为$\rho_e$的矩阵。然后，由上述模型计算出$n\times q$的矩阵$\mathbf{Y}$。

我们生成三组数据，包括训练集、验证集和测试集。

表1 方法的比较

Methods	Data reduction method (low rankness)	Variable selection (sparsity)	Explains interrelation between responses	Features
RRR	√	—	√	Restricts the rank of the regression coefficient matrix
SRRR	√	√	√	Uses a group lasso penalty to allow row-wise sparsity
SARRS	√	√	√	Suitable when the number of predictors exceeds the sample size
SPLS	√	√	√	Uses PLS to impose sparsity
REmMap	—	√	√	Each response has different relevant predictors
PCR	√	—	—	Projects the predictors into a lower-dimensional space
Group lasso	—	√	—	Enables variable selection considering group structure
Random forest	—	√	—	Used to rank the importance of predictor variables

训练集用于拟合基于各种方法的模型。验证集用于调整模型内部的参数和估计噪声方差。最后，我们使用测试集的数据来评估所构建模型的性能。

为探索在不同情况下这些方法的适用性，我们采用几种不同的情况来进行仿真研究。首先，由于研究人员有时很难获得足够的样本进行试验，因此我们希望同时在小样本容量和大样本容量的情况下测试这些方法的性能；其次，我们还对变量数量对实验造成的影响感兴趣。真实数据如微生物学数据和遗传学数据往往包括高维预测变量或响应变量。基于以上考虑，我们仿真了以下6种情况，其中，n为数据的样本容量，p和q分别为$\mathbf{X}$和$\mathbf{Y}$中变量的个数。

案例1：小样本容量，$n < p$

案例1a：$n = 20$；$p = 100$；$q = 25$

案例1b：$n = 20$；$p = 25$；$q = 25$

案例1c：$n = 20$；$p = 25$；$q = 100$

案例2：大样本容量，$n > p$

案例2a：$n = 200$；$p = 100$；$q = 25$

案例2b：$n = 200$；$p = 25$；$q = 25$

案例2c：$n = 200$；$p = 25$；$q = 100$

在R中使用spls、rrpack、remMap、pls、glmnet和randomForest软件包对所讨论的仿真过程和方法进行了编码。SARRS方法的代码由各软件包的作者提供。附录A中指定和列出了计算过程。

2.2.2. 各种方法的评价准则

在每种情况下，我们重复了仿真过程500次，使用以下指标来衡量和比较上述多元回归方法的性能。

（1）**均方误差（MSE）**：MSE用来表示这些方法的预测精度，具体定义如下：

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{Y}}_i - \mathbf{Y}_i)^2 \tag{10}$$

式中，$\hat{\mathbf{Y}}_i$是$\mathbf{Y}_i$的预测值。

（2）**R^2**：该指数是指可由预测变量解释的响应变量中方差所占的比例。它通常用来描述模型与数据的拟合程度。R^2数值越大表明模型的拟合程度越好。

（3）**泰尔不平等系数（TIC）**：TIC是另一个用来反映拟合值与真实值差异的指标。它的定义如下：

$$\mathrm{TIC} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{Y}}_i - \mathbf{Y}_i)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i^{\,2}} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}Y_i^{\,2}}} \tag{11}$$

TIC取值范围为0 ~ 1，TIC越小，模型的预测精度越高。

（4）**敏感性（TPR）和特异性（SPC）**：TPR和SPC常用来评价变量选择的准确性。TPR是选择真正相关变量的能力，由正确选择次数与相关输入变量总数的比值计算得出。SPC是选择真正无关变量的能力，它由正确选择次数与不相关输入变量总数的比值计算得出。一种方法如果既具有高TPR又具有高SPC就意味着它可以准确地选取出相关变量。

（5）**曲线下的面积（AUC）**：AUC也用于衡量正确选择真实相关变量的比率[33]。

（6）**总体评分**：采用上述评价指标计算总体评分指数。性能较好的方法整体得分较高。

2.3. 应用于真实的种群级微生物数据

为了说明上述方法对于现实世界问题的实用性，我们将其应用于Falony等[34]工作的比利时佛兰芒肠道菌群项目（FGFP；发现群体：$n = 1106$）的数据。本研究旨在发现微生物菌群变异与宿主特征、地理和药物摄入等环境因素之间的关系。讨论了将66个基于临床和问卷的变量作为预测变量，74个微生物群作为选择后的响应变量。

我们通过CV将数据随机分成训练集和测试集。我们还建立了一个模型来拟合数据，并通过上述指标来评价其性能。我们将这个过程重复50次来检验变量选择的稳定性。

3. 结果

3.1. 仿真研究揭示了每种方法的不同性质

在仿真中，我们对不同的案例应用SRRR（带有群lasso惩罚和自适应加权群lasso惩罚）、SARRS（带有群lasso惩罚和群MCP惩罚）、SPLS、REmMap、PCR、群lasso和随机森林，并使用CV来调整每种方法的低秩参数。它们的总体表现如图1所示。

热图显示，所有的方法在案例1的性能都比案例2差，这与我们的预测一致。此外，很明显，在所有案例中都最为适用的是SARRS（带有群MCP惩罚），当样本容量增多时，SRRR（带有自适应加权群lasso惩罚）和SPLS同样适用，且性能都很好。

每种方法的效果都是通过上述标准来衡量的，案例1的结果如图2所示。

在案例1a中，样本容量非常小，预测变量数目大于样本容量；因此，大多数方法都没有很好的预测和变量选择效果。除PCR无法选择相关变量外，各方法的SPC约为0.75，TPR约为0.55，表明选择不足。然而，与传统方法（PCR、群lasso和随机森林）相比，本文讨论的新方法都具有更好的效果，特别是带有群MCP惩罚的SARRS方法。该方法的MSE和TIC最低，R^2、SPC和AUC最高。这一结果与方法论部分的论述一致，其中特别强调了当预测变量数目远远大于样本容量时，SARRS是最合适且最准确的方法。

在案例1b和1c中，预测变量数目更接近样本容量。我们发现，由于R^2较高和TIC较低，所有模型对仿真数据

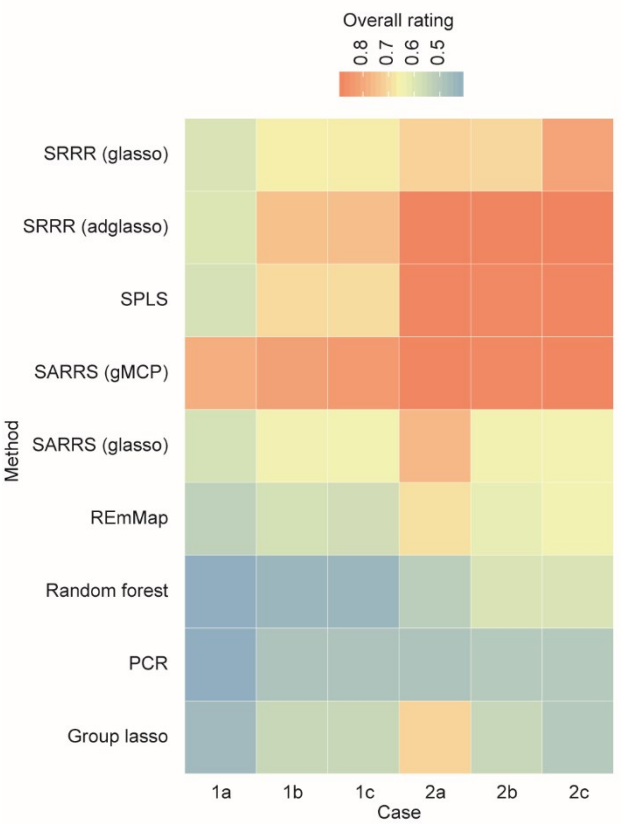

图1. 所有方法的总体评价，以热图的形式显示。x轴表示不同的案例，y轴表示不同的方法。每个单元格的颜色代表相应的总体评分。总体评分越高，表现越好。glasso：群lasso惩罚；adglasso：自适应加权群lasso惩罚；gMCP：群MCP惩罚。

的拟合效果都优于案例1a。从图中还可以看出SRRR在预测精度方面的优势，因为SPLS和REmMap相比SRRR和SARRS具有更大的MSE。此外，在变量选择方面，我们可以看到SARRS（带有群MCP惩罚）、SRRR（带有自适应加权群lasso惩罚）和SPLS的效果较好，其SPC值要高得多，表明其在选择真正的相关变量和避免过度选择之间取得了平衡。

在案例2中，我们研究了大样本容量的情况；如图3

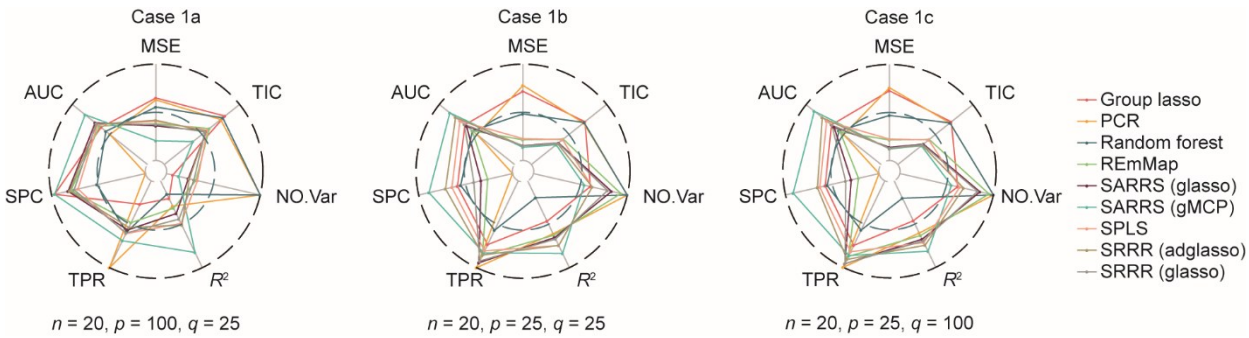

图2. 对案例1a、1b和1c中所有方法的效果评估，以雷达图形式显示。圆心为0。对于R^2、TIC、TPR、SPC和AUC，圆周为1。因此，如果一种方法在响应矩阵中的R^2、TPR、SPC和AUC较高，TIC和MSE较低，我们认为该方法效果较好。NO.Var指标给出了每种方法选择的变量数目，其中圆心表示0，边表示预测变量的数量。

所示，很明显所有的方法都比案例1的效果更好。

我们首先讨论案例2a，在这种情况下，预测变量远多于响应变量。在预测效果方面，我们感兴趣的新方法在系数矩阵中的MSE都非常低，甚至接近0，平均TIC约为0.27，平均R^2约为0.72，表明模型拟合良好。此外，传统方法在这方面的表现仍然不佳。在变量选择方面，所有方法的TPR和AUC均高于案例1。SARRS（带有群MCP惩罚）、SRRR（带有自适应加权群lasso惩罚）和SPLS也具有极高的SPC，说明它们能够准确地选择所有相关变量，剔除所有无关变量。

在案例2b中，我们将预测变量的数目减少到与响应变量相同，从而增大了样本容量与变量数目的差值。在此情况下，所有新方法的预测效果都很好。然而，在变量选择方面，这些方法的结果呈现两极化。表现最好的仍然是SRRR（带有自适应加权群lasso惩罚），其次是SARRS（带有群MCP惩罚）和SPLS，其选择精度接近1。而对于其他方法，SPC普遍较低，说明存在过度选择问题。由于案例1c类似于案例1b，因此案例2c也类似于案例2b。但是，与案例1b和1c相比，样本容量的增加使得案例2b和2c在预测精度和变量选择方面表现得更好。

3.2. 应用于实际种群水平的微生物组数据

案例研究的数据特征与上述仿真研究中的案例2b一致，其中样本容量（$n = 1106$）远大于变量数目（$p = 66$，$q = 74$）。基于之前的讨论，我们知道最适合应用于该数据集的方法是SRRR（带有自适应加权群lasso惩罚）。因此，我们构建了一个SRRR模型来分析环境指标与细菌组成之间的关系，并讨论其预测精度和变量选择结果。结果如图4所示。

模型的平均TIC为0.56，高于在仿真研究中案例2b的0.26。但由于我们知道真实的数据比仿真的数据噪声更大，因此我们认为，虽然这一TIC是可以接受的，但就充分的模型预测而言其并不令人信服。然而，在对每个响应变量进行更仔细的检查后，我们发现TIC较低的变量在以往许多研究报告中就出现过，包括FGFP [34]，如*Faecalibacterium*（TIC为0.24）、*Blautia*（0.32）、*Bacteroides*（0.33）、*Roseburia*（0.35）和*Ruminococcus*（0.40）。这些是生产丁酸的关键细菌，它们的低丰度与许多疾病有关，因为微生物组产生的低丁酸会引发较高水平的炎症和代谢紊乱[35–36]。因此，SRRR选择的预测因子可以很好地解释这些变量，并且也可以通过SRRR计算出的系数矩阵来预测。

最后，我们检验了变量选择的鲁棒性。由于我们将CV流程重复了50次，因此在80%以上的案例中选择的预测变量是最有意义的。图4（b）显示了最常选择的34个变量，包括性别、吸烟者、红细胞计数（RBC）、肌酸酐、粪便评分、平均红细胞血红蛋白浓度（MCHC）和多种药物。这一结果与先前一项研究中所讨论的药物作用的重要性一致[37]。

4. 结论

随着各个研究领域的数据量和维度的增长，生物医学研究仍然是最重要和发展最快的领域之一。当从数据中提取最大值时，在不同度量或组学水平之间获取正确、有用和有意义的关联是一项重大挑战[38]。在此，我们研究了一些有代表性的方法，包括RRR和其他多元回归方法的扩展形式，并使用仿真数据和以微生物组为中心的真实数据来说明这些方法的优势和局限性，这可能对微生物组和其他相关组学数据的未来应用具有指导意义。

我们总共研究了9种方法-参数组合，其中包括7种方法；此外，对其中两种组合使用了两种不同惩罚。我们仿真了不同样本容量/维度的数据，并比较了在维度上存在/

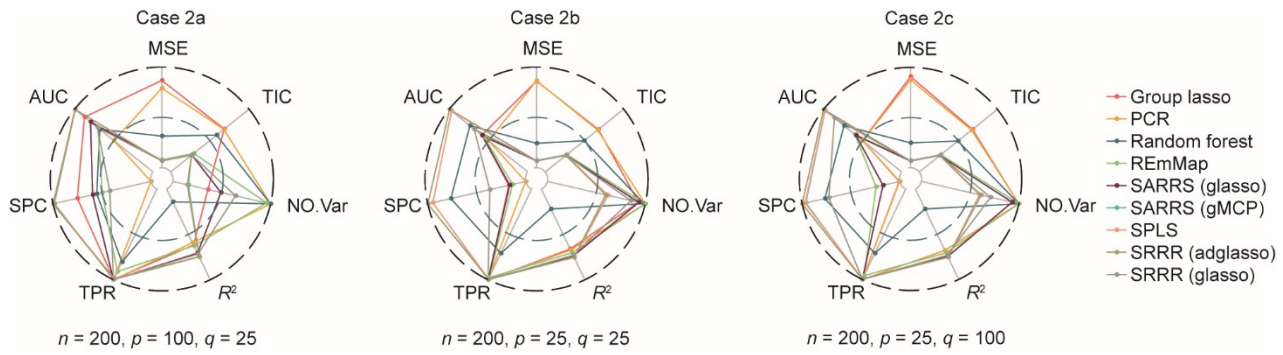

图3. 对案例2a、2b和2c中所有方法的效果评估，以雷达图形式显示。圆心为0。对于R^2、TIC、TPR、SPC和AUC，圆周为1。因此，如果一种方法在响应矩阵中的R^2、TPR、SPC和AUC较高，TIC和MSE较低，我们认为该方法效果较好。NO.Var指标给出了每种方法选择的变量数目，其中圆心表示0，边表示预测变量的数量。

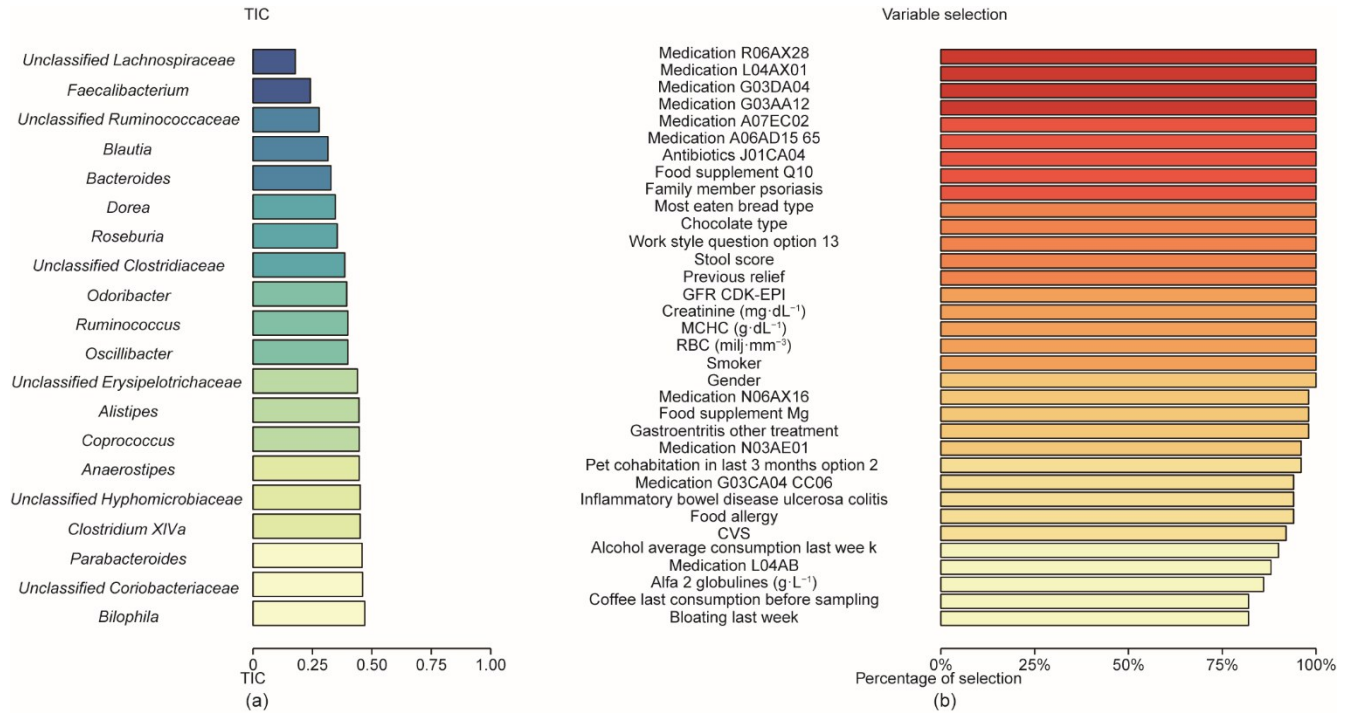

图4. 对SRRR（带有自适应加权群lasso惩罚）应用于实际种群水平微生物组数据的效果评估。（a）使用TIC指数来展示SRRR模型对每个响应变量的预测效果的条形图。图中显示了20个TIC值最低的变量。（b）显示每个预测变量在50次交叉验证中选择的百分比图。

不存在较大差异的预测变量和响应变量。从案例1和案例2的对比结果可以看出大样本容量的重要性，这将大大提升所有方法的效果。特别地，与案例1相比，SPLS在案例2中具有更好的预测精度，表明其在样本量较大时更适用。在类似于案例1中小样本容量的情况下，最好的方法是带有群MCP惩罚的SARRS，该方法在预测和变量选择方面都有优异的性能。当样本量较大时，如案例2，SARRS（带有组MCP惩罚）、SRRR（带有自适应加权群lasso惩罚）和SPLS均表现良好；通过进一步观察发现，SRRR（带有自适应加权群lasso惩罚）的效果略好于其他两种方法。

我们利用公布的FGFP数据来为一个真实场景选择最好的方法，从中可获得微生物组数据和研究中确定的环境因素。由于两类变量的维度大致相似，我们决定使用带有自适应加权群lasso惩罚的SRRR。首先，我们确定了环境变化最能解释的菌群。所确定的菌群证实了先前的论断，即产生丁酸盐的细菌对人体健康非常重要，且有可能与这些环境因素有关。另外，由于环境因素被认为是预测变量（即选定的与菌群有关的特征），我们还设法重现了已发表的研究中最重要的特征，这再次证明了所选方法的可靠性和鲁棒性。

研究人员在拟合模型时应谨慎选择合适的惩罚。例如，在SRRR方法中，当$n > p$时，自适应加权群lasso惩罚提高了预测精度和变量选择。当$n < p$时，与未加权组相比，自适应加权组的TPR较低，但SPC较高。这可以解释为，当我们在SRRR计算过程中引入权重时，之前被过滤掉的变量在其惩罚项中具有较大的权重，并且不再包含在模型中。因此，带有未加权惩罚的SRRR会选择更多的变量，导致SPC较高。

综上所述，我们检验了几种多元回归方法的适用性，并检测了它们在不同的组学情景下的效果，而在现实中，这些情景可能在样本量和维度上存在巨大差异。基于此，我们能够推荐最佳方法。诚然，我们的初步分析在现阶段无法进一步扩展，无法将不同指标（如物种）之间的系统信息纳入多组学数据中，因为这需要关于这些指标之间连通性和相似性的先验信息。我们还使用了一个著名的微生物组数据集，证明我们的选择方法可以在很大程度上概括由单变量分析获得的结果，并促进了在变量和组合特征选择方面的思考。这些发现将有助于在未来更大规模的组学研究中方法的选择，包括以微生物组为中心的研究。

致谢

本项目得到国家重点研发计划（2018YFC2000500）、中国科学院战略重点研究项目（XBD29020000）、国家自然科学基金项目（31771481和91857101）、中国科学院重点部署

项目（KFZD-SW-219）“中国科学院微生物组计划”的资助。

Compliance with ethics guidelines

Xiaoxi Hu, Yue Ma, Yakun Xu, Peiyao Zhao, and Jun Wang declare that they have no conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2020.05.028.

References

[1] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001;291(5507):1304–51.

[2] Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet 2010;11(1):31–46.

[3] Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, DiversityKnight R., stability and resilience of the human gut microbiota. Nature 2012;489(7415):220–30.

[4] Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. PLOS Comput Biol 2015;11(5):e1004226.

[5] Tsilimigras MCB, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Ann Epidemiol 2016;26(5):330–5.

[6] Izenman AJ. Modern multivariate statistical techniques: regression, classification, and manifold learning. New York: Springer-Verlag; 2008.

[7] Kharratzadeh M, Coates M. Sparse multivariate factor regression. In: Proceedings of the 2016 IEEE Statistical Signal Processing Workshop; 2016 Jun 26–29; Palma de Mallorca, Spain; 2016.

[8] Binder JJ. On the use of the multivariate regression model in event studies. J Account Res 1985;23(1):370.

[9] Kim KA, Jung IH, Park SH, Ahn YT, Huh CS, Kim DH. Comparative analysis of the gut microbiota in people with different levels of ginsenoside Rb1 degradation to compound K. PLoS ONE 2013;8(4):e62409.

[10] Peng Y, Li SN, Pei X, Hao K. The multivariate regression statistics strategy to investigate content-effect correlation of multiple components in traditional Chinese medicine based on a partial least squares method. Molecules 2018;23(3):545.

[11] Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med 2019;25(6):968–76.

[12] Smith L. A tutorial on principal components analysis. Technical report. Dunedin: University of Otago; 2002 Feb. Report No.: OUCS-2002-12.

[13] Gleason PM, Boushey CJ, Harris JE, Zoellner J. Publishing nutrition research: a review of multivariate techniques—part 3: data reduction methods. J Acad Nutr Diet 2015;115(7):1072–82.

[14] Paliy O, Shankar V. Application of multivariate statistical techniques in microbial ecology. Mol Ecol 2016;25(5):1032–57.

[15] ter Braak CJF. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 1986;67(5):1167–79.

[16] Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. Anal Chim Acta 1986;185:1–17.

[17] Chun H, Keles_ S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J R Stat Soc Series B Stat Methodol 2010;72(1):3–25.

[18] Bunea F, She Y, WegkampMH. Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. Ann Stat 2012;40(5):2359–88.

[19] Mukherjee A. Topics on reduced rank methods for multivariate regression [dissertation]. Ann Arbor: University of Michigan; 2013.

[20] D’Ambra L, Amenta P, Gallo M. Dimensionality reduction methods. Metodoloski Zveski 2005;2(1):115–23.

[21] Izenman AJ. Reduced-rank regression for the multivariate linear model. J Multivariate Analysis 1975;5(2):248–64.

[22] Hoffmann K, Schulze MB, Schienkiewitz A, Nothlings U, Boeing H. Application of a new statistical method to derive dietary patterns in nutritional epidemiology. Am J Epidemiol 2004;159(10):935–44.

[23] Cespedes EM, Hu FB. Dietary patterns: from nutritional epidemiologic analysis to national guidelines. Am J Clin Nutr 2015;101(5):899–900.

[24] Vounou M, Nichols TE, Montana G; Alzheimer’s Disease Neuroimaging Initiative. Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. NeuroImage 2010;53(3):1147–59.

[25] Vounou M, Janousova E, Wolz R, Stein JL, Thompson PM, Rueckert D, et al. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer’s disease. NeuroImage 2012;60(1):700–16.

[26] Chen L, Huang JZ. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. J Am Stat Assoc 2012;107(500):1533–45.

[27] Chen L, Huang JZ. Sparse reduced-rank regression with covariance estimation. Stat Comput 2016;26(1–2):461–70.

[28] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Series B Stat Methodol 2006;68(1):49–67.

[29] Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc 2006;101(476):1418–29.

[30] Ma Z, Sun T. Adaptive sparse reduced-rank regression. 2014. arxiv:1403.1922.

[31] Huang J, Breheny P, Ma S. A selective review of group selection in highdimensional models. Stat Sci 2012;27(4):481–99.

[32] Peng J, Zhu J, Bergamaschi A, Han W, Noh DY, Pollack JR, et al. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. Ann Appl Stat 2010;4(1):53–77.

[33] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29–36.

[34] Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, et al. Populationlevel analysis of gut microbiome variation. Science 2016;352(6285):560–4.

[35] Wan Y, Wang F, Yuan J, Li J, Jiang D, Zhang J, et al. Effects of dietary fat on gut microbiota and faecal metabolites, and their relationship with cardiometabolic risk factors: a 6-month randomised controlled-feeding trial. Gut 2019;68(8):1417–29.

[36] Sanna S, van Zuydam NR, Mahajan A, Kurilshikov A, Vila AV, Võsa U, et al. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. Nat Genet 2019;51(4):600–5.

[37] Maier L, Pruteanu M, Kuhn M, Zeller G, Telzerow A, Anderson EE, et al. Extensive impact of non-antibiotic drugs on human gut bacteria. Nature 2018;555(7698):623–8.

[38] Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C. Computational metaomics for microbial community studies. Mol Syst Biol 2013;9(1):666.