arXiv:2601.05842v1 [stat.AP] 9 Jan 2026

用于小麦籽粒产量多变量基因组预测的高光谱时间序列数据的潜在因子方法

JONATHAN F. KUNST[1],KILLIAN A.C. MELSEN[1], WILLEM KRUIJER[1], JOS ´E CROSSA[2],CHRIS MALIEPAARD[3],FRED A. VAN EEUWIJK[1], 以及 CAREL F.W. PEETERS[1]

Abstract. **关键信息：我们开发并应用了一种结合普罗克鲁斯特旋转的两阶段因子分析方法，用于将时间序列次级表型整合到多变量基因组预测中**。高维时间序列表型数据在植物育种项目中日益普遍。然而，对此类数据进行分析并将其整合用于遗传分析和基因组预测仍然十分困难。本文展示了如何对高光谱次级表型数据的遗传相关矩阵进行结合普罗克鲁斯特旋转的因子分析，从而帮助提取与试验内预测相关的特征。我们使用了国际玉米小麦改良中心（CIMMYT）2014—2015年高产精英小麦试验的一个子集，其中包含1,033个基因型。这些基因型在整个季节的多个时间点接受了三种灌溉处理，并通过载人飞机搭载高光谱传感器进行测量，传感器捕获了385—850纳米光谱范围内的62个波段。我们利用潜变量进行多变量基因组预测，以提高三种不同灌水处理下小麦籽粒产量的试验内基因组预测能力（PA）。通过在多变量基因组预测模型中整合高光谱数据的潜变量，与单变量基因组预测相比，我们能够使PA在绝对值上提高0.1至0.3（以相关性为尺度）。此外，我们还展示了试验中哪些时间点较为重要，以及这些时间点与植物生长阶段之间的关系。本文展示了如何将领域知识与数据驱动方法相结合，以提高PA，并从高通量表型分析平台的传感器数据中获得新的洞见。

Key words: 因子分析；基因组预测；高光谱数据；普罗克鲁斯特旋转；次要表型

1. Introduction

过去十年间，基因组预测以多种方式助力植物育种。通过捕捉大量微效加性遗传效应，基因组预测提高了那些拥有密集标记集的作物的复杂多基因性状（如产量）预测准确度（Jannink 等，2010）。相较于传统表型方法，它降低了获得遗传增益所需的成本和时间（Wartha 和 Lorenz，2021）。此外，基因组预测方法的灵活性推动了植物育种项目中的亲本选择和多环境评估（Jarquin 等，2020；Lopez-Cruz 和 De Los Campos，2021）。

单变量预测相对简单，但当有额外的次级表型信息可用时，多变量基因组预测更为适用。当然，也可以将次级表型信息作为预测变量加入单变量预测模型。然而，从理论角度来看，多变量模型更为合理，因为遗传效应会同时影响次要性状和目标性状。本质上，高遗传力的次要性状

([1]) Mathematical & Statistical Methods group (Biometris), Wageningen University & Research, Wageningen, The Netherlands([2]) Global Wheat Program,
International Maize and Wheat Improvement Centre (CIMMYT), Tex-coco, Mexico([3]) Plant Breeding, Wageningen University & Research, Wageningen, The Netherlands*E-mail addresses*: jonathan.kunst@wur.nl.

在性状之间存在较强遗传相关时，有助于预测遗传力较低的复杂性状（Thompson and Meyer，1986）。例如，多性状基因组预测可用于多个疾病特征（如是否存在及其形状，Jia and Jannink，2012），或谷物品质性状和产量（Michel et al.，2019）。因此，过去十年来，将多种性状整合到基因组预测模型中一直是一个活跃的研究领域（Krause 等人，2019；Dos Santos et al.，2020；Arouisse 等，2021；Runcie et al.，2021；Melsen 等，2025）。

高通量表型分析技术与基因组预测方法同步发展。这些技术有助于从细胞到个体的多个层面评估基因型（Van Dijk 等，2021）。例如，配备传感器的无人机（UAV）使跨季节开展田间试验评估变得更加可行。同时，配备自动化成像和处理施用系统的可控环境设施，能够提供大量有关单株植物处理响应的信息。这些进展使每项实验都产生了海量表型数据。

由于高通量表型数据具有高维特性，将其用于多变量基因组预测颇具挑战。使用非结构化协方差矩阵描述性状间残差时，拟合包含许多性状的模型可能会因奇异性和多重共线性而遇到问题（Antonio 等，2022）。因此，实践中需要采用某种降维或正则化方法。例如，可通过岭回归或 LASSO回归，以次要表型数据回归目标性状，随后将预测结果用于双变量预测模型（Lopez-Cruz 等，2020；Arouisse 等，2021）。尽管提高预测能力（PA）本身就是一个目标，但方法的可解释性也是需要考虑的重要方面。我们通常希望了解次要性状信息与复杂遗传性状之间的关系，以及这种关系的哪些方面对改善目标性状的预测至关重要。此外，了解哪些性状和时间点比较重要，可能有助于提高未来表型测定工作的效率，因为植物育种者可以更有选择性地决定测量哪些性状以及何时进行表型鉴定。虽然正则化方法能够提供可解释的模型或系数，但它们往往依赖难以成立的稀疏性假设，并可能导致解不稳定（Giannone 等，2021）。降维方法将原始特征投影到数据的低维表示中。这类投影不必满足稀疏性假设，也可能避免过度惩罚（Manthena 等，2022），但潜在的缺点是可解释性降低。本文提出一种可解释的降维方法，用于处理随时间变化的高维次级表型，以支持多变量基因组预测。

因子分析是许多领域广泛采用的降维技术，在植物育种领域也日益受到重视。它最初用于处理大型基因型×环境矩阵，如今则用于降低表型数据集的维数并概括相关性状（Piepho，1998年；Runcie 等，2021年；Paix˜ao 等，2022；Fialho 等，2023；Melsen 等，2025）。潜在因子的构建可以采用无监督方式，并以次级高维性状（如高光谱数据）的遗传相关矩阵为依据。此外，还可以通过旋转技术对这些潜在因子进行进一步的生物学解释。例如，通过 Procrustes 旋转（Sch¨onemann，1966年），我们可以将不同时间点的因子结构对齐，展示次要性状的表示在时间上相对稳定。简言之，所有时间点的因子结构都会旋转至所选时间点的因子结构；所选时间点可以依据领域知识确定，也可以通过数据驱动的方式选定。在此，关于小麦的领域知识可为选择提供指导：我们选择抽穗期作为锚点，将其他时间点旋转至该时间点，因为抽穗日期对籽粒产量有很大影响（Araus 等，2008）。

本文介绍了如何应用经过 Procrustes 旋转的因子分析，将高维时间序列表型数据解读并整合到基因组预测模型中。我们展示了该方法如何应用于 CIMMYT 三项高产精英试验中采集的高光谱数据；这些试验

采用了不同的灌溉方案（Krause 等人，2019）。通过逐步建模，我们说明了次要性状信息如何帮助提高预测准确度，以及在何时开展表型鉴定最为关键。因此，我们提供了一种计算效率高的动态因子模型拟合方法，同时通过指出重要的时间点和性状，帮助育种工作制定更高效的表型鉴定策略。

2. Materials and Methods

我们提出一种两阶段方法，用于利用时间性、高维次要表型进行多变量基因组预测。首先，我们采用一种特定的动态因子分析旋转方法，生成次级特征在时间上保持一致的低维表示（潜在因子的因子得分）。随后，将这些低维表示作为额外性状用于多变量基因组预测。我们的方法以图示形式呈现于图 1。下文将先介绍研究所依据的数据和符号体系，再详细说明两个建模阶段。

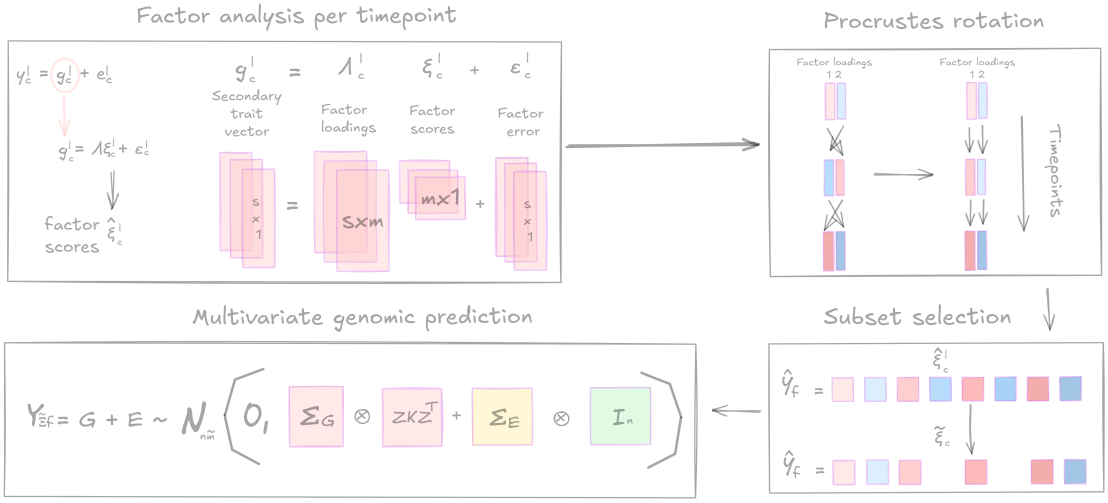

Figure 1. 我们降维和多变量基因组预测流程的图示。从次要性状的遗传协方差矩阵（红色）出发，我们估计各时间点的低维因子得分和载荷。使用 Procrustes 旋转对因子载荷重新排序，以追踪各时间点的因子得分。接下来，我们使用子集选择来确定ˆ˜因子得分（ξ）与目标性状（ˆ$\mathbf{y}f$）之间的关系。随后，通过（基因组）关系矩阵（$\mathbf{K}$），将选定的因子得分（$\boldsymbol{\Xi}$）和目标性状（ˆ$\mathbf{y}f$）用于对未观测基因型进行多变量基因组预测。下文各节将完整阐述符号体系和建模方法。

2.1. 动机数据：CIMMYT 精英产量试验。我们使用在墨西哥城奥布雷贡市诺曼·E·博洛格研究站生成的 CIMMYT 精英小麦产量试验数据集（Krause 等人，2019）。在该数据集中，次要性状为高光谱波长反射率，目标性状为籽粒产量。我们聚焦于其中一部分数据，包括 2014–2015 年的试验；这些试验采用三种干旱程度逐渐加重的灌溉处理：最佳平坦（沟灌 5 次）、中度干旱（沟灌 3 次）和重度干旱（沟灌 2 次）。在这些处理下，对 1，033 个基因型（每个基因型有 3 次重复）进行了评估，并使用飞机在多个时间点（7 至 10 个日期，具体取决于

处理）采集高光谱数据。光谱范围为 $385-850$ 纳米，间隔为 7.5 纳米，因此每个时间点测得 62 个波段的反射率。反射率基于每个试验中小区中部区域的平均值。此外，还使用 8，519 个单核苷酸多态性（SNP）标记对群体进行了基因分型。详情请参阅 Krause 等人（2019）。

2.2. **符号说明**。我们使用实验设计已解析的表型数据，做法与 Lopez-Cruz 等人（2020）相同。对于基因型 $c=1,\ldots,g$，我们有 $j=1,\ldots,s$ 个次要性状，涉及 $i=1,\ldots,n$ 个小区，每个小区有 $q=1,\ldots,r$ 个重复，并包含 $l=1,\ldots,\tau$个时间点。在我们的数据情境中，我们考虑平衡设计（尽管这绝非下文所提方法的必要条件），其中有 $g=1,033$、$s=62$和 $r=3$。由此得到 $n=gr=3,099$ 个小区；根据处理不同，最佳平坦条件下为 $\tau=10$，中度干旱条件下为 $\tau=8$，重度干旱条件下为 $\tau=7$。

对于向量，我们使用粗体小写符号（即 $\boldsymbol{a}$ 表示一个向量）；对于矩阵，则使用粗体大写符号（即 $\mathbf{A}$ 表示一个矩阵）。必要时，我们会指定数据矩阵中的行，例如，$\boldsymbol{a}_{(q|c)}$ 表示基因型 c 的重复 q 对应的数据向量。我们通常用 $\mathbf{K}$表示亲缘关系矩阵。此外，我们分别用 $\boldsymbol{\Sigma}$ 和 $\mathbf{R}$ 表示协方差矩阵和相关矩阵。我们用下标 P、G和 E 分别表示表型、遗传和残差协方差矩阵及相关矩阵。例如，$\boldsymbol{\Sigma}_E$ 表示残差协方差矩阵。上标 l 表示时间点。因此，例如，$\boldsymbol{\Sigma}_E^l$ 表示时间点 l对应的残差协方差矩阵。我们用 $\mathbf{I}_n$ 表示维数为（$n\times n$）–的单位矩阵。

在运算符号方面，我们用 $\otimes$表示克罗内克积，用 $\circ$表示哈达玛积。我们用 $\mathrm{diag}(a_{11},\ldots,a_{ss})$ 表示一个 $(s\times s)$ 维对角矩阵，其对角元素为$a_{11},\ldots,a_{ss}$。矩阵的按列向量化记为 $\mathrm{vec}(\mathbf{A})$，$\mathrm{vec}^{-1}(\mathbf{A})$ 表示其逆运算。此外，如果$\mathbf{A}$ 是方阵，我们用$\mathbf{A}^{1/2}$ 表示矩阵平方根，用$\mathbf{A}^{-1/2}$ 表示$\mathbf{A}^{-1}$ 逆矩阵的平方根。另外，我们用 $I(\cdot)$表示参数为 $(\cdot)$ 的指示函数。最后，弗罗贝尼乌斯范数记为 $\|\cdot\|_{\mathcal{F}}$。

2.3. **高维次级性状的投影**。以下各节将介绍如何把高维性状投影到低维潜变量上。首先，我们说明如何将表型数据视为遗传效应与残差效应的组合。随后，我们估计每个时间点次级性状的遗传协方差，并进一步估计各时间点的因子载荷和提取因子得分。之后，我们使用普罗克鲁斯特旋转和平滑来追踪不同时间点的因子得分。

2.3.1. 次要性状的基本模型。我们将每种处理下完整的次级表型矩阵表示为$\mathbf{Y}\in\mathbb{R}^{n\times s\tau}$。按时间点划分的次级表型矩阵表示为$\mathbf{Y}^l\in\mathbb{R}^{n\times s}$。假设矩阵$\mathbf{Y}^l$ 可分解为遗传效应（$\mathbf{G}^l$）和残差效应（$\mathbf{E}^l$），也就是说，$\mathbf{Y}^l$ 是遗传效应与残差效应之和。因此，$\mathrm{vec}(\mathbf{Y}^l)$服从以下多元正态分布：

$$\mathrm{vec}(\mathbf{Y}^l)=\mathrm{vec}(\mathbf{G}^l)+\mathrm{vec}(\mathbf{E}^l)\sim\mathcal{N}_{ns}\big(\mathbf{0},\boldsymbol{\Sigma}_G^l\otimes\mathbf{Z}\mathbf{K}\mathbf{Z}^\top+\boldsymbol{\Sigma}_E^l\otimes\mathbf{I}_n\big),$$

其中，$\boldsymbol{\Sigma}_G^l\in\mathbb{R}^{s\times s}$ 和 $\boldsymbol{\Sigma}_E^l\in\mathbb{R}^{s\times s}$ 分别为遗传协方差矩阵和残差协方差矩阵，$\mathbf{Z}\in\mathbb{R}^{n\times g}$表示关联矩阵，$\mathbf{K}\in\mathbb{R}^{g\times g}$。我们根据观测到的基因型（例如训练集中的基因型）的最佳线性无偏估计量（BLUE）的协方差矩阵，以及小区水平残差的协方差矩阵，得到 $\boldsymbol{\Sigma}_G^l$ 和 $\boldsymbol{\Sigma}_E^l$：

$$\bar{\mathbf{y}}_c^l=\mathbf{g}_c^l+\bar{\boldsymbol{\epsilon}}_c^l=\mathbf{g}_c^l+\frac{1}{r}\sum_{q=1}^{r}\boldsymbol{\epsilon}_{(q|c)}.$$

假设基因型 c 和重复 q 的小区水平残差 $\boldsymbol{\epsilon}_{(q|c)}^l$ 独立同分布，因此残差协方差矩阵可表示为 $r^{-1}\boldsymbol{\Sigma}_E^l$（Kruijer 等人，2020），从而得到 $\bar{\mathbf{y}}_c^l\sim\mathcal{N}_s\left(\mathbf{0},\ \boldsymbol{\Sigma}_P^l\right)$，其中 $\boldsymbol{\Sigma}_P^l=\boldsymbol{\Sigma}_G^l+r^{-1}\boldsymbol{\Sigma}_E^l$。协方差

矩阵的估计方法见 Melsen 等人（2025）。现在，我们需要关注每个时间点的 $\mathbf{\Sigma}_G^l$ 和$\mathbf{\Sigma}_E^l$，前者将用于后续的因子分析，后者则需要与潜在因子得分一起用于多变量BLUP方程。

2.3.2. 遗传效应的因子模型。 我们基于每个时间点所有观测到的次要性状的空间校正后基因型均值的遗传相关矩阵进行因子分析。次要性状的遗传成分可分解为:

$$\mathbf{g}_c^l = \mathbf{\Lambda}^l \boldsymbol{\xi}_c^l + \boldsymbol{\varepsilon}_c^l,$$

其中，$\mathbf{\Lambda}^l \in \mathbb{R}^{s\times m}$ 为 s 个次要性状在 m 个因子上的因子载荷矩阵，$\boldsymbol{\xi}_c^l \in \mathbb{R}^{m\times 1}$为基因型 c 的潜在因子，ϵ为因子模型误差。这里我们作出以下假设（Peeters 等人，2019年；Melsen 等人，2025年）：(i) $\text{rank}(\mathbf{\Lambda}^l) = m < s$，(ii) $\boldsymbol{\varepsilon}_c^l \sim \mathcal{N}_s\left(\mathbf{0},\ \mathbf{\Psi}^l\right)$，其中 $\mathbf{\Psi}^l = \mathbb{E}\left(\boldsymbol{\varepsilon}_c^l \boldsymbol{\varepsilon}_c^{l\top}\right)$ 和 $\mathbf{\Psi}^l \equiv \text{diag}\left(\psi_{11}^l,\ \ldots\ldots,\ \psi_{ss}^l\right)$，且 $\psi_{jj}^l > 0$、$\forall jj$；(iii) $\boldsymbol{\xi}_c^l \sim \mathcal{N}_m\ (\mathbf{0},\ \mathbf{I}_m)$；(iv) $\boldsymbol{\xi}_c^l \perp\!\!\!\perp \boldsymbol{\varepsilon}_{c'}^l,\ \forall c,\ c'$；(v) $\boldsymbol{\xi}_c^l \perp\!\!\!\perp \bar{\boldsymbol{\epsilon}}_{c'}^l,\ \forall c,\ c'$；以及(vi) $\boldsymbol{\varepsilon}_c^l \perp\!\!\!\perp \bar{\boldsymbol{\epsilon}}_{c'}^l,\ \forall c,\ c'$。根据这些假设，我们将每个时间点的遗传协方差矩阵（$\mathbf{\Sigma}_G^l$）分解为:

$$\mathbf{\Sigma}_G^l\left(\mathbf{\Theta}^l\right) = \mathbb{E}\left[\left(\mathbf{\Lambda}^l\boldsymbol{\xi}_c^l + \boldsymbol{\varepsilon}_c^l\right)\left(\mathbf{\Lambda}^l\boldsymbol{\xi}_c^l + \boldsymbol{\varepsilon}_c^l\right)^\top\right] = \mathbf{\Lambda}^l\mathbf{\Lambda}^{l\top} + \mathbf{\Psi}^l,$$

其中，$\mathbf{\Theta}^l = \{\mathbf{\Lambda}^l, \mathbf{\Psi}^l\}$ 表示时间点 l处模型推导出的载荷和独特性矩阵。我们根据各时间点的遗传相关矩阵估计载荷矩阵及后续的因子得分，因为介于 $[-1, 1]$ 之间的标准化载荷易于解释。不过，也可以使用协方差矩阵进行估计，因为该估计过程与尺度无关，但这样得到的结果较难解释。给定估计得到的遗传协方差矩阵 $\hat{}\ \mathbf{\Sigma}_G^l$，我们将其缩放为相关矩阵:

$$\hat{\mathbf{R}}_G^l = \left(\hat{\mathbf{\Sigma}}_G^l \circ \mathbf{I}_s\right)^{-1/2}\hat{\mathbf{\Sigma}}_G^l\left(\hat{\mathbf{\Sigma}}_G^l \circ \mathbf{I}_s\right)^{-1/2}.$$

2.3.3. 各时间点的参数估计。 进行因子分析时，需要确定因子 m 的数量（即因子维度）。我们通过使用 m的代理指标 m^* 来确定这一数量。此处采用碎石图拐点检测的一种数值替代方法，即加速因子（Raˆıche 等，2013）。该方法找到高加速出现前的特征值，并以 Kaiser准则规定下限（即特征值 > 1，见附录 A）。这一方法遵循拐点之后信息量递减的理念。对于每个时间点 l，我们通过最小化以下差异函数，估计因子模型参数集合 $\hat{}\ \mathbf{\Theta}_{m^*}^l = \{\hat{}\mathbf{\Lambda}^l,\ \hat{}\mathbf{\Psi}^l\}_{m^*}$（J¨oreskog，1967）：

$$F\left[\hat{\mathbf{R}}_G^l\left(\mathbf{\Theta}\right)^l; \hat{\mathbf{R}}_G^l; m^*\right] = \ln\left|\mathbf{\Lambda}^l\mathbf{\Lambda}^{l\top} + \mathbf{\Psi}^l\right| + \text{tr}\left[\hat{\mathbf{R}}_G^l\left(\mathbf{\Lambda}^l\mathbf{\Lambda}^{l\top} + \mathbf{\Psi}^l\right)^{-1}\right] - \ln|\hat{\mathbf{R}}_G^l| - s,$$

ˆ这要求$\mathbf{R}_G$ 是正定且良态的样本矩阵。因此得到:

$$\hat{\mathbf{\Theta}}_{m^*}^l \equiv \underset{\mathbf{\Lambda}^l, \mathbf{\Psi}^l}{\arg\min}\, F\left[\mathbf{R}_G^l\left(\mathbf{\Theta}\right); \hat{\mathbf{R}}_G^l, m^*\right].$$

此外，我们对载荷矩阵进行方差最大旋转（Kaiser，1958），因为这种旋转能得到近似稀疏的载荷矩阵，从而提高可解释性。因此，旋转后，次要性状在某个特定因子上的载荷较高，而在其他因子上的载荷较低。

2.3.4. *Procrustes* 旋转和平滑，以确保时间点之间的一致性。 估计出每个时间点 l的载荷矩阵 $\hat{}\mathbf{\Lambda}^l$ 后，我们希望追踪因子得分随时间的变化。由于方差最大旋转准则对符号切换和标签切换具有不变性，如果每个时间点的旋转仅基于无约束的方差最大旋转，我们可能会失去载荷和得分的可解释性。基本上，对于每个随时间变化的载荷矩阵，我们希望找到一个带符号置换矩阵，使所有这类矩阵具有相同的反射（即列顺序相同，且所有列的极性相同）。我们采用 Procrustes 旋转来解决这一问题。通过 Procrustes

旋转，我们可以将每个载荷矩阵旋转至选定的目标矩阵 $\hat{\mathbf{\Lambda}}^{l\prime}$（Sch¨onemann，1966年）附近。因此，我们可以将所有随时间变化的载荷矩阵和因子得分表示为与具有生物学意义的目标时间点（例如抽穗日期）相同的反射。这样，我们便能观察单个试验中因子在整个季节内的动态变化。实际操作中，这意味着要找到一个正交旋转矩阵 $\mathbf{T}$，使载荷矩阵之间的差异在最小二乘意义下达到最小（Sch¨onemann，1966年）：

$$\mathbf{T}^l \equiv \arg\min_{\mathbf{H}} \left\{ \left\| \hat{\mathbf{\Lambda}}^l \mathbf{H} - \hat{\mathbf{\Lambda}}^{l\prime} \right\|_{\mathcal{F}} : \mathbf{H}^\top \mathbf{H} = \mathbf{I}_{m^*} \right\}.$$

请注意，Procrustes旋转是一种严格的旋转准则，因为它可能会消除部分（抽样）变异，而不只是对$\hat{\mathbf{\Lambda}}^l$各列的顺序和极性进行置换。然而，$\mathbf{T}^l$ 将接近带符号置换矩阵 $_{\pm}\mathbf{P}^l$。我们可以使用 Peeters（2012，附录 2A）中描述的平滑算法，从$\mathbf{T}^l$ 中提取带符号置换矩阵 $_{\pm}\mathbf{P}^l$ 。这样，乘积 $\hat{\mathbf{\Lambda}}^l {}_{\pm}\mathbf{P}^l$ 便能确保 $\hat{\mathbf{\Lambda}}^l$ 与目标 $\hat{\mathbf{\Lambda}}^{l\prime}$具有相同的反射方向。现在，我们可以通过近似稀疏因子表示来表征次要性状，同时追踪它们随时间的变化。

2.3.5. 估计各时间点的因子得分。每个时间点的（Procrustes）旋转因子得分 $\hat{\mathbf{\Xi}}^l$，通过改进的 Thomson 回归方法估计（Thomson，1939年；Melsen 等，2025年）。简言之，我们将次要表型投影到较低维数的因子空间中，得到：

$$\hat{\mathbf{\Xi}}^l = \bar{\mathbf{Y}}^l \left(\hat{\mathbf{\Psi}}^l + r^{-1} \hat{\mathbf{\Sigma}}_E^l \right)^{-1} \hat{\mathbf{\Lambda}}^l \left[\mathbf{I}_{m^*} + \hat{\mathbf{\Lambda}}^{l\top} \left(\hat{\mathbf{\Psi}}^l + r^{-1} \hat{\mathbf{\Sigma}}_E^l \right)^{-1} \hat{\mathbf{\Lambda}}^l \right]^{-1}.$$

随后可将这些因子得分直接用于后续建模步骤。另一种做法是追踪这些因子得分随时间的变化，从而为建模提供多种选择。其中一种选择是使用平滑样条可视化随时间变化的趋势（P´erez-Valencia et al., 2022; Boer, 2023），以展示因子与性状数据（如产量）之间的关系（图 2）。此外，如有需要，也可以用样条特征替代因子得分，详见附录 B。

2.4. **多变量基因组预测**。 获得各时间点的因子得分后，我们根据贝叶斯信息准则（Schwarz, 1978）最小化的原则，纳入次要性状因子的最佳子集。在训练集中，通过对平均目标性状 $\mathbf{y}_f$ 与因子得分进行回归，确定最佳子集。我们从 τm^* 个次要性状因子得分开始。经过子集选择后，得到候选因子 $\tilde{m} \leq \tau m^*$，并构成矩阵 $\tilde{\mathbf{\Xi}}$。随后，我们将选定的因子得分与目标性状组合如下：

$$\mathbf{Y}_{\tilde{\Xi} f} = \begin{bmatrix} \tilde{\mathbf{\Xi}} & \mathbf{y}_f \end{bmatrix},$$

其中，$\tilde{\mathbf{\Xi}}$ 为包含小区水平因子得分的 $n \times \tilde{m}$ 矩阵，$\mathbf{y}_f$为目标性状的 $n \times 1$ 列向量。该矩阵同样由遗传成分和残差成分组成，如下所示：

$$\mathbf{Y}_{\tilde{\Xi} f} = \mathbf{G}_{\tilde{\Xi} f} + \mathbf{E}_{\tilde{\Xi} f}$$

Ass假设遗传成分与残差成分相互独立，我们得到 t:

$$\mathrm{vec}(\mathbf{Y}_{\tilde{\Xi} f}) = \mathrm{vec}\left(\mathbf{G}_{\tilde{\Xi} f}\right) + \mathrm{vec}\left(\mathbf{E}_{\tilde{\Xi} f}\right) \sim \mathcal{N}_{n\tilde{m}}\big(\mathbf{0}, \mathbf{\Sigma}_G \otimes \mathbf{ZKZ}^\top + \mathbf{\Sigma}_E \otimes \mathbf{I}_n\big),$$

其中，$\mathbf{ZKZ}^\top$ 和$\mathbf{I}_n$ 分别是$\mathbf{G}^{\tilde{\Xi} f}$ 和$\mathbf{E}^{\tilde{\Xi} f}$的行协方差。同样， $\mathbf{\Sigma}_G$ 和 $\mathbf{\Sigma}_E$分别是$\mathbf{G}^{\tilde{\Xi} f}$ 和$\mathbf{E}^{\tilde{\Xi} f}$的列协方差。这些协方差矩阵按块划分如下：

$$\mathbf{\Sigma}_G = \begin{bmatrix} \mathbf{\Sigma}_G^{\tilde{\Xi}\tilde{\Xi}} & \boldsymbol{\sigma}_G^{\tilde{\Xi} f} \\ \boldsymbol{\sigma}_G^{f\tilde{\Xi}} & \sigma_G^f \end{bmatrix}, \quad \mathbf{\Sigma}_E = \begin{bmatrix} \mathbf{\Sigma}_E^{\tilde{\Xi}\tilde{\Xi}} & \boldsymbol{\sigma}_E^{\tilde{\Xi} f} \\ \boldsymbol{\sigma}_E^{f\tilde{\Xi}} & \sigma_E^f \end{bmatrix},$$

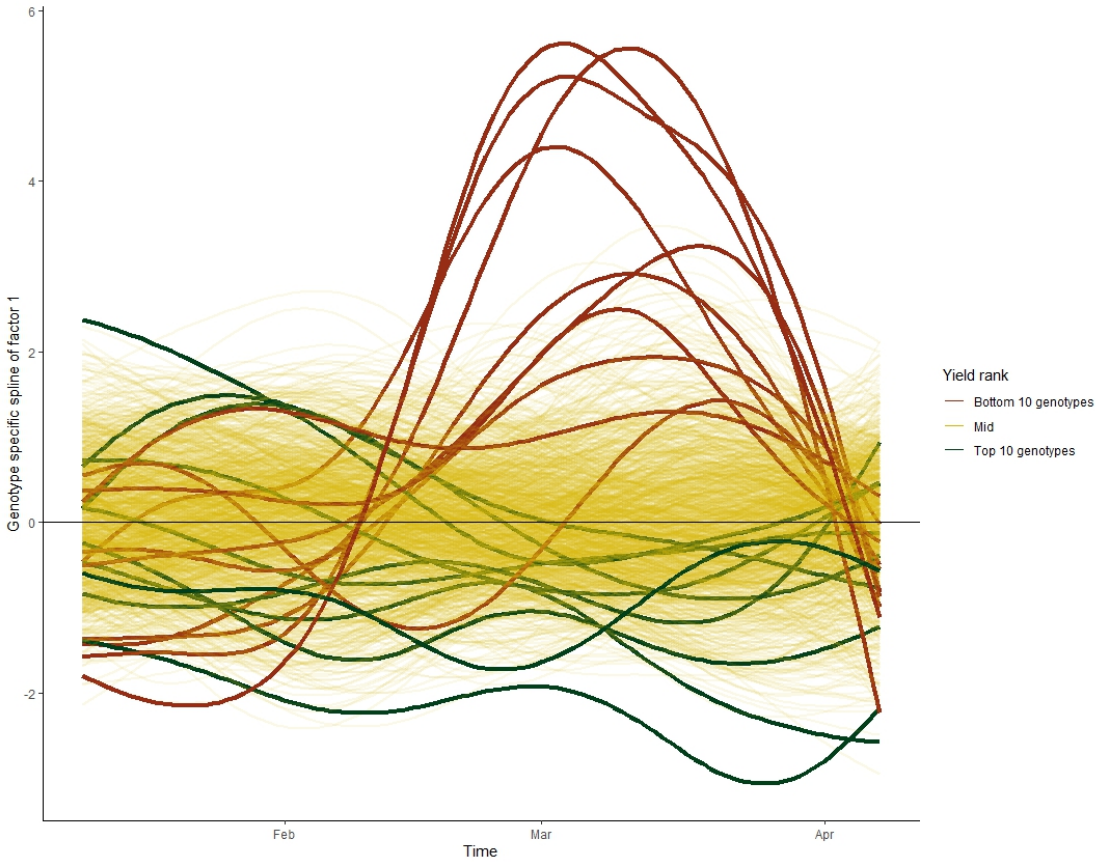

FIGURE 2. 基于 Optimal Flat 处理中的因子 1，按基因型使用平滑样条的示意图。从产量来看，排名最低的 10 个基因型（红色）在 3 月左右开始与表现最佳的 10 个基因型（绿色）出现分化。此外，$y = 0$处的黑线表示群体样条。因此，可以将因子空间也视为赋范空间。

其中，$\boldsymbol{\Sigma}_G^{\tilde{\Xi}}$ 和 $\boldsymbol{\Sigma}_E^{\tilde{\Xi}}$ 分别是因子得分的遗传协方差和残差协方差；σ_G^f 和 σ_E^f 分别表示目标性状的遗传标量方差和残差标量方差。非对角元素是因子得分与目标性状之间的协方差，分别记为 $\boldsymbol{\sigma}_G^{\tilde{\Xi}f} = \left(\boldsymbol{\sigma}_G^{f\tilde{\Xi}}\right)^\top$ 和 $\boldsymbol{\sigma}_E^{\tilde{\Xi}f} = \left(\boldsymbol{\sigma}_E^{f\tilde{\Xi}}\right)^\top$。请注意，若不进行降维，这些协方差矩阵的规模会大得多，从而导致多变量基因组预测的计算成本高昂。这种多变量基因组预测方法意味着，由于亲缘关系矩阵是基于标记的，因此整合了次要性状和基因组数据。有关交叉验证情境1（次要性状数据仅对训练集可用）和情境2（次要性状数据对训练集和测试集均可用）下的完整多变量BLUP方程，请参见附录C（Runcie 和 Cheng，2019）。

2.5. **递进比较**。 我们通过比较三个模型来展示纳入次要性状因子如何有助于提高预测能力（PA）；这些模型在模型复杂度和所用数据方面逐步增加（表 1）。第一个模型是单变量 gBLUP，仅使用标记数据和籽粒产量。第二个模型是拼接遗传潜在因子BLUP（glfBLUP），其因子基于时间点构建，方法是像 Melsen 等人（2025）那样拼接波长-时间点组合。此外，相关矩阵经过冗余过滤，保留

并对次要性状进行正则化。完整说明请参见 Melsen 等人（2025）。第三个模型是本研究所述的 glfBLUP，使用在各时间点构建的因子，并分别采用和不采用 Procrustes 旋转。我们比较了在渐进生长阶段（例如从营养生长期到抽穗期和灌浆期，包括这些阶段）可获得次要性状信息时各方法的表现，如 Krause 等人（2019）所示。我们使用 CIMMYT 数据集比较上述方法在 CV1（次要性状仅在训练集中可用）和 CV2（次要性状在训练集和测试集中均可用）下的预测准确度（PA）：将每个年份-试验组合子集划分为 2/3 的训练集和 1/3 的测试集，并进行 100 次重复。此外，我们展示因子的可解释性，以指出预测籽粒产量的关键时间点和波长。

TABLE 1. 所比较模型的概述、可获得数据的生育阶段，以及是否使用普罗克鲁斯特旋转。以基于标记和产量的单变量 gBLUP 作为基准，因此生育阶段不适用

模型	生育阶段	普罗克鲁斯特旋转
单变量 gBLUP	-	-
拼接 glfBLUP	营养生长期	-
拼接 glfBLUP	营养生长期、抽穗期	-
拼接 glfBLUP	all	-
每个时间点的 glfBLUP Varimax 因子	营养生长期	no
每个时间点的 glfBLUP Varimax 因子	营养生长期、抽穗期	no
每个时间点的 glfBLUP Varimax 因子	all	no
每个时间点的 glfBLUP Procrustes 因子	营养生长期	yes
每个时间点的 Procrustes 因子 glfBLUP	营养生长期、抽穗期	yes
每个时间点的 Procrustes 因子 glfBLUP	all	yes

3. RESULTS

我们使用 2014—2015年 CIMMYT 精英产量试验小麦数据集的一个子集，展示如何将 Procrustes 因子分析应用于对多个时间点采集的高光谱数据进行建模。此外，我们利用提取的因子得分，通过子集选择将高光谱数据与籽粒产量相关联，随后将选定的因子得分用于多变量基因组预测。我们将其预测能力（PA）与单变量 gBLUP 和多变量 gBLUP 的预测能力进行比较；后两者使用基于拼接次要性状（波长—时间点组合）的因子得分，方法与 Melsen 等人（2025）相同。

3.1. 干旱处理对籽粒产量的影响。我们发现，随着干旱条件加剧，籽粒产量呈下降趋势（图 3 b）；其中，最佳平坦处理的基因型平均籽粒产量最高，为 5.6 吨/公顷（t/ha），广义遗传力为 0.61；而重度干旱条件下，籽粒产量为 2.7 t/ha，广义遗传力为 0.39。同样，基于籽粒产量基因型均值计算的皮尔逊相关，在水分处理间差异越大时越低（图 3 a）。其中，最佳平坦处理与中度干旱处理之间的皮尔逊相关最高（0.46），而最佳平坦处理与重度干旱处理之间的相关最低（0.26）。

3.2. 高光谱潜在因子的解释。在进行 Procrustes 旋转之前，除 2015-03-23 和 2015-04-07 外，各时间点的载荷结构已基本一致；在这两个时间点中，最佳平坦和中度干旱处理内的载荷均在因子之间切换（图4）。具体而言，时间点中波长范围为 400 至 700 纳米的波长

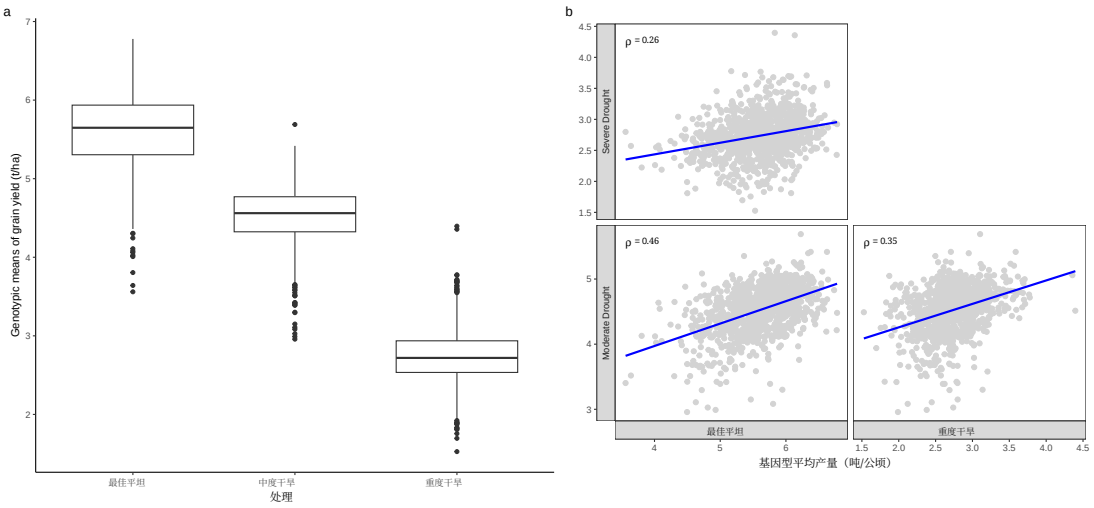

FIGURE 3. 基因型平均产量（最佳线性无偏估计量）的箱线图（a）和散点图（b），产量单位为吨/公顷（t/ha）；每个散点图坐标轴上标示相应处理。每对处理的皮尔逊相关系数以 ρ 表示。

在 2015-03-23 之前，波长往往在因子 1 上有较高载荷，在因子 2 上载荷较低；而 700 至 850 纳米的波长往往在因子 2 上有较高载荷，在因子 1 上载荷较低。2015-03-23 之后，这种情况反转。正如预期，Procrustes 旋转和平滑解决了这个问题（图4）。在大多数时间点，两个因子之间呈现出清晰的区别：因子 1 表示 400 至 700 纳米范围内的光谱，而因子 2 表示 700 纳米以上的波长。此外，在不同处理下，因子 2 中约 560 纳米处始终存在一个小峰值。另一个相似之处是，中度干旱和最优平坦处理的因子载荷均在最后一个时间点发生变化；中度干旱的最后一个时间点为3月23日，最优平坦处理的则为4月7日。我们发现，因子 2 在 400 至 600 纳米范围内的因子载荷有所增加。重度干旱处理未观察到这种变化，可能是因为干旱导致表型鉴定提前停止。由于不同处理下的因子结构相对相似，我们能够对各处理的因子得分进行相关性分析（图 5）。与基因型均值籽粒产量的情况一样，环境间因子得分均值的相关程度会随着处理之间相似性的降低而减小。例如，最佳平坦和中度干旱处理的因子 1 均值相关系数为 0.74，而与重度干旱处理的相关系数为 0.54。中度和重度干旱处理之间的相似性在因子 2 中最为明显，相关系数为 0.76。这揭示了因子得分 1 和 2 中保留的不同信息。

3.3. **子集选择**。 为探究无监督因子得分与产量之间的关系，我们进行了子集选择。用于交叉验证的 100 个数据集中，纳入的时间点因子各不相同。然而，在 2月9日和25日、3月10日和4月7日，各数据集均一致选出了同一组时间点因子（图 6）。这些日期对应抽穗期之后的时间点。令人惊讶的是，只有在最佳平坦处理下，最后一个时间点才始终被选为因子 1；而在中度和重度干旱处理下，任何因子都没有出现这种情况。这可能表明不同环境下的生理响应存在差异，例如干旱条件下衰老开始得更早

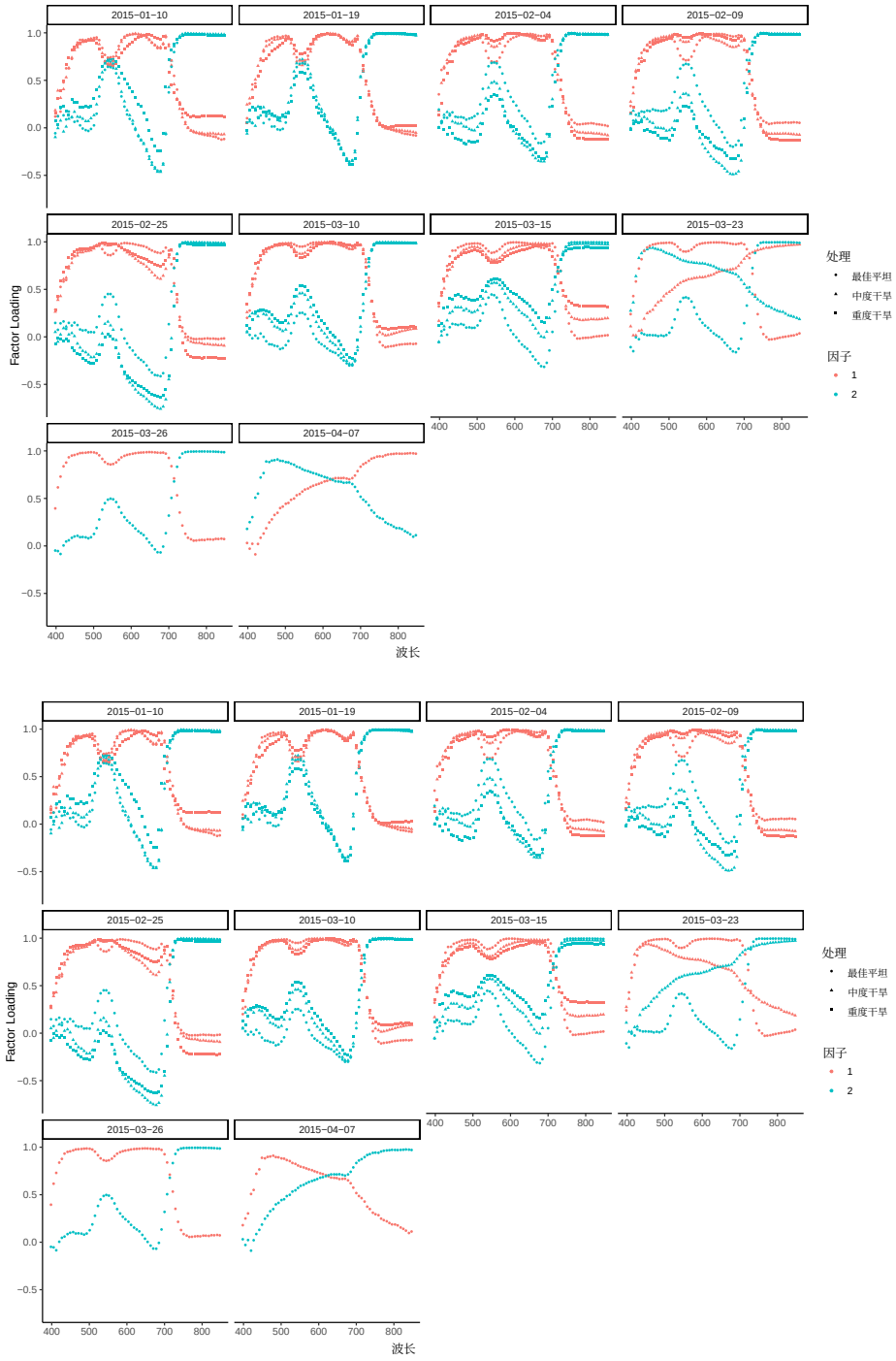

FIGURE 4. 每个时间点和处理下，高光谱反射率在因子 1 和因子 2 上的波长反射率载荷；普罗克鲁斯特旋转前（上图）和旋转后（下图）

。此外，在中度干旱情景下，抽穗期的时间点也常常仅纳入因子 1。

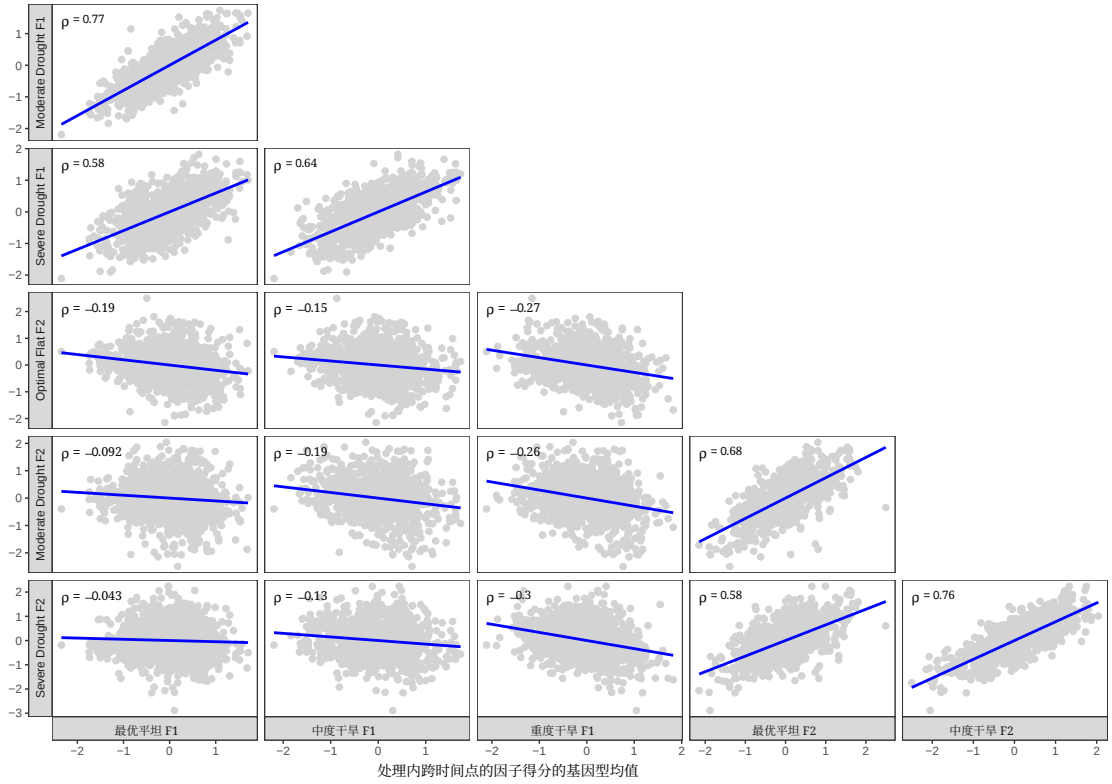

FIGURE 5. 每种处理内各因子得分的基因型均值的散点图和皮尔逊相关系数（ρ）。为提高可读性，坐标轴标签中的 Factor 缩写为 'F'

3.4. 不同情景下的多变量基因组预测。在交叉验证 2（CV2）情景下，所有处理均较基准单变量 gBLUP 有所改进（图 7）。在该情景中，训练集和测试集中均有次要性状。尽管在最优平坦处理中，采用拼接性状（波长-时间点）组合的 glfBLUP 表现更优，但在中度和重度干旱处理中并非如此。在考虑全部时间点时，在中度和重度干旱处理中，使用每个时间点的因子得分优于 gBLUP 以及采用拼接性状的 glfBLUP。然而，当仅能获得截至抽穗期（含抽穗期）的时间点时，这一优势便会消失。此外，我们还对经 Procrustes 旋转的因子得分拟合了样条函数，随后将提取的样条特征作为次要性状用于 glfBLUP（附录 B）。这并未提高预测准确度（PA）。提取样条特征这一额外步骤可能导致了信息损失。然而，平滑样条对于因子得分的可视化仍然有用，因为低产基因型在各时间点的次要性状响应可能有所不同（图 2）。在交叉验证 1（CV1）情景下，次要性状仅用于训练集，所有方法在不同处理中的预测准确度（PA）相近。

4. DISCUSSION

高维表型数据在大田作物植物育种研究中的应用日益普遍。因此，越来越需要开展将这些数据整合到遗传分析和基因组预测中的研究。在此，我们展示了一种计算上易于实现的方法，用于

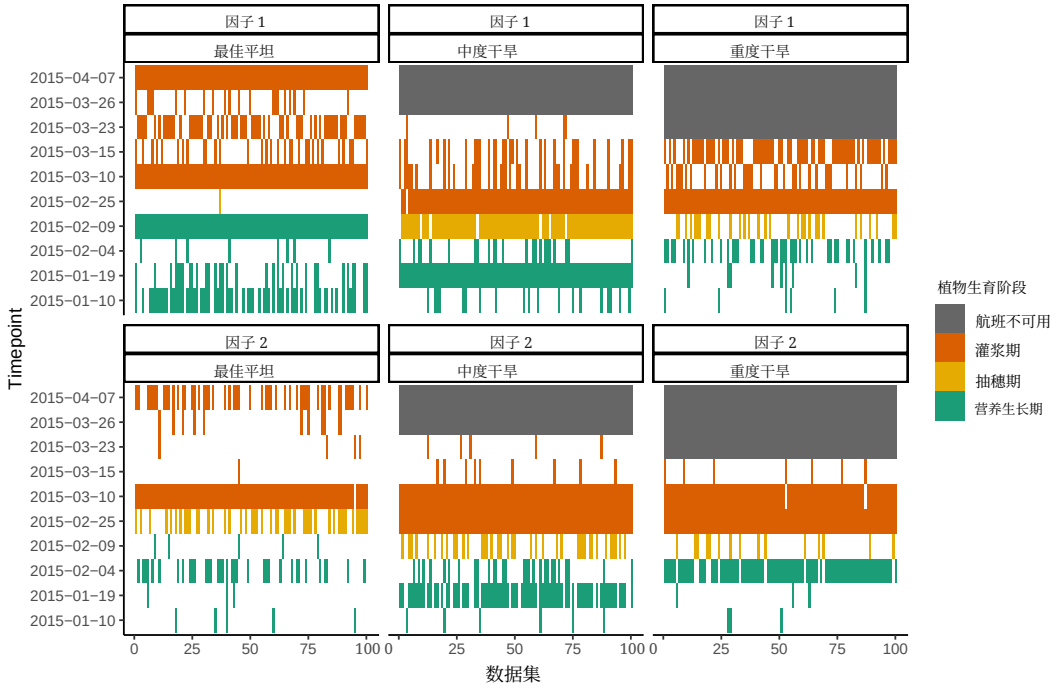

FIGURE 6. 在所有时间点均可用时，CV2 情景下子集选择后纳入的因子。垂直短划线表示该训练测试集中最终模型纳入的时间点因子。时间点采用年-月-日格式，植物生育阶段依据 Krause 等人（2019）。通过交叉验证创建的数据集，用于预测能力模型比较。

将时间序列高光谱表型数据纳入单个试验的多变量基因组预测中。尽管高光谱数据已广泛应用于植物科学领域，但我们提出了一种通用方法，可利用因子分析提取有意义的特征，并在整个植物生长周期中对其进行追踪，同时保留可解释性。这与当前在表型数据上应用因子分析的方法形成对比，后者尚未涉及多个时间点（Paix˜ao et al., 2022；Fialho etal., 2023），或难以解释（Runcie et al., 2021）。

高光谱数据通常会通过对特定波长应用函数进行转换，以提取植被指数，例如广为人知的归一化差异植被指数（NDVI）（Sellers，1985年）。这类指数广泛用于近距离测量与生物量相关的性状。然而，选择使用哪些指数或创建新指数并非易事，而且可能造成信息损失，因为这类方法只利用数据光谱中被认为相关的部分。例如，Krause 等人（2019年）指出，光谱中的蓝光区域也可能包含有价值的植物生理信息，而使用 NDVI 时这些信息会丢失。本研究通过对高光谱数据采用低维表示，并将其与目标性状（例如籽粒产量）相关联，让数据自行揭示规律。我们对高光谱时间序列数据的遗传相关矩阵进行因子分析，以提取具有生物学意义的特征。借助普罗克鲁斯特旋转，我们得以高效地保留时间序列动态。此外，我们还展示了这些特征在不同季节和处理条件下相对稳定。由于该方法具有普适性，它或许能揭示其他传感器数据中相关的特征和模式；这些数据尚未像高光谱数据那样得到广泛研究。

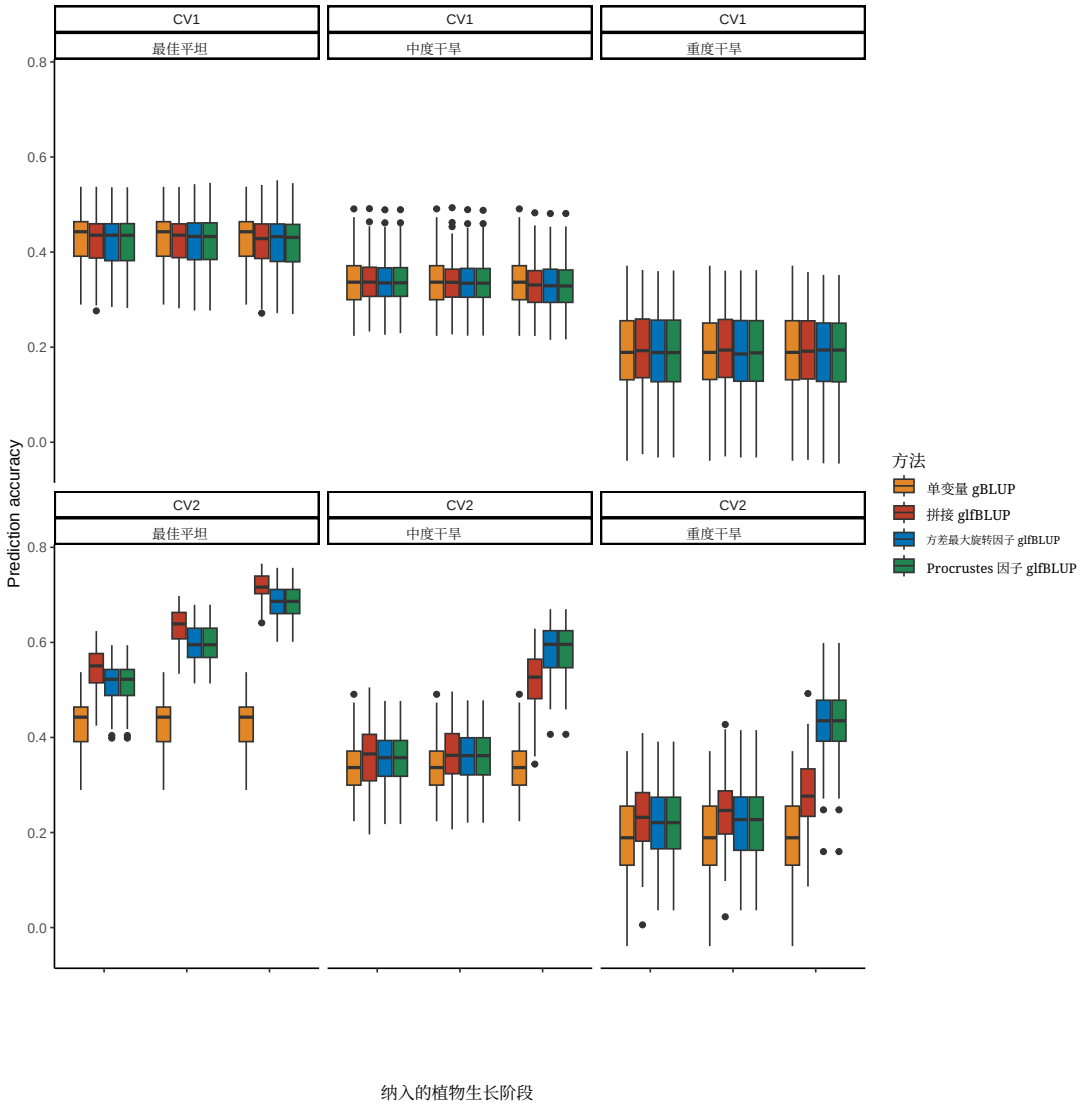

FIGURE 7. 每种模型类型下，针对每种交叉验证（CV）情景、处理和植物生育阶段，预测的基因型平均产量与观测的基因型平均产量之间的皮尔逊相关系数。营养生长期：仅纳入营养生长期的次要性状。营养生长期和抽穗期：纳入截至抽穗期（含抽穗期）的次要性状。全部：纳入所有可用植物生育阶段的次要性状。

尽管普罗克鲁斯特旋转能使因子在不同时间保持对齐，但它并不能保证每个因子在每个生育阶段都反映相同的生物学特征。小麦从营养生长期到抽穗期和灌浆期会发生显著变化，其高光谱信号也会随之改变。我们在较晚时间点发现的因子结构变化也暗示了这一点。解读这些因子的一个实用方法，是考察它们与 NDVI（参见 Melsen 等，2025年）等熟悉性状、冠层温度或叶绿素之间的关系，以及其关键波长如何随时间变化。通过进行子集选择，并将目标性状从籽粒产量改为熟悉的性状，即可实现这一点。如果这些模式符合已知的生理阶段——例如绿色程度达到峰值

在干旱条件下抽穗前或衰老前——那么这些因子很可能捕捉到了有意义的生物学变化。

按时间点进行因子分析，使我们得以了解哪些时间点对预测小麦籽粒产量至关重要。使用全部高光谱波长反射率，按时间点拟合因子模型，可采用标准估计技术，因为基因型数量（1033）远多于波长数量（62）。但是，如果波长数量超过基因型数量 $s > g$，仍可通过正则化步骤拟合因子模型，Melsen 等人（2025）便采用了这种方法。我们发现，在所有处理下，开花时间附近的因子得分均被选中。尽管开花时间是小麦中的一个重要指标已是普遍认知（Kole，2013），但我们在选择时间点时并未利用这一先验知识。因此，即使我们事先不了解开花时间的重要性，该分析也会揭示这一点。本文采用了原作者对开花时间的判定标准，即当 50% 的小区处于抽穗期时，判定为抽穗（Krause 等人，2019）。这一范围再次凸显了预测开花时间的重要性：开花时间会随处理而变化，同时又是籽粒产量的重要指标。由于开花时间等性状可能因群体而异，因此应在不同群体中研究变量选择的稳健性。

数量遗传学理论表明，遗传力较低的目标性状可以通过选择遗传相关且遗传力较高的次要性状得到改良（Thompson 和 Meyer，1986年；Runcie 和 Cheng，2019年）。这是开展多变量基因组预测的一个关键目标，因为次要性状信息有助于预测目标性状。迄今为止，这种方法主要已被证明有助于交叉验证情景 2（CV2），在该情景中，训练集和测试集基因型均具有次要性状信息。然而，如何纳入次要性状信息决定了预测能力（PA）的提升幅度。一种方法是在因子分析之前连接时间点和波长（Melsen 等，2025年）。在最优处理中，我们发现，基于此类连接的时间点-波长性状构建的因子进行多变量预测，其表现优于使用时间点因子进行的多变量预测。相比之下，在中度和重度干旱处理中，使用时间点因子进行多变量基因组预测可获得更高的预测能力（PA）。这种变化可能是因为，在针对每个时间点构建因子时，子集选择能够在相关时间点之间进行选择。而对于使用连接的时间点-波长性状构建的因子，情况并非如此，因为多个时间点可能载荷于同一个因子上，从而难以在时间点之间进行子集选择。

在本研究中，我们采用无监督因子分析进行降维，并通过最佳子集选择选取时间点因子，以便在多变量基因组预测中整合高维表型数据。也可以采用一种类似于 Arouisse 等人（2021）的方法。在这种情况下，所有因子得分都会对目标性状进行回归，随后得到的表型预测值将作为双变量基因组预测模型的输入。这样便无需像本研究中那样进行最佳子集选择。

该数据集中的飞行和测量频率较低，且间隔不规则。虽然这不会直接影响因子得分的使用，但可能会影响平滑样条等其他时间序列方法。本研究使用了完整样条的特征，因此选出的特征可能并未带来额外的预测准确度（附录 2）。加入先验作物生理学知识，或有助于从这些样条函数中提取并选择相关的发育阶段。

正如相关性分析所示，次级潜在性状有助于表征处理，为开展多环境分析开辟了道路。与利用产量数据进行相关性分析相比，基于高光谱数据的因子显示出处理间更强的相关性，因此可能更好地反映处理间的基因型响应。此外，正交因子之间（例如因子 1 与因子 2 之间）的相关性接近于零，表明这两个因子各自包含独特的信息。除此之外

将每个处理所考虑的性状从 62 个减少到 2 个潜在性状，除了上述解释之外，还有助于降低大型多环境试验分析的计算负担，使其更易于实现。随后可根据环境亲缘关系进行整合，但因子分析框架中的其他方案也可能很有前景（Costa-Neto 等人，2020）。

总之，因子分析是一种灵活的方法，可用于整合高维次要性状数据，并为多变量基因组预测构建具有可解释性的表征。结果展示了次要性状数据的可解释性，以及其在多变量基因组预测中的成功整合。因此，该方法有助于加深我们对植物育种试验中高维数据的理解。

数据与代码。用于生成高光谱数据集以及本文所报告结果的脚本可在 https://github.com/KunstJF/glfBLUP-Procrustes 获取。glf-BLUP《方法论》已在一个 R 包中实现，可通过 https://github.com/KillianMelsen/glfBLUP 获取。高光谱数据集可向 J. Crossa 合理申请获取

References

Antonio, O., L

´ ´ 2022 *Multivariate Statistical Machine Learning Methods for Genomic Prediction*

opez, M., Montesinos Lopez, A., and Crossa, J. ( ). . Springer, Cham, 瑞士。Araus, J. L., Slafer, G. A., Royo, C., and Serret, M. D. (2008). 谷类作物产量潜力与胁迫适应的育种。 *Critical Reviews in Plant Sciences*, 27(6):377–412. Arouisse, B., Theeuwen, T. P. J. M., Van Eeuwijk, F. A., and Kruijer, W. (2021). 利用高维次级表型改进基因组预测。 *Frontiers in Genetics*, 12:667358. Boer, M. P. (2023). 使用稀疏混合模型形式的张量积 P 样条。 *Statistical Modelling*, 23 5 6 465 479(-): – 。Costa-Neto, G., Fritsche-Neto, R., and Crossa, J. (2020). 非线性核、显性和环境分型数据提高了多环境试验中基于基因组的预测准确性。 *Heredity*, 126(1):92–106. Dos Santos, J. P. R., Fernandes, S. B., McCoy, S., Lozano, R., Brown, P. J., Leakey, A. D. B., Buckler, E. S., Garcia, A. A. F., and Gore, M. A. (2020). 用于生物量高粱发育性状基因组预测的新型贝叶斯网络。 *G3 Genes|Genomes|Genetics*, 10(2):769–781. Fialho, I. C., Azevedo, C. F., Nascimento, A. C. C., Teixeira, F. R. F., De Resende, M. D. V., and Nascimento, M. (2023). 在水稻中应用因子分析进行基因组预测，考察不同密度的标记面板。*Euphytica*, 219(9):88. Giannone, D., Lenza, M., and Primiceri, G. E. (2021). 利用大数据进行经济预测：稀疏性幻象。 *Econometrica*, 89(5):2409–2437. Jannink, J.-L., Lorenz, A. J., and Iwata, H. (2010). 植物育种中的基因组选择：从理论到实践。 *Briefings in Functional Genomics*, 9(2):166–177. Jarquin, D., Howard, R., Crossa, J., Beyene, Y., Gowda, M., Martini, J. W. R., Covarrubias Pazaran, G., Burgue˜no, J., Pacheco, A., Grondona, M., Wimmer, V., and Prasanna, B. M. (2020). 用于多环境试验的基因组预测增强型稀疏测试。 *G3 Genes|Genomes|Genetics*, 10(8):2725–2739. Jia, Y. and Jannink, J.-L. (2012). 多性状基因组选择方法提高遗传价值预测准确性。 *Genetics*, 192(4):1513–1522. J¨oreskog, K. (1967). 对最大似然因子分析的若干贡献。 *Psychometrika*. Kaiser, H. F. (1958). 因子分析中解析旋转的最大方差准则。 *Psychometrika*, 23(3):187–200. Kole, C., editor (2013). *Genomics and Breeding for Climate-Resilient Crops*. Springer Berlin Heidelberg, 柏林, 海德堡。Krause, M. R., Gonz´alez-P´erez, L., Crossa, J., P´erez-Rodr´ıguez, P., Montesinos-L´opez, O., Singh, R. P., Dreisi- gacker, S., Poland, J., Rutkoski, J., Sorrells, M., Gore, M. A., and Mondal, S. (2019). 用于小麦籽粒产量基因组预测的高光谱反射率衍生关系矩阵。 *G3 Genes|Genomes|Genetics*, 9 4 1231 1247 ( ): – 。Kruijer, W., Behrouzi, P., Bustos-Korts, D., Rodr´ıguez- ´Alvarez, M. X., Mahmoudi, S. M., Yandell, B., Wit, E., and van Eeuwijk, F. A. (2020). 具有直接和间接遗传效应的网络重建。*Genetics*, 214:781–807. Lopez-Cruz, M. and De Los Campos, G. (2021). 使用稀疏选择指数进行最佳育种值预测。*Genetics*, 218(1):iyab030. Lopez-Cruz, M., Olson, E., Rovere, G., Crossa, J., Dreisigacker, S., Mondal, S., Singh, R., and Campos, G. D. L. (

2020 *Scientific*). *Reports* 10 1 8195

使用高光谱图像数据进行育种值预测的正则化选择指数。, ( ): 。

Manthena, V., Jarqu´ın, D., Varshney, R. K., Roorkiwal, M., Dixit, G. P., Bharadwaj, C., and Howard, R. (2022). 评估基因组预测的降维方法。 *Frontiers in Genetics*, 13. Melsen, K. A. C., Kunst, J. F., Crossa, J., Krause, M. R., van Eeuwijk, F. A., Kruijer, W., and Peeters, C. F. W. (2025). 利用高维次级表型改进基因组预测：遗传潜因子方法。 *Biometrical Journal*, 67(5). Michel, S., L¨oschenberger, F., Ametz, C., Pachler, B., Sparry, E., and B¨urstmayr, H. (2019). 通过多性状基因组选择综合利用籽粒产量、蛋白质含量和蛋白质品质，进行面包小麦育种*Theoretical and Applied Genetics* 132 10 2767 2780 , ( ): – 。Paix˜ao, P. T. M., Nascimento, A. C. C., Nascimento, M., Azevedo, C. F., Oliveira, G. F., Da Silva, F. L., and Caixeta, E. T. (2022). 在中粒咖啡育种中应用因子分析进行基因组选择研究。*Euphytica*, 218(4):42. Peeters, C. F. W. (2012). *Bayesian Exploratory and Confirmatory Factor Analysis*. 博士论文，乌得勒支大学。Peeters, C. F. W., ¨ Ubelh¨or, C., Mes, S. W., Martens, R., Koopman, T., Graaf, P. d., Velden, F. H. P. v., Boellaard, R., Castelijns, J. A., Beest, D. E. t., Heymans, M. W., and Wiel, M. A. v. d. (2019). 基于影像组学数据的稳定预测。arXiv:1903.11696 [stat]. Piepho, H.-P. (1998). 在品种试验中使用因子分析方差—协方差结构进行经验最佳线性无偏预测： *Theoretical and Applied Genetics*, 97(1-2):195–201. P´erez-Valencia, D. M., Rodr´ıguez- ´Alvarez, M. X., Boer, M. P., Kronenberg, L., Hund, A., Cabrera-Bosquet, L., Millet, E. J., and Eeuwijk, F. A. V. (2022). 高通量表型数据时空分析的两阶段方法。 *Scientific Reports*, 12(1):3177. Raˆıche, G., Walls, T. A., Magis, D., Riopel, M., and Blais, J.-G. (2013). 卡特尔碎石检验的非图形解法。 *Methodology*, 9(1):23–29. Runcie, D. and Cheng, H. (2019). 多性状基因组预测方法交叉验证中的陷阱与补救措施。 *G3 Genes|Genomes|Genetics*, 9(11):3727–3741. Runcie, D. E., Qu, J., Cheng, H., and Crawford, L. (2021). MegaLMM：用于数千个性状基因组预测的超大规模线性混合模型。 *Genome Biology*, 22(1):213. Schwarz, G. (1978). 估计模型的维数。 *The Annals of Statistics*, 6(2):461–464. Sch¨ onemann, P. H. (1966). 正交普罗克鲁斯特问题的广义解法。 *Psychometrika*, 31(1):1– 10. Sellers, P. J. (1985 *International Journal of Remote*). *Sensing* 6 8 1335 1372 冠层反射率、光合作用与蒸腾作用。, ( ): – 。Thompson, R. and Meyer, K. (1986). 多性状选择育种值估计理论方面综述。 *Livestock Production Science*, 15(4):299–313. Thomson, G. (1939). *The factorial analysis of human ability*. 伦敦大学出版社，伦敦，英国。Van Dijk, A. D. J., Kootstra, G., Kruijer, W., and De Ridder, D. (2021). 植物科学与植物育种中的机器学习。 *iScience*, 24(1):101890. Wartha, C. A. and Lorenz, A. J. (2021). 公共部门植物育种计划中基因组选择的实施：现状与机遇。 *Crop Breeding and Applied Biotechnology*, 21(spe):e394621S15.

5. Statements and Declarations

致谢。 我们感谢Margaret Krause和CIMMYT生成并发布原始数据。此外，我们感谢生产生态学与资源保护研究生院（PE&RC）通过PE&RC研究生项目2022年项目征集资助本研究，项目编号为22145。

经费。 本研究由生产生态学与资源保护研究生院（PE&RC）资助。

利益冲突。 作者声明不存在需要披露的相关财务或非财务利益。

作者贡献。 概念化：C.F.W.P. 和 J.F.K；方法论与软件：K.A.C.M. 和 J.F.K.；撰写—初稿：J.F.K；撰写—审阅与编辑：J.F.K.、K.A.C.M.、W.K.、J.C.、M.R.K.、C.M.、F.A.E.、C.F. W.P.。所有作者均已阅读并同意发表此稿件版本。

Appendix A. Factor dimension determination

进行因子分析时，需要确定因子数量（即维数）。上限可通过加速因子确定（Raˆıche et al., 2013），这是一种替代视觉碎石图拐点检测的方法，其中维数取在加速度最高点之前的位置。这意味着根据最后一个特征值和 $(i+1)^{th}$ 的最优坐标进行线性回归，并计算特征值 δ_j 的二阶导数：

$$\hat{\delta}_j = a_{j+1} + b_{j+1}(j),$$

其中， a 是截距， b 是斜率

$$b_{j+1} = \frac{\delta_s - \delta_{j+1}}{s-j-1} \text{ and } a_{j+1} \equiv \delta_{j+1}(j+1) - b_{j+1}(j+1).$$

二阶导数近似如下：

$$f''(j) = f(j+1) - 2f(j) - f(j-1)$$

其中， $f''(j)$ 是特征值 (δ) j 的二阶有限差分。同时，我们采用了下界 Kaiser准则，其中 $k \geq 1$, 从而得到如下选择： m^* 作为 m ，如 (Raˆıche et al., 2013) 所述：

$$m^* = \sum_j I(\delta_j \geq 1 \wedge j < s') \text{ with } s' \equiv \arg\max_j (af_j).$$

Appendix B. Extracting and using spline parameters

估算因子得分后，我们将 s 次要性状的维数降至 m^* 个因子，这些因子适合通过平滑样条对时间序列进行建模（P´erez-Valencia et al.，2022）。利用 statgenHTP软件包拟合群体和基因型特异性平滑样条，这得益于 LMMsolver软件包的高效计算（Boer，2023）。每个小区 i、基因型 c和群体π 的 Procrustes平滑因子得分 $\hat{}\ \xi^l_{\pi ci}$ 均可通过分层平滑样条在时间点 l 上进行建模：

$$\hat{\xi}_{\pi ci}(l) = \phi_\pi(l) + \phi_{\pi c}(l) + \phi_{\pi ci}(l) + \epsilon_{\pi ci}(l), \quad \epsilon_{\pi ci}(l) \sim \mathcal{N}_n\left(0, \sigma^2(l)\right),$$

其中，群体 π 样条为 $\phi_\pi(l)$，基因型 c 特异性样条为 $\phi_{\pi c}(l)$，小区i 特异性样条为 $\phi_{\pi ci}(l)$。基因型特异性样条可视为相对于群体样条的偏差。为利用这些样条辅助产量预测，我们针对每个因子提取基因型样条的关键参数。例如，营养生长期与籽粒灌浆期之间区间的曲线下面积，因为此时可能观察到较大的基因型方差。其他参数提取区间可通过目视检查基因型特异性样条来确定。随后将这些特征用作基因组预测中的次要性状。然而，在多变量 gBLUP方程中，以样条特征替代因子得分作为次要性状并未观察到明显益处（图 8）。

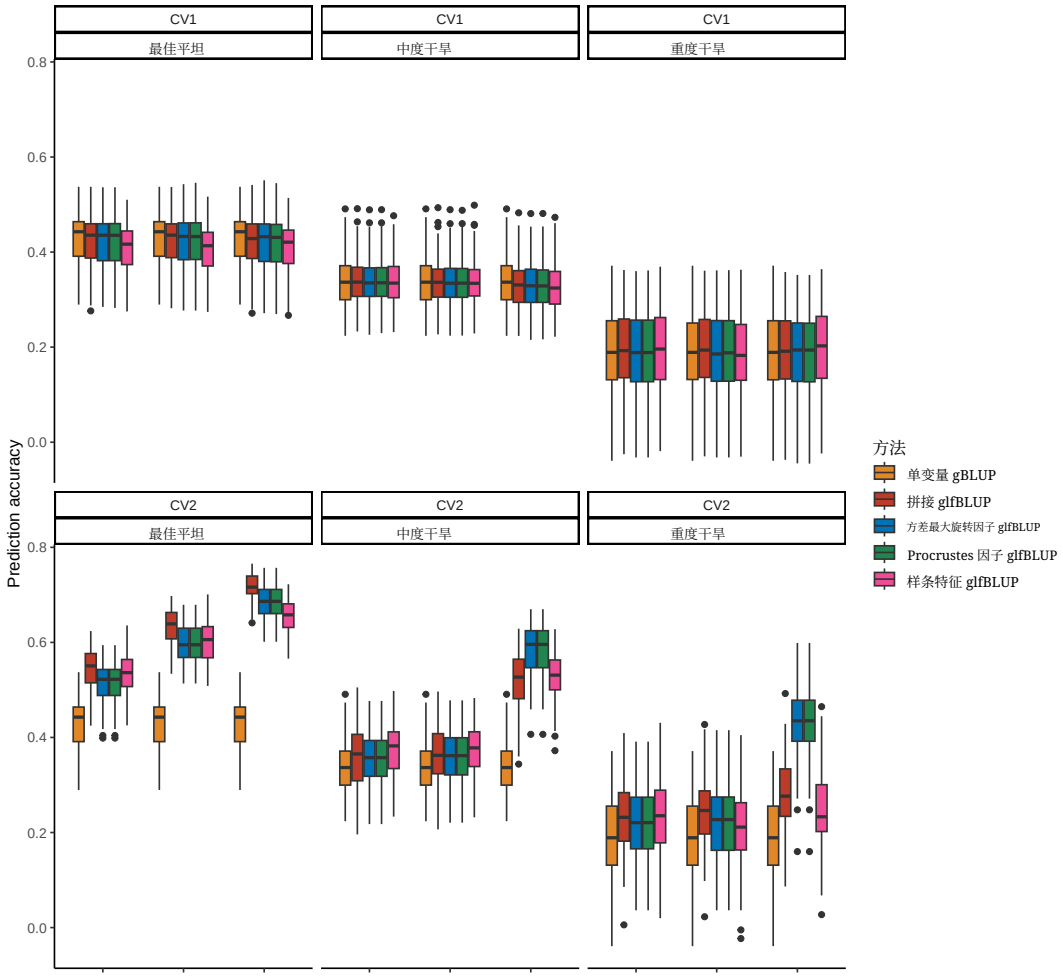

FIGURE 8. 各模型类型下，每种交叉验证（CV）方案、处理和植物生育阶段中，按基因型计算的预测平均产量与观测平均产量之间的皮尔逊相关系数。营养生长期：仅纳入营养生长期阶段的次要性状。营养生长期与抽穗期：纳入截至抽穗期（含抽穗期）的次要性状。全部：纳入所有可用植物生育阶段的次要性状。

Appendix C. Univariate and multivariate BLUP equations

我们给出一般的单变量和多变量最佳线性无偏预测方程，其中 f 表示目标性状， w 表示次要性状，下标 o 表示训练集（例如，已观测的表型），下标 u 表示测试集（例如，未观测的表型）。因此，我们将训练集-测试集划分后的表型矩阵表示为：

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_o \\ \mathbf{Y}_u \end{bmatrix}.$$

首先，我们介绍目标性状 $\mathbf{y}^f$ 的单变量 gBLUP，然后将其扩展为适用于 CV1 和 CV2 的多变量 gBLUP。对于单变量 gBLUP，我们使用以下符号：

$$\begin{aligned}
\hat{\mathbf{V}} &= \hat{\sigma}^f_G \otimes \mathbf{K}_o + \hat{\sigma}^f_E \otimes \mathbf{I}_n \\
\hat{\mathbf{g}}^{(uni)}_{fo} &= (\hat{\sigma}^f_G \otimes \mathbf{K}_o)\hat{\mathbf{V}}^{-1}\big(\mathbf{y}^f_o - \mathbf{X}_o\hat{\boldsymbol{\beta}}^f]\big) \\
\hat{\mathbf{g}}^{(uni)}_{fu} &= \mathbb{E}(\mathbf{g}_{fu}|\mathbf{y}_{fo}) \\
&= (\hat{\sigma}^f_G \otimes \mathbf{K}_{uo})\hat{\mathbf{V}}^{-1}\big(\mathbf{y}^f_o - \mathbf{X}_o\hat{\boldsymbol{\beta}}^f\big) \\
&= \mathbf{K}_{uo}\mathbf{K}_o^{-1}\hat{\mathbf{g}}^{(uni)}_{fo},
\end{aligned}$$

其中，$\hat{}\mathbf{g}^{(uni)}_{fo}$ 和$\hat{}\mathbf{g}^{(uni)}_{fu}$ 分别是训练集和测试集的目标性状单变量最佳线性无偏预测值。另请注意，由于$\hat{}\sigma^f_G$和$\hat{}\sigma^f_E$是标量，因此$\hat{}\mathbf{V} = \hat{}\sigma_{Gf}\mathbf{K}_o + \hat{}\sigma_{Ef}\mathbf{I}_n$。我们将单变量 gBLUP 模型扩展为多变量版本。这里，我们使用根据目标性状（f）和次要性状（w）的数据（即因子得分）估计的遗传（$\boldsymbol{\Sigma}_G$）和残差（$\boldsymbol{\Sigma}_E$）协方差矩阵：

$$\begin{aligned}
\boldsymbol{\Sigma}_G &= \begin{bmatrix} \boldsymbol{\Sigma}^w_G & \boldsymbol{\sigma}^{wf}_G \\ \boldsymbol{\sigma}^{fw}_G & \sigma^f_G \end{bmatrix} = \begin{bmatrix} \boldsymbol{\sigma}^{w\cdot}_G \\ \boldsymbol{\sigma}^{f\cdot}_G \end{bmatrix} \\
\boldsymbol{\Sigma}_E &= \begin{bmatrix} \boldsymbol{\Sigma}^w_E & \boldsymbol{\sigma}^{wf}_E \\ \boldsymbol{\sigma}^{fw}_E & \sigma^f_E \end{bmatrix} = \begin{bmatrix} \boldsymbol{\sigma}^{w\cdot}_E \\ \boldsymbol{\sigma}^{f\cdot}_E \end{bmatrix}.
\end{aligned}$$

在 CV1 情况下（测试集中未观测到次要性状）：

$$\begin{aligned}
\hat{\mathbf{V}} &= \hat{\boldsymbol{\Sigma}}_G \otimes \mathbf{K}_o + \hat{\boldsymbol{\Sigma}}_E \otimes \mathbf{I}_n \\
\hat{\mathbf{g}}^{(CV1)}_{fo} &= (\boldsymbol{\sigma}^{f\cdot}_G \otimes \mathbf{K}_o)\hat{\mathbf{V}}^{-1}\,\mathrm{vec}\left(\mathbf{Y}_o - \mathbf{X}_o\hat{\boldsymbol{\beta}}\right) \\
\hat{\mathbf{g}}^{(CV1)}_{fu} &= \mathbb{E}(\mathbf{g}_{fu}|\mathbf{Y}_o) \\
&= \left(\boldsymbol{\sigma}^{f\cdot}_G \otimes \mathbf{K}_{uo}\right)\hat{\mathbf{V}}^{-1}\,\mathrm{vec}\left(\hat{\mathbf{Y}}_o - \mathbf{X}_o\hat{\boldsymbol{\beta}}\right) \\
&= \mathbf{K}_{uo}\mathbf{K}_o^{-1}\hat{\mathbf{g}}^{(CV1)}_{fo},
\end{aligned}$$

$\hat{}$其中 $\mathbf{Y}_o$ –$\mathbf{X}_o\boldsymbol{\beta}$ 已向量化。CV2 情景下的多变量预测（测试集中观测到次要性状）可直接获得（Arouisse 等，2021），也可通过两步法获得（Runcie 和 Cheng，2019）。我们倾向于采用两步法，因为该算法可用于获得全部 CV1 训练集最佳线性无偏预测，$\hat{}\mathbf{G}^{(CV1)}_o$:

$$\begin{aligned}
\hat{\mathbf{V}} &= \hat{\boldsymbol{\Sigma}}_G \otimes \mathbf{K}_o + \hat{\boldsymbol{\Sigma}}_E \otimes \mathbf{I}_n \\
\hat{\mathbf{G}}^{(CV1)}_o &= \mathrm{vec}^{-1}(\hat{\boldsymbol{\Sigma}}_G \otimes \mathbf{K}_o)\hat{\mathbf{V}}^{-1}\,\mathrm{vec}(\mathbf{Y}_o - \mathbf{X}_o\hat{\boldsymbol{\beta}}).
\end{aligned}$$

然后，我们将这些训练集最佳线性无偏预测转换为中间的 CV1 测试集最佳线性无偏预测，并将其与测试集的观测次级特征相结合，最终获得测试集目标性状的 CV2 最佳线性无偏预测：

测试集目标性状：

$$\begin{aligned}\hat{\mathbf{V}} &= \hat{\boldsymbol{\Sigma}}_G \otimes \mathbf{K}_u^{-1} + \hat{\boldsymbol{\Sigma}}_E \otimes \mathbf{I}_n \\ \hat{\mathbf{g}}_{fu}^{(CV2)} &= \mathbb{E}(\mathbf{g}_{fu}|\mathbf{Y}_{wu}, \mathbf{Y}_o) \\ &= \mathbf{K}_{uo}\mathbf{K}_o^{-1}\hat{\mathbf{g}}_{fo}^{(CV1)} + (\boldsymbol{\sigma}^{wf} \otimes \mathbf{K}_u^{-1})\hat{\mathbf{V}}^{-1}\,\mathrm{vec}(\mathbf{Y}_{wu} - \mathbf{K}_{uo}\mathbf{K}_o^{-1}\hat{\mathbf{G}}_o^{(CV1)}).\end{aligned}$$