arXiv:2603.24820v1 [stat.ME] 25 Mar 2026

稳健双块同步降维

Sven Serneels[1,2]

[1] SiC Systems，美国田纳西州富兰克林[2] 比利时安特卫普大学数学系

摘要

本文介绍稳健双块（RTB）同步降维，这是首种能够对两个变量块进行同步降维的统计稳健方法，并允许分别精细调整每个块的模型复杂度。本文提出了该新方法的稠密版本和稀疏版本。稀疏 RTB 是首个能够分别为每个块选择模型复杂度和稀疏程度的稳健估计量。因此，即使存在离群值，RTB 也能从每个数据块中最优地提取并概括相关信息。作为推论，这些估计量可以重新组合为多元回归的单一回归系数估计量，即使每个块中的变量数超过样本数，该估计量仍然适用。大规模模拟研究表明，新方法能够抵抗不同类型的离群值，同时在各种维度设置下保持估计效率。稠密方法和稀疏方法均呈现出这些特性。本文还通过两个示例数据集进一步展示了这些方法的性能，并提出了一种简便的算法，将其发布在开源代码库中。

关键词： 变量选择、多元稳健回归、稳健双块降维

1 引言

随着数据维数不断增加，将信息压缩为紧凑的低维表示变得愈发重要。当数据由两个共同观测的变量块构成时，这种压缩应保留块间关系：降维空间应捕捉从一个块建模另一个块所需的全部信息。同步双块充分降维（SDR）正好能够实现这一点，为两个块的联合结构提供有理论依据且简约的概括。

形式上，设$\mathbf{X} \in \mathbb{R}^{n\times p}$和$\mathbf{Y} \in \mathbb{R}^{n\times q}$分别表示由 n 个 p维自变量和 q维因变量的观测值构成的矩阵。SDR旨在寻找满足以下条件的潜变量$\mathbf{T} = \mathbf{XW}$

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{T}, \tag{1}$$

其中，满足条件的子空间称为中心子空间，而 ⫫表示统计独立。在中心子空间的各种估计量中，偏最小二乘（PLS）是历史最悠久且计算效率最高的方法之一，其理论基础已得到充分确立。具体而言，Cook 等人(2013)推导出，当$\mathbf{Y}$ 为单变量时，PLS1（单变量响应变量）可为中心子空间 $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ 提供一致估计量。其之所以能够实现这一点，是因为PLS1的权重向量张成了$\mathbf{\Sigma}_{\mathbf{XX}}\mathbf{\Sigma}_{\mathbf{XY}}$的克雷洛夫子空间 $\mathbf{\Sigma}^{-1}$ ，该子空间在温和条件下收敛至 $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$。

方程（1）仅对$\mathbf{X}$ 进行降维。更丰富的目标是同时对两个块进行降维，寻找线性组合 $\mathbf{U} = \mathbf{YV}$ 和 $\mathbf{T} = \mathbf{XW}$ ，使得

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{T}, \tag{2a}$$
$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{U}. \tag{2b}$$

由于难以直接确立独立性，因此将这些条件放宽为零协方差。令 $\mathbf{P} = \mathbf{X}^T\mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}$和 $\mathbf{Q} = \mathbf{Y}^T\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}$分别表示 X载荷和 Y载荷，残差补空间为 $\mathbf{E} = \left(\mathbf{I}_n - \mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\right)\mathbf{X}$ 和 $\mathbf{F} = \left(\mathbf{I}_n - \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\right)\mathbf{Y}$，则对两个块进行同步降维要求

$$\operatorname{cov}(\mathbf{E}, \mathbf{F}) = \mathbf{0}, \tag{3a}$$
$$\operatorname{cov}(\mathbf{UQ}^T, \mathbf{F}) = \mathbf{0}, \tag{3b}$$
$$\operatorname{cov}(\mathbf{TP}^T, \mathbf{E}) = \mathbf{0}. \tag{3c}$$

偏最小二乘（PLS）回归也是多元回归情境下广泛使用的一种同步降维方法。人们常常不加说明地假定，多元偏最小二乘（也称为 PLS2）能为（3）中的两个中心子空间提供估计值。然而，尽管 Cook 等人 (2013)证明 PLS1得分空间是中心子空间 $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$的一个估计量，Cook 等人 (2023)表明，这一结论并不适用于 PLS2（多元响应变量）。PLS2 仅对 $\mathbf{X}$ 块进行降维，同时将$\mathbf{Y}$ 视为固定，因此无法估计联合的双块中心子空间（2）。这促使他们开发了双块同步降维 （以下简称 twoblock，也称为“XY-PLS”），该方法同时从预测变量块$\mathbf{X}$ 和响应变量块 $\mathbf{Y}$中提取潜在成分，为双块 SDR 问题提供估计量。该方法分两个连续的消胀阶段进行：首先从 h_x 中提取成$\mathbf{X}$ 分，使其与$\mathbf{Y}$ 的协方差最大；然后从 h_y 中提取成分$\mathbf{Y}$ ，使其与$\mathbf{X}$的协方差最大。该方法为两个块提供真正的同步 SDR 估计值，并可直接估计回归系数。研究表明，其性能优于 PLS2 和各种单变量回归变体。

作者（Serneels，2025年）最近提出了稀疏双块降维，这是原始方法的稀疏扩展：在每次消去步骤中对权重向量进行软阈值处理。这使得能够在$\mathbf{X}$ 和$\mathbf{Y}$块中进行变量选择，从而得到具有竞争力预测性能且易于解释的模型。除了可作为稀疏PLS2和稀疏典型相关分析（CCA）的灵活替代方法外，这也是首个能够分别精细调节模型复杂度以及各变量块稀疏程度的方法。

稠密和稀疏双块方法都依赖经典的位置和形状估计，因此容易受到异常观测的影响。在化学计量学及其他应用领域，数据中经常包含非典型案例，例如预测变量空间中的杠杆点、响应变量中的垂直离群值，或两者兼有。这些非典型案例的存在会严重扭曲估计子空间和回归系数。有鉴于此，本文提出了稳健双块（RTB）降维。本文提出的RTB算法采用源自M估计的迭代重加权方法。有关回归情境下稳健估计的详细综述，请参见例如Maronna等人（2019年）。作者此前已在偏最小二乘领域成功应用这种使降维和回归算法稳健化的方法，由此提出了如今广泛采用的偏鲁棒M回归（PRM，Serneels等人（2005））及其稀疏扩展SPRM（Hoffmann等人，2015）。本文提出的算法以这两种稳健估计量为基础，并引入了适用于双块情境的多变量重加权方案。根据各自得分空间中的马氏型距离，分别计算$\mathbf{X}$ 和$\mathbf{Y}$ 块的个案权重，然后在估计双块分解时用这些权重降低非典型案例的权重，或将其剔除。由此得到的方法是首个稳健双块同步降维方法，这一说法同时适用于稠密和稀疏变体。

本文其余部分安排如下。第2节介绍新的 RTB 方法，以及一种易于实现的算法。第3节展示一项模拟研究，评估 RTB 在各种污染和维数情景下的表现。第4 节通过两个已知适合采用多元估计且包含离群值的真实数据集，展示该方法的应用。最后，第5 节总结全文，并展望未来进一步研究的方向。

2 方法

在介绍新的稳健估计量之前，我们将简要概述现有的非稳健替代方法。

2.1 稠密和稀疏 twoblock 维数约简

设预测变量矩阵$\mathbf{X}$ 和被预测变量矩阵$\mathbf{Y}$ 均已中心化并标准化。双块方法（Cook 等人，2023）随后依次提取成分。从本质上说，该方法采用受 NIPALS（Wold，1966年）启发的算法，对原始被预测变量和预测变量数据计算潜变量，从而提取$\mathbf{X}$ 的潜在空间；随后再次应用该算法，此时预测变量与被预测变量的角色互换。最终，将两组估计值重新组合为一组回归系数。

算法总结见算法 1，其中 $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_{h_x}]$ 和 $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_{h_y}]$ 分别收集了$\mathbf{X}$ 块和$\mathbf{Y}$ 块的权重向量。

在稀疏变体（Serneels，2025年）中，每个权重向量都经过软阈值化（算法1中的步骤7和14），而 η_x，$\eta_y \in [0, 1)$ 则控制相应块的稀疏程度。当 $\eta_x = \eta_y = 0$时，即可得到稠密双块算法。在算法中及其他地方，$\odot$ 表示哈达玛（逐元素）矩阵乘积。

算法1 双块降维 要求：X $\in \mathbb{R}^{n\times p}$、Y $\in \mathbb{R}^{n\times q}$、成分 h_x, h_y1：对 X $\to$ X_0，和 Y $\to$$Y_0$ 进行中心化和尺度化 2：$\tilde{X}^{(1)} \leftarrow X_0$；$\tilde{Y}^{(1)} \leftarrow$ Y_0— X 块消去 — 3：对于 $i = 1, \ldots, h_x$，执行 4：$S_{XY}^{(i)} \leftarrow \tilde{X}^{(i)\top}$ Y_0 / n5：$w_i \leftarrow$ S 的主左奇异向量$_{XY}^{(i)}$6：如果为稀疏，则 $w_i \leftarrow$ $\text{sign}(w_i) \odot \max\left(|w_i| - \eta_x \max|w_i|, 0\right)$7：结束如果 8：$t_i \leftarrow \tilde{X}^{(i)} w_i$；$p_i \leftarrow \tilde{X}^{(i)\top} t_i / (t_i^\top t_i)$ 9：$\tilde{X}^{(i+1)} \leftarrow \tilde{X}^{(i)} - t_i p_i^\top$10：结束循环 — Y 块消去 — 11：对于 $j = 1, \ldots, h_y$，执行 12：$S_{YX}^{(j)} \leftarrow \tilde{Y}^{(j)\top}$ X_0 / n13：$v_j \leftarrow$ S 的主左奇异向量$_{YX}^{(j)}$14：如果为稀疏，则 $v_j \leftarrow$ $\text{sign}(v_j) \odot \max\left(|v_j| - \eta_y \max|v_j|, 0\right)$15：结束如果 16：$u_j \leftarrow \tilde{Y}^{(j)} v_j$；$q_j \leftarrow \tilde{Y}^{(j)\top} u_j / (u_j^\top u_j)$ 17：$\tilde{Y}^{(j+1)} \leftarrow \tilde{Y}^{(j)} - u_j q_j^\top$18：结束循环 — 回归系数 — 19：$\hat{B}_{\text{scaled}} \leftarrow W(W^\top X_0 W)^{-1}(W^\top X_0)^\top(X_0^\top Y_0 V)V^\top$ ^ ^ 20：将 B_{scaled} 重新缩放至原始数据尺度 $\to$B ^21：返回 B、W、V、T、U、P、Q

2.2 RobustTwoblock（RTB）

为了使 twoblock 算法实现鲁棒化，我们选择了迭代重加权方法。此新方法基于迭代重加权在单变量偏最小二乘中的成功应用，后者催生了稠密（Serneels 等人，2005年）和稀疏（Hoffmann 等人，2015年）偏最小二乘部分鲁棒 M 回归估计量。虽然也可以探索其他鲁棒化途径，但迭代重加权具有多方面的优点：计算上易于处理，已证明能够抵御多种偏离正态分布的情况，并且能够得到适用于变量数多于样本数情形的估计量。在化学计量学和生物信息学等多个应用领域，可能会遇到 $p > n$， $q > n$或两者兼有的数据。该方法必须能够处理此类数据，这一前提排除了其他看似可行的方法，例如采用基于子集选择的估计值，如 MCD（Rousseeuw，1985年）或 LTS（Rousseeuw，1984年）。同样，广义空间符号预处理（Serneels 等人，2024年）最初是计算广义空间符号协方差矩阵（Raymaekers 和 Rousseeuw，2019年）的一个步骤，但无法应用于变量数多于样本数的数据。

考虑 r 是产生可操作估计量的前提，当 $p > n$ o r

$q > n$，RTB 将双块估计纳入迭代重加权方案。其核心思想是根据各自得分空间中的离群程度分配案例权重，并在加权数据上重新估计模型，直至收敛。

初始化。在迭代重加权和M估计的背景下，众所周知，估计值最终的稳健性很大程度上取决于初始值的选择：如果初始值在存在离群值时本身会失效，那么重加权方法仍可能失效。这始于数据的中心化和尺度化。理想情况下，数据应以高度稳健、且位于数据凸包内的位置估计为中心。同样，也应选择高度稳健的尺度估计量。这两个预处理步骤的良好选择是 ℓ_1 中位数位置估计量（有关其计算方面的讨论，参见例如Fritz等人(2012年）和 τ_2 尺度估计量（Maronna和Zamar，2002年），不过也可以采用其他方案。如果计算效率至关重要，则可以选择逐列中位数和中位数绝对偏差。

迭代重加权的另一个重要选择是权重函数。可以采用连续加权函数，例如Fair函数。然而，Fair函数会降低分布中位置不完全相同的所有个案的权重，因此所得权重可能难以解释；由于对非离群个案的降权过度，其预测性能也可能不尽如人意。非连续降权函数则设定一个截断值，超过该值的个案会被视为离群值，并开始降权；这类函数通常易于实现，所得个案权重具有很强的可解释性，也能带来良好的统计稳健性。常见的非连续降权函数包括Huber函数和Hampel函数。本文的模拟与实例部分均采用后者。Hampel函数定义如下：

$$\psi_H(d_i) = \begin{cases} 1 & \text{if } d_i \le c_1, \\ \dfrac{c_1}{d_i} & \text{if } c_1 < d_i \le c_2, \\ \dfrac{c_1\,(c_3 - d_i)}{d_i\,(c_3 - c_2)} & \text{if } c_2 < d_i \le c_3, \\ 0 & \text{if } d_i > c_3, \end{cases} \tag{4}$$

其中，情形 $d_i \in \mathbb{R}^+$ 可以是任意正标量，但在下文中，通常根据数据通过某种距离度量计算得出。Hampel函数依赖于三个截断值，这些值会根据相应情形的偏离程度，逐步加大降权力度。这些截断值是算法中可调的参数，可根据相应的分位数进行设定，例如距离平方的 χ^2 分布的分位数。此前，这种方法已成功应用于相关领域，例如 SPRM回归或逐单元稳健 M 回归（Filzmoser等人，2020）。

初始权重的计算过程在算法 2中作了形式化说明。该过程分别独立应用于 $\mathbf{X}$ 块和 $\mathbf{Y}$ 块。

接下来，RTB初始化步骤如下：

1. 对$\mathbf{X} \to \mathbf{X}_s$ 和$\mathbf{Y} \to \mathbf{Y}_s$进行稳健中心化和尺度化。

2. $\mathbf{w}^{(X)} \leftarrow$ STARTING WEIGHTS($\mathbf{X}_s$, ψ); $\mathbf{w}^{(Y)} \leftarrow$ STARTING WEIGHTS($\mathbf{Y}_s$, ψ).

算法 2初始权重

要求： 缩放后的数据块 $\mathbf{Z}_s \in \mathbb{R}^{n\times d}$，加权函数 ψ

$\mathbf{d} \leftarrow \left[\|\mathbf{z}_{r,\,i}\|_2 \,/\, \text{中位数}_i \,\|\mathbf{z}_{r,\,i}\|_2\right]_{i=1}^{n}$ ▷ 标准化的逐例距离

$\mathbf{w} \leftarrow \psi(\mathbf{d})$ ▷ 应用加权函数

$w_i \leftarrow \max(w_i,\, 10^{-6})$ 对所有 i ▷ 为较小权重设定下限

返回 w

算法 3稳健双块（RTB）

要求： $\mathbf{X} \in \mathbb{R}^{n\times p}$，$\mathbf{Y} \in \mathbb{R}^{n\times q}$，成分 h_x, h_y，加权函数 ψ

对 $\mathbf{X} \to \mathbf{X}_s$、$\mathbf{Y} \to \mathbf{Y}_s$ 进行稳健中心化和尺度化

$\mathbf{w}^{(X)} \leftarrow$ STARTING WEIGHTS$(\mathbf{X}_s, \psi)$

宽$^{(Y)} \leftarrow$ STARTING WEIGHTS$(\mathbf{Y}_s, \psi)$

$\mathbf{X}_w \leftarrow \text{diag}(\sqrt{\mathbf{w}^{(X)}})\mathbf{X}_s$; $\mathbf{Y}_w \leftarrow \text{diag}(\sqrt{\mathbf{w}^{(Y)}})\mathbf{Y}_s$

重复

拟合 twoblock$(\mathbf{X}_w, \mathbf{Y}_w) \to$ 得分 $\mathbf{T}$, $\mathbf{U}$, 系数 $\mathbf{b}$

$\mathbf{T}_{\text{uw}} \leftarrow \mathbf{T} \oslash \sqrt{\mathbf{w}^{(X)}}$; $\mathbf{U}_{\text{uw}} \leftarrow \mathbf{U} \oslash \sqrt{\mathbf{w}^{(Y)}}$

$\mathbf{w}^{(X)} \leftarrow \psi\big(d(\mathbf{T}_{\text{uw}})\big)$; $\mathbf{w}^{(Y)} \leftarrow \psi\big(d(\mathbf{U}_{\text{uw}})\big)$

$\mathbf{X}_w \leftarrow \text{diag}(\sqrt{\mathbf{w}^{(X)}})\mathbf{X}_s$; $\mathbf{Y}_w \leftarrow \text{diag}(\sqrt{\mathbf{w}^{(Y)}})\mathbf{Y}_s$

直到 $\|\mathbf{b}\|^2$ 收敛

将系数还原至原始尺度

返回 $\hat{}\mathbf{B}$, $\mathbf{w}^{(X)}$, $\mathbf{w}^{(Y)}$

3. 构造加权数据：$\mathbf{X}_w = \text{diag}(\sqrt{\mathbf{w}^{(X)}})\,\overline{\mathbf{X}_s}$和$\mathbf{Y}_w = \text{diag}(\sqrt{\mathbf{w}^{(Y)}})\,\overline{\mathbf{Y}_s}$。

迭代重新加权 。重复执行直至收敛（或达到最大迭代次数）：

1. 使用中心化方式设为“均值”、尺度化方式设为“None”的设置，将双块模型（稠密或稀疏）拟合至$\mathbf{X}_w$ 和$\mathbf{Y}_w$ （因为数据已进行稳健预处理）。2. 从拟合模型中提取$\mathbf{X}$块得分$\mathbf{T}$ 和$\mathbf{Y}$块得分$\mathbf{U}$，将其除以 $\sqrt{\mathbf{w}}$以取消加权。3. 更新$\mathbf{X}$块权重：计算未加权得分的稳健马氏距离，并应用加权函数 ψ 以获得更新后的$\mathbf{w}^{(X)}$。4. 类似地，根据$\mathbf{Y}$ 得分更新$\mathbf{Y}$块权重，以获得更新后的$\mathbf{w}^{(Y)}$。5. 使用更新后的权重重新构造$\mathbf{X}_w$ 和$\mathbf{Y}_w$ 。6. 检查收敛情况：$|\|\mathbf{b}^{(t)}\|^2 - \|\mathbf{b}^{(t-1)}\|^2|/\|\mathbf{b}^{(t-1)}\|^2 < \tau$。

最终估计值。收敛后，回归系数会重新缩放到原始数据尺度。观测值 i 的合并案例权重为 $w_i = w_i^{(X)} \cdot w_i^{(Y)}$。

稳健双块算法总结见算法 3。

2.3 性质

正如下一节的模拟研究将展示的那样，RTB继承了twoblock（稠密或稀疏）的降维和变量选择特性，同时增强了对离群值的稳健性。对**X**块和**Y** 块进行独立加权，使该方法能够处理杠杆点（通过$\mathbf{w}^{(X)}$降低**X** 中的离群值权重）、垂直离群值（通过$\mathbf{w}^{(Y)}$降低**Y**中的离群值权重）以及联合污染，因为两个权重共同构成组合权重$\mathbf{w} = \mathbf{w}^{(X)} \odot \mathbf{w}^{(Y)}$。

对于干净数据，采用分布分位数作为标准截断值（$p_1 = 0.95$, $p_2 = 0.975$, $p_3 = 0.999$）的 Hampel加权函数会为绝大多数观测值赋予权重 1，因此 RTB 收敛到接近经典双块方法的解。预计污染程度较高时，可以使用更激进的截断值（例如 $p_1 = 0.75$, $p_2 = 0.90$, $p_3 = 0.95$）。下文呈现的结果将采用后一种设置。

3 模拟研究

本模拟研究将评估该方法的稳健性和稀疏性。模拟的设置能够生成符合潜变量模型的数据，且模型复杂度已知，因此适合采用各种双块方法进行建模，无需调节相应参数。在六种不同情景中加入离群值：离群值比例分为两种，且加入的离群值可以仅为垂直离群值、仅为杠杆点，或同时包含这两种类型。为评估稀疏方法的变量选择性能，将按两种比例向**X**块中添加噪声变量：一种情景中添加的噪声变量远少于信息变量，另一种情景中更多 的非信息变量多于信息变量。最后，还将在两种情景下评估**X** 块数据的维度，其中一种情景 $n < p$，另一种情景 $p > n$。每种设置的结果均为两百次运行的平均值。总计而言，本模拟研究评估了 42 种情景，共需进行 33200 次独立运行。

3.1 设计

数据由以下真实潜变量模型生成：

$$\mathbf{X} = \mathbf{T}\mathbf{P}^\top + \mathbf{E}, \quad \mathbf{T} \in \mathbb{R}^{n \times k},\ \mathbf{P} \in \mathbb{R}^{p_{\text{signal}} \times k},\ E_{ij} \sim \mathcal{N}(0, \sigma_e^2), \tag{5}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{C} + \mathbf{F}, \quad \mathbf{C} \in \mathbb{R}^{k \times q},\ F_{ij} \sim \mathcal{N}(0, \sigma_f^2), \tag{6}$$

其中**P** 的列为正交归一列，$T_{ij} \sim \mathcal{N}(0, 1)$，且 $C_{ij} \sim \mathcal{N}(0, 1)$。真实回归系数为$\mathbf{B}_{\text{true}} = [\mathbf{P}\mathbf{C};\ \mathbf{0}]$, 针对非信息变量补零。非信息变量作为 $\mathcal{N}(0, \sigma_e^2)$噪声的列附加。

我们设定 $n = 100$、$k = 3$、$q = 4$、$\sigma_e = \sigma_f = 0.5$，并改变：

- **维数**: $p_{\text{信号}} = 20$（包含$p_{\text{噪声}} \in \{0, 10, 80\}$）和$p_{\text{信号}} = 150$（包含$p_{\text{噪声}} \in \{0, 50, 250\}$），涵盖 $p < n$和 $p > n$两种情形。

- **污染**：0%、10%或20%的样本存在离群值，仅在**X** 中、仅在**Y** 中，或两者中均有离群值；离群值通过幅度为10的加性偏移引入。

比较了四种方法，每种方法均使用真实的 k 作为**X** 成分的数量，并使用 $\min(k, q)$ 作为**Y** 成分的数量：稠密和稀疏经典双块方法（采用标准尺度化），以及采用 Hampel 加权的稠密和稀疏稳健双块（RTB）方法，使用中位数/MAD 预处理、标准截断值，以及 Hampel 函数截断值集合 $\{p_1 = 0.75, p_2 = 0.90, p_3 = 0.95\}$。在此我们指出，采用更严格的降权截断值设定（例如 $\{p_1 = 0.95, p_2 = 0.99, p_3 = 0.999\}$ ）会得到稳健性较低但效率很高的估计量。然而，所示估计量的模拟结果在效率方面已经相当令人信服，因此不再展示采用更严格参数集合所得的结果。

为评估回归系数的准确性，报告了估计回归系数与实际回归系数之差的弗罗贝尼乌斯范数 $\mathrm{MSE}(\mathbf{B}) = \frac{1}{pq}\|\hat{\mathbf{B}} - \mathbf{B}_{\mathrm{true}}\|_F^2$，该数值取 200 次独立重复的平均值。为评估稀疏方法的变量选择表现，报告了变量选择的 F_1 分数。

3.2 结果

系数估计准确度。图1 展示了全部 42 种情景下系数估计值的均方误差（MSE），并按 2×3 网格排列。稠密和稀疏 twoblock、稠密和稀疏 RTB 这四种方法并列展示。顶行对应 $p_{\mathrm{signal}} = 20 (p \le n)$，底行对应 $p_{\mathrm{signal}} = 150$ $(p > n)$，无信息噪声变量的数量从左到右递增。

在所有维度设置下，均可观察到以下几种一致的规律：

1. **干净数据**。在未受污染的数据（0%，无污染）上，四种方法的表现相近；与相应的经典双块变体相比，RTB 仅造成适度的效率损失（稠密方法的均方误差比为 ≈ 1.1–1.3；稀疏方法为 ≈ 1.1–1.2）。2. **X 离群值**。**X** 中的杠杆点会使稠密和稀疏双块方法的均方误差大幅膨胀（例如，在$p = 20$下，污染率为 10% 时，稠密方法的均方误差从 0.021 升至 0.124；稀疏方法则从 0.030 升至 0.135）。两种 RTB 变体基本不受影响，均方误差仍接近干净数据基线。3. **Y 离群值**。**Y** 中的垂直离群值是所有方法面临的最具挑战性的情景。与经典双块方法相比，RTB 可显著降低均方误差（例如，在 $p = 20$ 下，污染率为 10% 时，稠密方法的均方误差比为 0.31；稀疏方法为 0.48 ）。当噪声变量占多数时（$p_{\mathrm{noise}} = 80$），稀疏 RTB 的保护效果尤为显著（污染率为 10% 时，均方误差为 0.014，而稀疏双块方法为 0.062）。4. **联合 XY 离群值**。当两个块均受到污染时，均方误差比与仅有**X**离群值的情形相似，因为**X**块的权重在重新加权中占主导地位。稠密和稀疏 RTB 均保持优势。

5. **高维设定**。 在 $p > n$ 情形下，RTB 的优势依然存在，相对增益也相近。对于 $p = 400$ ($p_{\mathrm{signal}} = 150$, $p_{\mathrm{noise}} = 250$)，在 **X** 污染下，稠密 RTB 的比率为 0.64–0.66，稀疏变体的表现也相当。

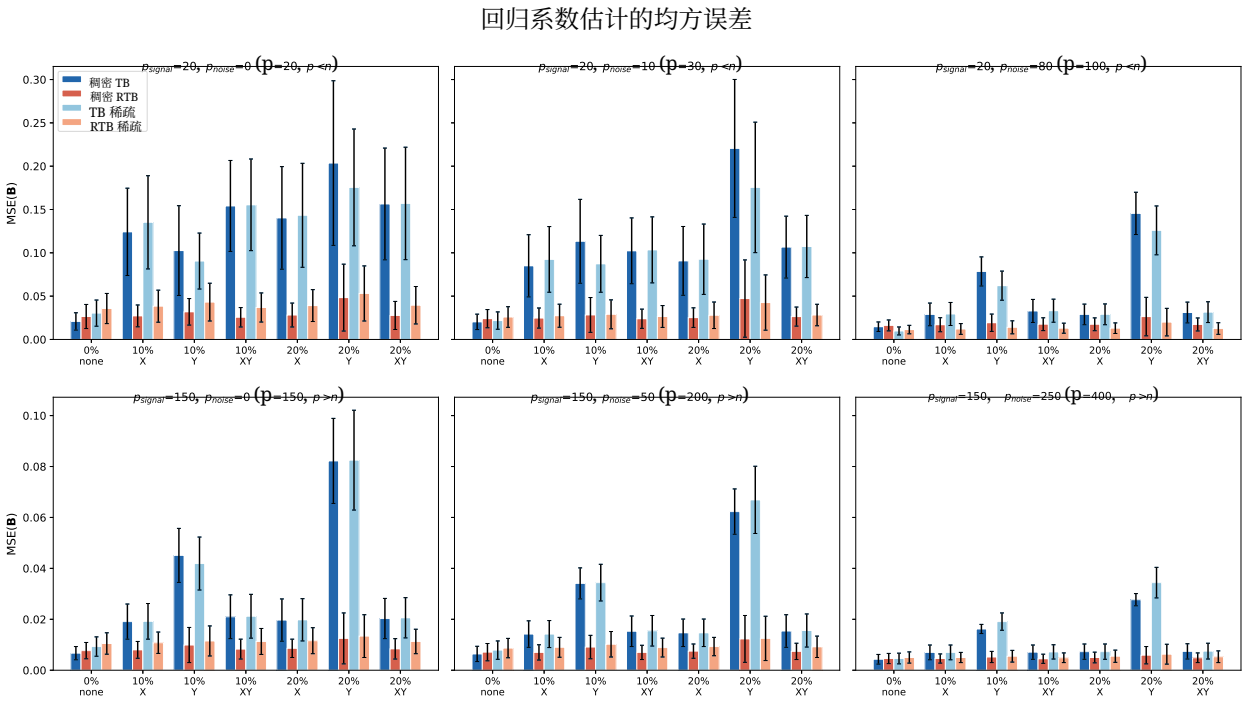

图 1：42 个模拟场景下回归系数估计的均方误差（每个场景重复 200 次）。深蓝色：稠密 TB；深红色：稠密 RTB；浅蓝色：TB 稀疏；鲑红色：RTB 稀疏。上排：$p \leq n$；下排：$p > n$。每个面板中，x 轴显示污染比例和类型。误差线表示一个标准差。

6. **噪声变量**。加入非信息变量不会从整体上改变各方法的相对表现，不过均方误差的绝对值会下降（因为按 pq进行了归一化）。存在大量噪声变量时，稀疏方法的均方误差往往低于相应的稠密方法（例如 $p_{\text{noise}} = 80$：干净数据下，稀疏 TB 为 0.010，稠密 TB 为 0.015），这体现了变量选择的优势。

图 2 以均方误差比热图的形式，总结了所有情景下稠密与稀疏 RTB 相对于对应 twoblock 变体的相对效率。数值低于 1（绿色）表示 RTB 优于 twoblock；数值高于 1（红色）则表示相反。twoblock 唯一胜出的情景是干净数据情形（比值 ≈ 1.07–1。27），这反映了适度的稳健性代价。在任何形式的污染下，RTB 均能持续取得远低于 1 的比值，其中在污染率为 20% 且存在 **Y** 离群值时，优势最为显著。稠密与稀疏变体呈现出高度一致的模式。

变量选择准确性。对于稀疏方法，图 3 报告了 F_1 变量选择得分，即精确率与召回率的调和平均数，用于衡量在信号变量和噪声变量的混合数据中识别真实信号变量的能力。

在干净数据上，两种稀疏方法取得相同的 F_1 得分，这证实了在不存在离群值的情况下，稳健预处理不会损害变量选择。然而，在污染下，出现了一个显著的现象：当

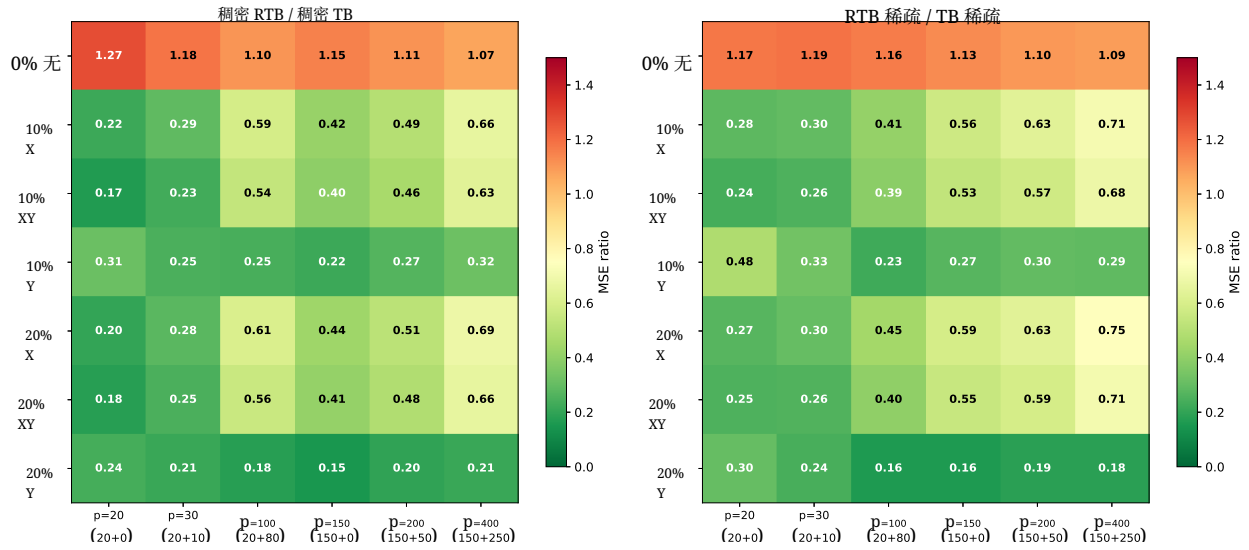

图2：所有模拟场景中，稠密（左）和稀疏（右）变体的 RTB 相对于 twoblock 的均方误差比。绿色单元格（< 1）表示 RTB 优于 twoblock；红色单元格（> 1）则表示相反。唯一的红色单元格对应未污染基线。

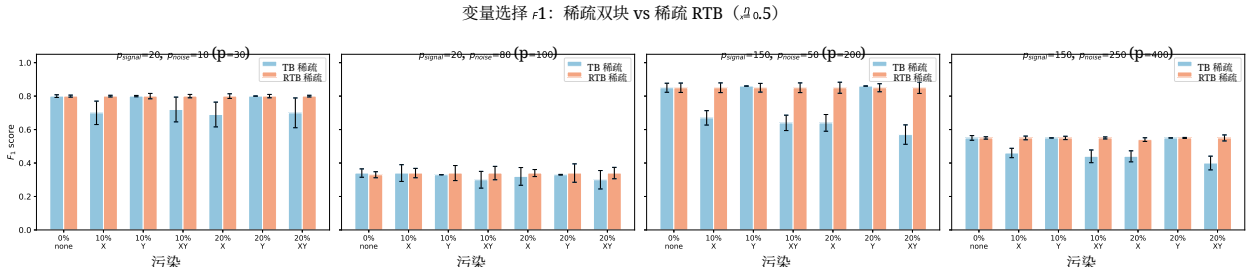

图 3：稀疏双块（浅蓝色）和稀疏 RTB（鲑红色）在不同污染情景下的变量选择 F_1 分数，针对包含噪声变量（$\eta_x = 0.5$）的配置。分数越高越好。

X 离群值或联合**X—Y** 离群值存在时（例如，在 **10–20%** 的污染水平下，$p_{\text{signal}} = 150$、$p_{\text{noise}} = 50$ 从 **0.85** 降至 **0.57–0.67**），稀疏 RTB 则将其 F_1 维持在接近干净数据的水平（≈ 0.85）。这表明，RTB 不仅能保护系数估计，还能在污染下保留稀疏方法的变量选择一致性。有趣的是，仅有**Y**离群值不会降低任何一种方法的变量选择性能，因为稀疏性施加于**X**块权重。

4 真实世界数据应用

本节将通过两个示例说明新方法的实用性。第一个数据集包含数量适中的样本，但变量数多于样本数。之所以纳入该数据集，是因为双块降维的开创性论文（Cook 等人，2025年）已指出，该数据集能够很好地展示同步降维的优势

。此外，Serneels（2025年）表明，稀疏双块维度的内在变量选择优于Cook 等人（2025年）中的结果。此外，众所周知，该数据集包含离群值，而此前的两篇论文均将其移除。本节将展示，尽管存在离群值，稳健稀疏双块的表现仍大致不逊于两种经典方法；对于四个响应变量中的一个，其表现甚至优于先前发表的结果。

第二个数据集考察了燃气轮机运行中的一组传感器读数。其案例数远多于变量数。所选变量具有代表性，但其中一两个变量对建模目标贡献甚微也是合理的推测。该数据集的响应变量已知会产生自然的垂直离群值，这有助于对本文提出的方法进行说明性比较。

4.1 饼干面团近红外数据

饼干面团数据集（Osborne 等人，1984年）由面团样品的近红外（NIR）光谱（$p = 700$ 个波长）组成，并以四种分析物（脂肪、蔗糖、面粉、水）的实验室测量值作为响应变量。训练集包含40个样本，测试集包含32个样本。已知该数据集包含异常观测值：训练集中的案例23和测试集中的案例21已被识别为离群值。

对于非稳健方法，将移除这些案例，并报告干净数据（$n_{\text{train}} = 39$、 $n_{\text{test}} = 31$）上的拟合结果。对于 RTB，训练时保留所有案例，以便通过重新加权自动处理离群值。

比较的方法。我们比较了 PLS2、PLS1（按响应变量分别建模）、twoblock（稠密和稀疏）以及 RTB（稠密和稀疏）。对于经典方法，使用干净数据通过交叉验证选择超参数。对于 RTB，则使用完整（受污染）数据通过交叉验证选择超参数，并采用 Hampel 加权（截断值 $p_1 = 0.75$, $p_2 = 0.90$,$p_3 = 0.95$）和中位数中心化。

交叉验证选出了以下配置：

- PLS2：6 个成分，不进行尺度化。
- PLS1：分别使用 7 个（脂肪）、6 个、6 个和 7 个（水）成分。
- 双块密集型： $h_x = 6$, $h_y = 2$，不进行尺度化。
- 双块稀疏： $h_x = 9$, $h_y = 2$, $\eta_x = 0.5$，采用标准尺度调整。
- 稠密 RTB： $h_x = 11$, $h_y = 2$，kstepLTS 尺度调整。
- RTB 稀疏： $h_x = 8$, $h_y = 2$, $\eta_x = 0.3$，kstepLTS 尺度调整。

结果。 表 1报告了各方法和响应变量的测试集 R^2 表现。非稳健方法使用干净数据进行训练和评估（39 个训练样本，31 个测试样本），而 RTB 使用完整数据进行训练（40 个训练样本，包括离群值），并在干净测试集上评估（31 个样本）。

表1：曲奇面团数据：按方法和响应变量划分的测试集 R^2 。非稳健方法使用干净数据训练；RTB 使用包含离群值的完整数据训练。

方法	Fat	蔗糖	面粉	水	Mean
PLS2（干净）	0.550	0.948	0.746	0.658	0.725
PLS1（干净）	0.979	0.935	0.730	0.916	0.890
双块密集型（干净）	0.596	0.940	0.821	0.379	0.684
双块稀疏（干净）	0.930	**0.962**	**0.931**	**0.948**	**0.943**
稠密 RTB（完整）	0.933	0.908	0.839	0.823	0.876
RTB 稀疏（完整）	**0.980**	0.914	0.876	0.904	0.918

有几个值得注意的现象。首先，基于干净数据的稠密 twoblock 在脂肪和水上的表现不佳（$R^2 = 0.596$ 和 0.379），这表明即使移除了已知离群值，经典方法仍会受到剩余数据结构的影响。其次，基于干净数据的稀疏双块取得了最佳整体表现（均值 $R^2 = 0.943$），展现了变量选择的价值。第三，使用包含离群值的完整 数据训练的稠密 RTB，取得了 0.876 的平均测试 R^2 表现，明显优于基于干净数据的稠密 twoblock。最后，基于完整数据的稀疏 RTB 在全部四种分析物上均表现出色（均值 $R^2 = 0.918$），与基于干净数据的稀疏双块相比颇具竞争力，而且无需事先识别或移除离群值。这既体现了该方法的实用性，也展示了其对离群值的稳健性和较高效率，其结果接近针对未受污染数据的经典估计量。

案例权重诊断。图4 显示了稠密 RTB 模型的综合案例权重。案例23的权重为 10^{-6} （实际上为零），这证实 RTB 会自动识别并降低这个已知离群值的权重。案例24的权重也被降低，表明它可能同样具有异常特征。所有其他案例的权重均为满权重（$w = 1$）。

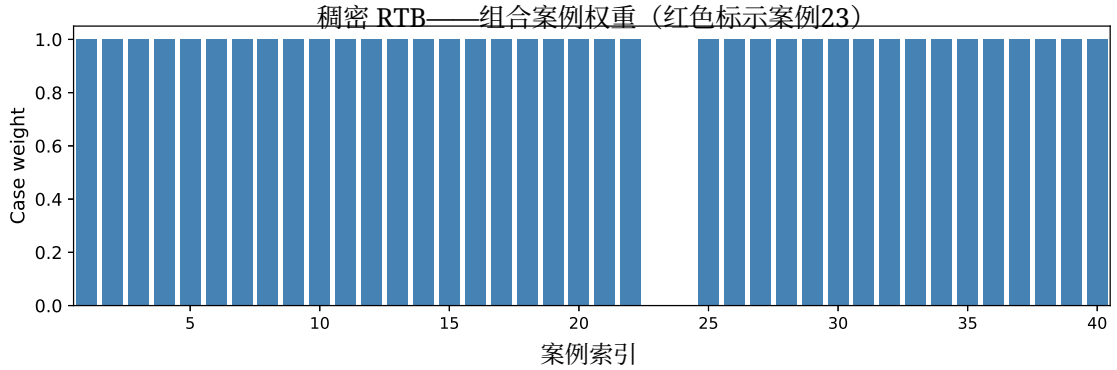

图4：曲奇面团训练数据的 RTB 稠密案例权重。案例23（红色）和案例24被赋予接近零的权重。

稀疏 RTB 模型产生了相同的权重模式（图5），这证实了稠密与稀疏变体中的异常值识别结果一致。事实上，

稀疏稳健方法在脂肪变量上的表现优于经典方法，这可能表明案例24仅在该变量上是一个垂直离群值。然而，实际上无法验证这一假设。

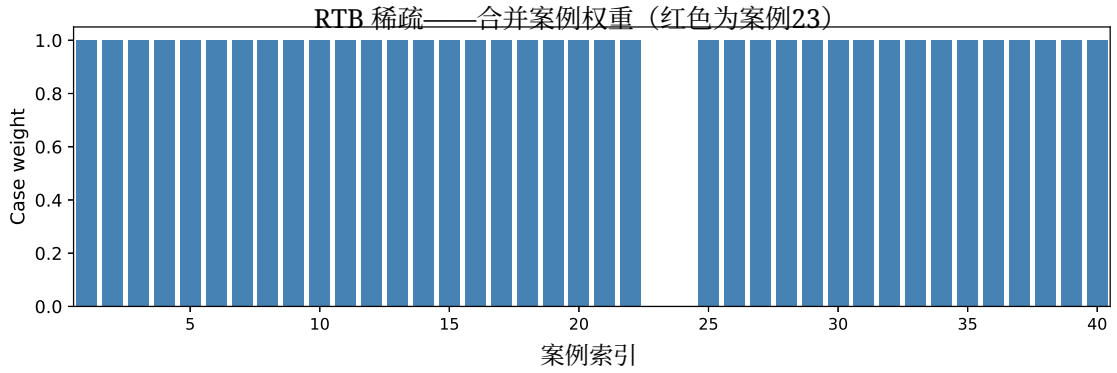

图 5：曲奇面团训练数据的 RTB 稀疏案例权重。案例23和24（红色：案例23）再次被赋予接近零的权重。

变量选择。 稀疏 RTB 模型（$\eta_x = 0.3$，$h_x = 8$）剔除了700个波长中的368个（52.6%），保留332个波长（47.4%）。图 6 展示了叠加了被剔除波长区域（蓝色阴影）的近红外光谱。主要剔除了四个区域：初始基线区域（波长0–26），以及中高波长范围内的三个区域（165–215、302–341、397–646）。这些区域对应的光谱信噪比较低，或对四种分析物的判别能力有限。值得注意的是，该变量选择结果与非稳健稀疏双块降维所得结果几乎无法区分，后者首次发表于（Serneels，2025年）。

4.2 燃气轮机数据

燃气轮机 CO 和 NOx 排放数据集（Kaya 等人，2019年）包含2011年至2015年期间从联合循环发电厂的一台燃气轮机采集的36,733条小时传感器读数。预测变量块由 $p = 9$ 过程变量组成：环境温度（AT）、环境压力（AP）、环境湿度（AH）、空气过滤器压差（AFDP）、燃气轮机排气压力（GTEP）、涡轮机入口温度（TIT）、涡轮机后温度（TAT）、涡轮机能量产出（TEY）以及压缩机出口压力（CDP）。双变量响应由 CO 和 NOx 排放浓度组成，这些浓度已知会产生自然垂直离群值。

为识别这些自然离群值，我们在二元响应空间中拟合了一个基于最小协方差行列式（MCD）估计量（Rousseeuw，1985年）的97.5%容许椭圆。在一个随机抽取的5,000个观测值的子样本中，该方法识别出1,064个离群值（21.3%），并将其与3,936个内点区分开来。内点按70/30的比例划分为干净训练集（$n_{\text{train}} = 2,755$）和干净测试集（$n_{\text{test}} = 1,181$）。随后构建了一个受污染训练集

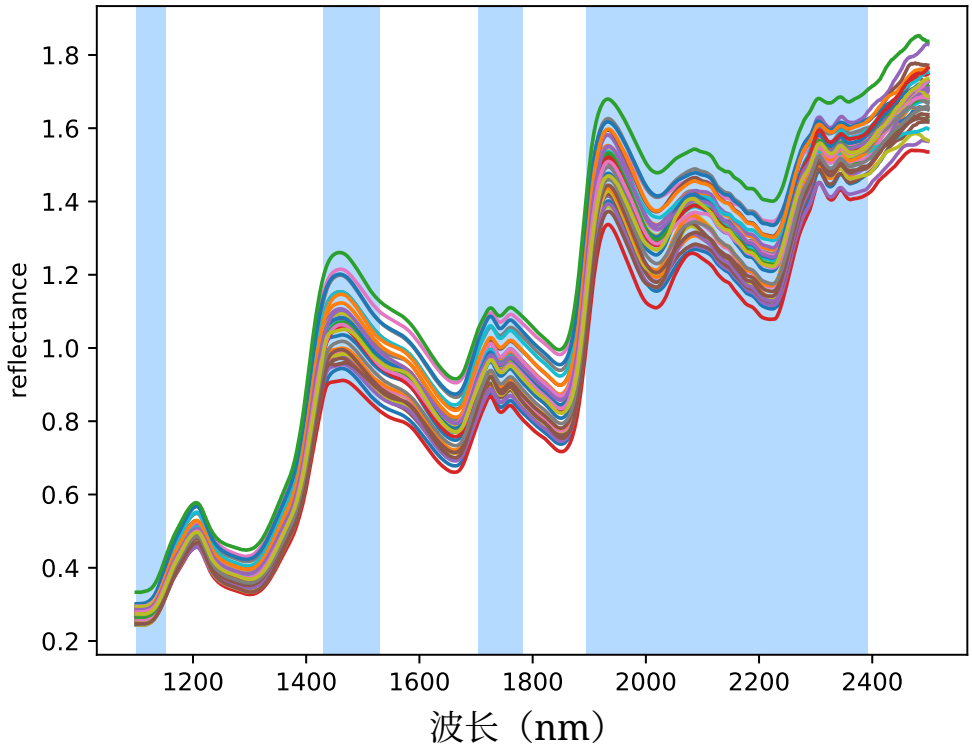

图6：曲奇面团近红外光谱，以及稀疏 RTB 模型剔除的波长区域（蓝色阴影）。

方法是注入检测出的离群值中的306个（目标污染率为 ≈10%），并特别纳入CO和氮氧化物各自都极端的样本。容许椭圆及由此得到的离群值分类如图7所示。

比较的方法。比较了五种方法，各方法的超参数均通过相应训练集上的 5 折交叉验证选定：

- PLS2：9 个成分。
- 双块密集型：$h_x = 9,\, h_y = 2$，不进行尺度化（干净）；$h_x = 7,\, h_y = 2$（受污染）。
- 双块稀疏：$h_x = 9,\, h_y = 2,\, \eta_x = 0.3$（干净）；$h_x = 5,\, h_y = 2,\, \eta_x = 0.7$（受污染）。
- 稠密 RTB：$h_x = 9,\, h_y = 2$（清洁）；$h_x = 7,\, h_y = 1$（受污染）。
- RTB 稀疏：$h_x = 9,\, h_y = 2,\, \eta_x = 0.3$（清洁）；$h_x = 5,\, h_y = 2,\, \eta_x = 0.7$（受污染）。

结果。表2 报告了两种训练情景下各方法的测试集均方误差。由于一氧化碳和氮氧化物变量的方差差异显著（$\sigma^2_{\mathrm{CO}} \approx 5.7$，$\sigma^2_{\mathrm{NOx}} \approx 135$），还报告了

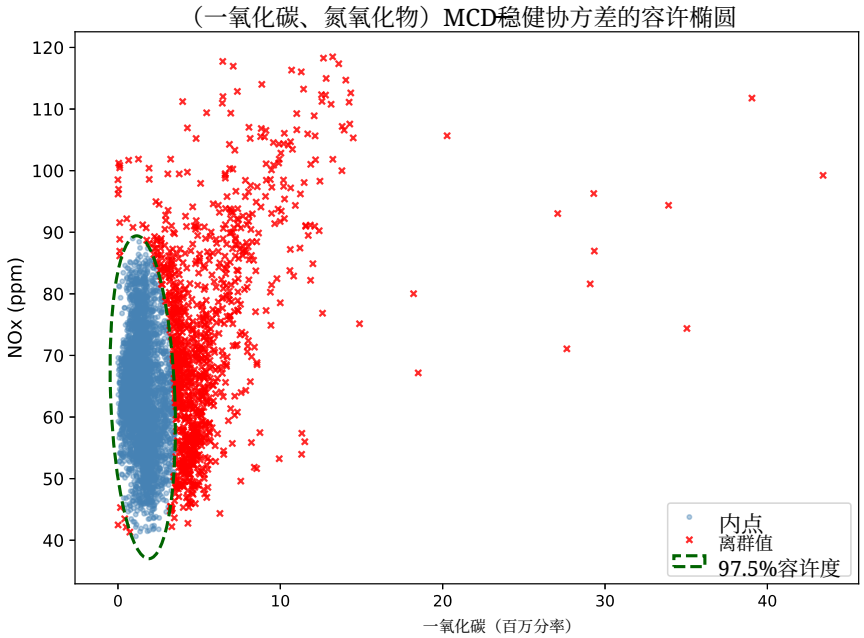

图7：（一氧化碳、氮氧化物）双变量散点图及97.5% MCD容许椭圆。红色十字：被判定为离群值的观测值；蓝色：用于干净训练集和测试集的内点。

加权均方误差，使用逆方差权重进行加权，确保两个响应变量对总体得分的贡献成比例。

表2：燃气轮机数据：按方法和响应变量划分的测试集均方误差。加权均值采用逆方差权重（CO：0.960，NOx：0.040）。

	干净训练			受污染训练		
方法	CO	NOx	加权均值	CO	NOx	加权均值
PLS2	0.318	30.40	1.532	1.111	47.26	2.974
稠密 TB	0.318	30.40	1.532	1.121	47.56	2.997
TB 稀疏	0.318	30.40	1.532	1.114	60.43	3.509
稠密 RTB	**0.318**	**30.38**	**1.532**	0.566	**34.99**	1.956
RTB 稀疏	**0.318**	**30.38**	**1.532**	**0.349**	38.97	**1.908**

在干净数据上，除 PRM 外的所有方法都取得了几乎相同的均方误差，RTB 未造成可测量的效率损失。需要指出的是，PRM 的效率损失有限。在存在污染的情况下，非稳健方法（PLS2、稠密 TB、TB 稀疏）的 CO 均方误差约增加一倍，而 RTB 的各个变体仍接近其在干净数据上的表现。值得注意的是，稀疏 RTB 在污染数据上的 CO 均方误差为 0.349，与干净数据上的结果（0.318）非常接近；同时，它在两个变量的加权均值均方误差方面都取得了最佳结果，表明存在离群值时，稀疏 RTB 能为两种排放物提供最佳的同时多变量模型。

案例权重诊断。 图8展示了在受污染训练集上拟合的 RTB 稠密模型的综合案例权重。306 个注入的离群值（红色柱形，接在 2,755 个干净案例之后）获得了大幅降低的权重，其中 47% 的权重被完全降至零（$w < 0.5$）。其余离群值获得了部分权重，这与并非所有位于容许椭圆之外的观测值都同样极端这一事实相符。

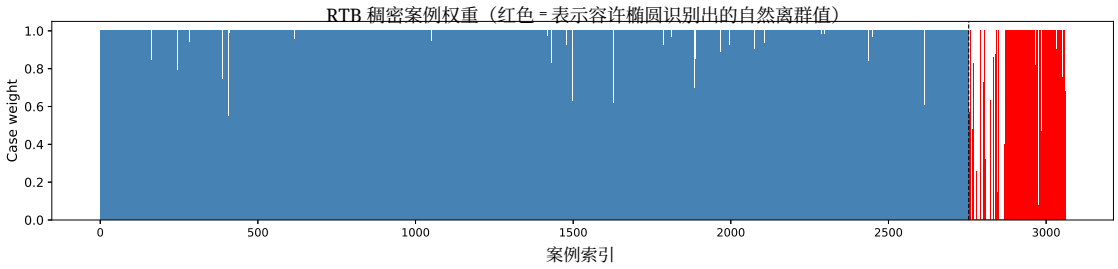

图 8：燃气轮机受污染训练集的 RTB 稠密案例权重。红色条形表示由 MCD 容许椭圆识别出的自然离群值。

变量选择。稀疏 RTB 模型（$\eta_x = 0.7$， $h_x = 5$）最终确实保留了受污染数据中的全部 9 个预测变量，这表明所有传感器读数都包含一些可用于预测排放变量对的信息。然而，图9显示，每个响应变量的系数分布截然不同。对于一氧化碳，主要预测变量是 TIT（涡轮进口温度）和 TEY（能量产出），而 AFDP、TAT 和 CDP 的系数接近零。对于氮氧化物，主要预测变量是 AT（环境温度），其次是 GTEP（排气压力）和 TIT，且各变量的系数分布更为分散。这种差异化结构体现了多变量双块建模的优势：共享潜在空间捕捉共同结构，而针对各响应变量的系数则揭示了哪些物理过程变量驱动着各类排放。五成分 RTB 模型最终保留了所有变量，却仍能优于稠密版本，这表明以稀疏方式提取成分仍有优势：对于稠密 RTB，所有**X** 块成分权重均非零，而对于稀疏双块，只有一个或少数几个变量的成分权重非零。最终，加权平均均方误差结果表明，即使最终模型使用了所有变量，稀疏成分提取仍可能带来优势。此外，稀疏成分权重也更易于解释。

5 结论与展望

本文介绍了稳健双块降维（RTB）。RTB 是首个双块同步降维的稳健变体，同时提供稠密和稀疏形式。因此，它也是首个能够分别调节每个变量块中的模型复杂度，且对离群值具有稳健性的同步 twoblock 降维方法。稀疏版本还是首个允许为每个块分别调节潜变量数量以及

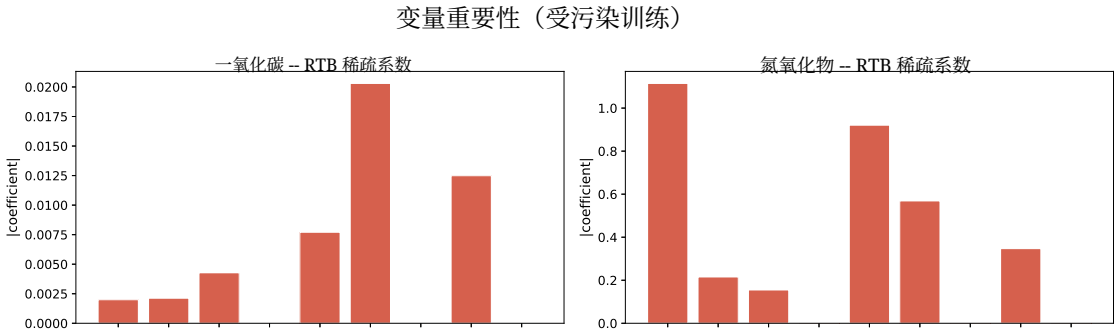

图 9：受污染的燃气轮机数据中，一氧化碳（左）和氮氧化物（右）的 RTB 稀疏绝对系数幅值。保留全部 9 个预测变量；颜色深浅表示系数幅值。

稀疏度参数的稀疏双块降维方法。需要指出的是，在稀疏稳健典型相关分析中，分别调节稀疏度参数早已成为可能（Wilms 和 Croux， 2016年），但典型相关分析固有的一项缺陷是，两个变量块都必须采用相同数量的潜在成分，即使两个块的维度相差很大也是如此。

通过对$\mathbf{X}$ 块和$\mathbf{Y}$ 块采用独立案例权重进行迭代重加权估计，RTB 能够抵御杠杆点、垂直离群值和联合污染，同时在干净数据上的估计效率接近经典双块方法。

这一点在一项广泛的模拟研究以及两个示例中得到了说明。模拟研究表明，在存在污染的情况下，RTB 的表现始终优于经典双块方法；离群值占比越高，这一优势越明显，而且这些优势也适用于高维 $p > n$ 设置。

本文提出的工作为未来研究留下了若干值得探索的方向。虽然该方法允许在每个变量块中选择不同的稀疏度，但本文重点关注的是降维和多元回归场景下的稳健性，以及$\mathbf{X}$块中的变量选择。未来工作还需以类似方式研究因变量块中的变量选择。该方法在这一方面的应用可见于生物信息学，例如将基因组数据与代谢组数据联系起来的模型；在这类模型中，分析前并不知道应在各个块中选择哪些变量，而离群值则可能自然地出现在任一块中。另一个值得探索的方向是将该方法扩展至时间序列数据。尽管未来研究有许多潜在方向可供设想，但希望本文所呈现的结果已能推动稳健双块降维方法得到更广泛的应用。

A 代码可用性

本文提出的所有方法均由Python 3 软件包 `twoblock`实现。该软件包已在 PyPI 上公开发布，其源代码可

在 GitHub 上获取。该软件包实现了方法 `twoblock` 和其稳健版本 `rtb`。本文讨论的超参数易于调整，因为这些方法兼容广泛使用的 `scikit-learn`API，并支持常见属性，例如 `fit`、 `predict` 和 `transform`。

B 数据可用性

所有数据均属公有领域。为方便起见， `twoblock` 软件包 `data/` 目录（https://github.com/SvenSerneels/twoblock/data）中附有一份副本。原始燃气轮机 CO 和 NOx 排放数据集（Kaya 等人，2019年）可从 UCI 机器学习知识库公开获取，网址为 https://archive.ics.uci.edu/dataset/551。本文结果所用的所有仿真代码和示例笔记本均包含在软件包代码库中。

C 生成式人工智能的使用

准备本文稿件时，使用了 Claude Opus 4.6（Anthropic 公司）来润色代码和 Jupyter 笔记本，并将部分结果整理到论文的表格和图中。

参考文献

Cook, R. D., Forzani, L., and Liu, L. (2023). 用于回归中响应变量向量和预测变量向量同时降维的偏最小二乘法。*Journal of Multivariate Analysis*, 196:105163.

Cook, R. D., Helland, I. S., 和 Su, Z.（2013）。包络与偏最小二乘回归。《皇家统计学会杂志：*B*辑（统计方法学）》， 75(5):851–877。Filzmoser, P., H¨oppner, S., Ortner, I., Serneels, S., 和 Verdonck, T.（2020）。逐单元稳健 M 回归。《计算统计与数据分析》，147:106944。Fritz, H., Filzmoser, P., 和 Croux, C.（2012）。多元 l 1 中位数算法的比较。《计算统计》， 27(3):393–410。Hoffmann, I., Serneels, S., Filzmoser, P., 和 Croux, C.（2015）。稀疏偏稳健 M 回归。*Chemometrics and Intelligent Laboratory Systems*， 149:50–59。Kaya, H., T¨ufekci, P., 和 Uzun, E.（2019）。燃气轮机 CO 和 NOx 排放数据集。UCI 机器学习知识库。数据集 #551，https://archive.ics.uci.edu/dataset/551。Maronna, R. A., Martin, R. D., Yohai, V. J., 和 Salibi´an-Barrera, M.（2019）。《稳健统计：理论与方法（附 *R*）》。约翰·威立父子出版公司。

Maronna, R. A. 和 Zamar, R. H.（2002）。高维数据集的位置与离散度稳健估计值。*Technometrics*，44(4):307–317。Osborne, B. G., Fearn, T., Miller, A. R., 和 Douglas, S.（1984）。近红外反射光谱在饼干和饼干面团成分分析中的应用。*Journal of the Science of Food and Agriculture*，35(1):99–105。Raymaekers, J. 和 Rousseeuw, P.（2019）。广义空间符号协方差矩阵。*Journal of Multivariate Analysis*，171:94–111。Rousseeuw, P. J.（1984）。最小中位数平方回归。*Journal of the AmericanStatistical Association*，79(388):871–880。Rousseeuw, P. J.（1985）。高崩溃点的多元估计。载于 Grossmann, W., Pflug, G., Vincze, I. 和 Wertz, W.（编），《数理统计与应用》*B*卷，第283–297页。Reidel Publishing Company，多德雷赫特。Serneels, S.（2025）。稀疏双块降维：稀疏PLS2和典型相关分析的一种多用途替代方法。*Journal of Chemometrics*，39:e70051。Serneels, S., Croux, C., Filzmoser, P., 和 Van Espen, P. J.（2005）。偏稳健 M 回归。*Chemometrics and Intelligent Laboratory Systems*，79(1–2):55–64。Serneels, S., Insolia, L., 和 Verdonck, T.（2024）。利用诱导鲁棒性的变换优雅地实现稀疏偏最小二乘的鲁棒化。*Statistics*，58(1):44–64。Wilms, I. 和 Croux, C.（2016）。鲁棒稀疏典型相关分析。*BMC SystemsBiology*，10:1–13。Wold, H.（1966）。通过迭代最小二乘程序进行非线性估计。载于 David, F.（编），*Papers in Statistics: Festschrift for J. Neyman*，第411–444页。Wiley。