arXiv:2303.04694v1 [math.ST] 8 Mar 2023

双边矩阵回归

Nayel Bettache[1] 和 Cristina Butucea[1]

[1]CREST、ENSAE、巴黎理工学院，5 avenue Henry Le Chatelier，91120 帕莱索，法国

2023年3月9日

摘要

双边矩阵回归模型 $Y = A^*XB^* + E$ 旨在通过同时考虑以下两种线性联系来进行预测：一是通过未知矩阵 B^* 建立的 X 列特征之间的联系，二是通过矩阵 A^* 建立的 X 行特征之间的联系。我们针对这一高维矩阵回归模型，通过秩惩罚最小二乘和核范数惩罚最小二乘提出低秩预测器。这两种准则都不是联合凸的；不过，我们提出了基于奇异值分解的显式预测器，并证明了最优预测界。我们给出了秩选择器一致的充分条件。我们还提出了一种完全数据驱动的秩自适应程序。仿真结果表明，在数据驱动的显式调节参数和噪声尺度参数选择下，该方法具有良好的预测效果，且秩选择一致。

关键词： 矩阵回归、多元响应回归、核范数惩罚、预言机不等式、秩惩罚、秩选择、双边矩阵回归。

1 引言

监督学习通常应用于大型数据库。矩阵回归假设，数据 Y 可以由矩阵 X的列所给出的一组特征及这些列的线性组合得到很好的解释。在实际生活中， Y的行通常可以由 X的行的线性组合来解释。

例如，经济数据库将经济指标存储为列特征，将国家作为行。这样的矩阵通常可以由一个更小的矩阵来解释，该矩阵大致包含较少数量的国家（地理位置或经济状况相近的群体代表）以及少数经济特征，或由所有这些指标生成的某些因子。我们希望使用特征 X a $m \times q$ 矩阵，预测更多国家的更多指标，即$i.e.$ Y a$n \times p$ 矩阵。

推荐系统旨在预测 n 客户对 p 物品的看法。我们可以利用关于若干 m 不同客户群体及其对若干 q 大型物品类别的偏好的公开数据，通过评估客户与总体中既定群体的

相关性以及物品在其类别中的权重来进行预测。我们还可以纳入多标签情形，即属于某个主类别的物品也与其他类别相关。

气象数据、医疗或药学数据等领域也有其他例子。

模型。我们观测到矩阵 $Y \in \mathbb{R}^{n\times p}$ 和设计矩阵 $X \in \mathbb{R}^{m\times q}$，二者通过**双边矩阵回归（2MR）**模型相关联，该模型涉及两个参数矩阵 $A^* \in \mathbb{R}^{n\times m}$和 $B^* \in \mathbb{R}^{q\times p}$:

$$Y = A^* X B^* + E, \tag{1}$$

其中，假设噪声矩阵 E 包含相互独立、均值为零的次高斯元素 $\sigma-$。

2MR 模型涵盖了矩阵回归和矩阵分解等已知模型，例如例如，若 $n = m$ 和 A^* 均为单位矩阵，则矩阵模型 (1) 就变为（单边）矩阵回归（MR）模型 $Y = XB^* + E$，参见 [18], [5], [17]。

现在假设 $m = q$，且设计矩阵 X 是秩为 m的单位矩阵，其中 m小于 n 和 p。我们的模型就变成了一个分解模型，用于描述在噪声干扰下观测到的信号 $M^* = A^*B^*$。其思路是恢复生成观测数据的低秩结构。在 [12] 中，作者研究了信号的结构化分解，并假设 A^* 的行和 B^* 的列具有相同的稀疏度参数，而未被观测到的 X的维度远小于 Y。

2MR 模型 (1) 与其他模型密切相关，但我们认为，它无法简化为这些性质不同的模型。事实上，注意到矩阵 Y 的元素 Y_{ij} 可以写作

$$Y_{ij} = \mathrm{Tr}(X \cdot B^*_{\cdot,j} A^*_{i,\cdot}) + E_{ij},$$

对于任意 i 属于 $[n]$，其中 $[n] = \{1, \ldots, n\}$，且对于任意 j 属于 $[p]$。因此，每个元素 Y_{ij} 都通过同一个设计矩阵 X，为秩为 1 的矩阵$B^*_{\cdot,\ j} A^*_{i,\ \cdot}$提供信息。这与迹回归模型 或 [19], [14], 研究的更一般的矩阵补全 不同；在后者中，不同的设计矩阵为参数矩阵 B^*A^* 提供信息。

模型 (1) 还可以写成向量回归模型的形式：分别将矩阵 Y、 X 和 E 的列堆叠成 vec（Y）、 vec（X）和 vec（E），得到

$$vec(Y)^\top = vec(X)^\top \cdot A^\top \otimes B + vec(E)^\top, \tag{2}$$

其中 $\otimes$ 表示两个矩阵的张量积。在这种关系下，我们利用大小为 mq 的行向量（特征矩阵的秩为 1），通过一个大小为 $(mq) \times (np)$ 的参数，预测大小为 np 的行向量；除非 A 和 B 的结构是平凡的，否则这种方法难以奏效。这种方法无法考虑特征以及矩阵 A^*、 B^* 的矩阵结构，因此效果不佳。

该模型由 [6] 引入时间序列领域，称为 1 阶自回归矩阵值模型（MAR(1)），$Y_t = A^* Y_{t-1} B^* + E_t$，在 $[T]$ 中的 t 时刻观测。在这种情况下， A^* 和 B^* 是谱半径严格小于 1 的方阵，以确保时间序列的稳定性（因此，X_t 是平稳且因果的）。作者提出了三种估计方法：首先，他们使用类似于 (2) 的向量形式，将 $vec(Y_t)^\top$ 的 T 行堆叠起来，并

他们利用最近克罗内克积（NKP）问题，根据 $A^{*\top} \otimes B^*$ 的全局最小二乘估计量，给出 A^* 和 B^* 的估计量；然后，他们的下一种方法是针对 A 和 B 最小化最小二乘

$$\min_{A,B} \frac{1}{T} \sum_{t=1}^{T} \|Y_t - AY_{t-1}B\|_F^2,$$

通过序贯程序进行：固定给定的 B，对 A 进行最小化；然后固定 A，对 B 进行最小化，并迭代执行；最后，他们在 E 的协方差矩阵具有特定结构的条件下，对 A 和 B 给出一个最大似然估计（MLE）程序，并同样采用序贯方式进行。理论结果表明，当 T 趋于无穷大且维度固定时，具有渐近正态性。然而，第一个程序较为繁琐，因为估计得到的矩阵非常大；另外两个程序则基于非凸最小化，且对于算法的极限点没有理论保证。

针对更一般的时间序列模型， [10] 对带有AIC和BIC惩罚项的最小二乘估计量和最大似然估计量进行了数值研究

$$Y_t = \sum_{\ell=1}^{L} A_\ell Y_{t-\ell} B_\ell + E_t, \quad t = 1, \ldots, T,$$

将其视为 $Y_t = A^* X_t B^* + E_t$，其中 X_t 是包含$L-$个过去观测矩阵 Y_{t-1}，……， Y_{t-L} 以及 $A^* =$（A_1，……， A_L）和 $B^* =$（$B_1^\top$，……， $B_L^\top$）的分块对角矩阵；它们是前一个方程中的拼接矩阵。

因此，我们的论文旨在应对高维数据和有限（非渐近）时间（例如 $T = 1$）的问题，从而为预测提供理论保证。

贡献。我们在第 2 节中证明，利用矩阵 $Y = U_Y \Sigma_Y V_Y^T$和 $X = U_X \Sigma_X V_X^T$的奇异值分解，可以将最小二乘程序简化为：将形式为 $A_0 \Sigma_X B_0$的预测器拟合到对角矩阵 Σ_Y上，并得到 A_0、 B_0和 A、B之间的显式关系。在对角形式下， A_0 和 B_0 的预测器有自然的选择。我们研究给定秩 r 时的这些预测器，并将其转换回 Y的原始空间，而不会损失预测速率。随后，我们给出一个数据依赖的秩选择器，并证明与之对应的预测器能够达到最优界。我们给出秩选择器保持一致的充分条件。最后，我们对该程序稍作修改，使其不再依赖噪声参数 σ，并证明在这种情况下可得到新的上界。在第 3 节中，我们研究核范数惩罚最小二乘，并证明它同样能够达到最优界。所有证明都集中在附录中的专门章节。最后，我们在第 4 节通过数值模拟展示这些运行速度快且形式显式的预测器具有出色的预测效果。

记号。对于任意整数 n 和 m，我们分别用 $n \wedge m$ 和 $n \vee m$ 表示 n 与 m 之间的最小值以及 n 与 m 之间的最大值。对于任意大小为 $n \times m$ 且秩为 r_M 的矩阵 M，我们将其奇异值分解（SVD）记为 $M = U_M \Sigma_M V_M^\top$，其中 U_M 属于 $\mathcal{O}_n$——大小为 $n \times n$ 的正交矩阵集合， V_M 属于 $\mathcal{O}_m$，并且

$$\Sigma_M = Diag_{n,m}(\sigma_k(M), 1 \le k \le r_M).$$

注意， $\sigma_1(M) \ldots, \sigma_{r_M}(M)$是 M的正奇异值，按降序排列； $n \times m$对角矩阵$Diag_{n,m}(\sigma_k(M)$ $1 \le k \le r_M)$的对角元素为列表中的

$n\wedge m$，其余位置为 0。此外，记 $\|M\|_F^2 = \sum_{k=1} \sigma_k(M)^2$ 为其弗罗贝尼乌斯范数，$\|M\|_{(2,q)}^2 = \sum_{k=1}^{q} \sigma_k(M)^2$ 为其 Ky-Fan $(2,q)$ 范数，$\|M\|_{op} = \sigma_1(M)$ 为其算子范数，$\|M\|_* = \sum_{k=1}^{n\wedge m} \sigma_k(M)$ 为其核范数，$M^\dagger$ 为其 Moore–Penrose 逆，r_M 为其秩，M^T 为其转置。对于任意矩阵 M_1 和 M_2，若它们属于 $\mathbb{R}^{n\times m}$，则 $\langle M_1, M_2\rangle_F$ 表示标准内积，即 $\langle M_1, M_2\rangle_F = \mathrm{Tr}(M_1^T M_2)$。对于任意 $r \in [r_M]$，我们将 $[M]_r$ 记为 M 在弗罗贝尼乌斯范数下的最佳秩 r 近似。在模型 (1) 中，记 r^* 为 $A^* X B^*$ 的秩。

2 秩惩罚学习

本节提出秩自适应预测器，并为其误差提供理论保证。首先，我们在已知参数矩阵秩的假设下给出显式预测器；随后，通过一个选择程序，可以得到依赖于数据的秩选择器及其对应的秩自适应预测器。尽管我们遵循了秩惩罚（单边）矩阵回归的经典结果，例如 [5], [9]and [3], ，但据我们所知，本文首次详细讨论固定秩的双边矩阵回归。令人惊讶的是，尽管该模型存在可识别性问题，仍然可以构造显式预测器。完成这一步后，我们再进行秩选择和秩自适应学习。

2.1 给定秩的预测

^ ^ 设 r 属于 $[n \wedge p \wedge r_X]$。下面构造显式预测器（A_r, B_r），它们是以下非凸最小化问题的解：

$$\min_{\substack{A,B:\\ \mathrm{rank}\, A\wedge \mathrm{rank}\, B\le r}} \|Y - AXB\|_F^2. \tag{3}$$

请注意，A 和 B 上的秩约束使用相同的值 r。事实上，我们的目标是构造一个针对信号 $A^* X B^*$ 的预测器，使其满足 $\mathrm{rank}(A^* X B^*) \le \min(r_{A^*}, r_X, r_{B^*})$。在证明结果的过程中，我们发现，风险上界仅通过 A^* 和 B^* 的较小秩取决于它们，而无法恢复两者中较大的秩。因此，将 A 和 B 设为具有相同秩，有助于将其作为一种降维技术，且不会对最终结果产生任何影响。

利用观测矩阵 Y 和设计矩阵 X 的奇异值分解，可以将模型 (1) 改写为

$$\Sigma_Y = A_0^* \cdot \Sigma_X \cdot B_0^* + E_0, \tag{4}$$

其中 $A_0^* = U_Y^T A^* U_X$、$B_0^* = V_X^T B^* V_Y$ 和 $E_0 := U_Y^T \cdot E \cdot V_Y$。特别地，当 E 的独立元素服从分布 $\mathcal{N}(0, \sigma^2)$ 时，E_0 也服从该分布，见引理 5.1。现在，Σ_Y 和 Σ_X 是对角矩阵，不一定是方阵，也不一定满秩。鉴于弗罗贝尼乌斯范数在左乘或右乘正交矩阵时保持不变，对于任意矩阵 $A \in \mathbb{R}^{n\times m}$ 和 $B \in \mathbb{R}^{q\times p}$，我们有

$$\|Y - AXB\|_F^2 = \|\Sigma_Y - A_0 \Sigma_X B_0\|_F^2,$$

其中 $A_0 = U_Y^T A U_X$ 和 $B_0 = V_X^\top B V_Y$ 是通过类似变换得到的，这些变换与用于关联真实底层参数的变换类似。

显然，矩阵 A 和 A_0 具有相同的秩，B 和 B_0 也是如此。因此，求解 (3) 等价于求解 A_{0r} 和 B_{0r} 的解：

$$\min_{\substack{A_0,B_0:\\ \operatorname{rank} A_0 \wedge \operatorname{rank} B_0 \le r}} \|\Sigma_Y - A_0 \Sigma_X B_0\|_F^2. \tag{5}$$

定理 2.1 令 $r \in [n \wedge p \wedge r_X]$

$$\hat{A}_{0r} = Diag_{n,m}(\sigma_k(Y), 1 \le k \le r \wedge r_Y) \quad and \quad \hat{B}_{0r} = Diag_{q,p}(\sigma_k(X)^{-1}, 1 \le k \le r). \tag{6}$$

那么，$(A_{0r}、B_{0r})$ 属于问题 (5) 的解集，并且预测器 $A_{0r}\Sigma_X B_{0r}$ 对于绝对常数 $C > 0$ 以及任意 $t > 0$ 满足预言机不等式

$$\|A_0^*\Sigma_X B_0^* - \hat{A}_{0r}\Sigma_X \hat{B}_{0r}\|_F^2 \le 9 \inf_{\substack{A_0,B_0:\\ \operatorname{rank} A_0 \wedge \operatorname{rank} B_0 \le r}} \|A_0^*\Sigma_X B_0^* - A_0\Sigma_X B_0\|_F^2 + 24C\sigma^2(1+t)^2 \cdot r(n+p),$$

以大于 $1 - 2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$ 的概率。

Next ，由 (5) 的显式解 (A_{0r}, B_{0r}) 可推导出 (3) 的显式解).

推论 2.2 令 $r \in [n \wedge p \wedge r_X]$

$$\hat{A}_r = U_Y \hat{A}_{0r} U_X^T \quad and \quad \hat{B}_r = V_X \hat{B}_{0r} V_Y^T, \tag{7}$$

其中 A_{0r} 和 B_{0r} 在 (6) 中定义。则 $(A_r、B_r)$ 是问题 (3) 的解，且预测器 $A_r X B_r$ 满足：对于某个绝对常数 $C > 0$，以及任意 $t > 0$，有如下预言机不等式

$$\|A^*XB^* - \hat{A}_r X \hat{B}_r\|_F^2 \le 9 \inf_{\substack{A,B:\\ \operatorname{rank} A \wedge \operatorname{rank} B \le r}} \|A^*XB^* - AXB\|_F^2 + 24C\sigma^2(1+t)^2 \cdot r(n+p),$$

其概率大于 $1 - 2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$。

定理 2.1 和推论 2.2 的证明见第 5 节。在证明中，我们将偏差表示为未知矩阵参数的函数：

$$\inf_{\substack{A,B:\\ \operatorname{rank} A \wedge \operatorname{rank} B \le r}} \|A^*XB^* - AXB\|_F^2 = \sum_{k=r+1}^{r^*} \sigma_k(A^*XB^*)^2 \cdot \mathbf{1}_{r<r^*}.$$

请注意，我们选择的这对预测器（A_{0r}，B_{0r}）并非唯一，而且很容易推导出问题（5）的解族。每个解族都可以转化为

问题（3）的一个解。事实上，考虑（$\alpha\hat{}A_{0r}$，$\frac{1}{\alpha}\hat{}B_{0r}$），其中 $\alpha > 0$ 任意。或者，令 λ_i 对所有 $i \le m \wedge q$ 都是任意正数，则

$$(\hat{A}_{0r} Diag_{m,m}(\lambda_1, \ldots, \lambda_{m\wedge q}), Diag_{q,q}(\lambda_1^{-1}, \ldots, \lambda_{m\wedge q}^{-1})\hat{B}_{0r})$$

给出相同的预测。我们来看看，对参数矩阵 A_0^*和 B_0^*应用相同的变换，也会得到相同的信号矩阵 $A_0^*\Sigma_X B_0^*$。事实上，该模型不可识别，因此，若没有进一步的强假设，我们只能希望学习全局信号，而不能学习模型的参数。

替代预测器。我们来定义第二对预测器（$\tilde{A}_r$，$\tilde{B}_r$）。$\tilde{A}_r$，$\tilde{B}_r$ 与 $\tilde{A}_0$，$\tilde{B}_0$ 完全相同，且具有相同的理论性质，但其优势在于 A 是满秩的，且不依赖于 r。定义

$$\tilde{A}_0 = I_{n,m} \quad \text{and} \quad \tilde{B}_{0r} = Diag_{q,p}\left(\frac{\sigma_k(Y)}{\sigma_k(X)}, 1 \le k \le r \wedge r_Y\right)$$

其中 $I_{n,\ m}$ 表示维数为 $n \times m$的单位矩阵，而 B_{0r} 的秩为 $r \wedge r_Y$。利用类似变换，我们得到

$$\tilde{A} = U_Y I_{n,m} U_X^T \quad \text{and} \quad \tilde{B}_r = V_X \tilde{B}_{0r} V_Y^T.$$

很容易看出，定理 2.1 对 A_0 和 B_{0r}成立，而推论 2.2 对 A 和 B_r成立。

2.2 秩自适应预测

本节中，我们提出秩自适应预测器 ($A_{\hat{r}}$，$B_{\hat{r}}$)，它们通过与 [5] 类似的模型选择程序，从族 $\{(A_r,\ B_r) : r \in [n \wedge p \wedge r_X]\}$ 中选出。首先，对于一般矩阵 M 和任意 $\lambda > 0$，定义 M 的 $\lambda-$ 秩如下：

$$r_M(\lambda) = 1 \vee \sum_{k=1}^{\operatorname{rank} M} \mathbf{1}_{\sigma_k(M)^2 \ge \lambda}.$$

对于给定的 $\lambda > 0$，令

$$\hat{r} := \arg \min_{r \in [n\wedge p \wedge r_X]} \left\{ \|Y - \hat{A}_r X \hat{B}_r\|_F^2 + \lambda r \right\}. \tag{8}$$

考虑式 (7) 中针对式 (8) 所定义的数据驱动的秩 $\hat{r}$ 引入的预测器。接下来的定理将预言机不等式推广到与估计秩 $\hat{r}$ 以及某个足够大的 $\lambda > 0$ 相对应的秩自适应预测器 ($A_{\hat{r}}$，$B_{\hat{r}}$)。

定理 2.3 与式 (8) 中的 $\hat{r}$ 以及满足以下条件的 λ 相对应的秩自适应预测器$(A_{\hat{r}}, B_{\hat{r}})$：存在某个绝对常数 $C > 0$，使得对任意 $t > 0$, $\lambda \ge 4C(1+t)^2\sigma^2(n+p)$，均满足预言机不等式

$$\|A^*XB^* - \hat{A}_{\hat{r}} X \hat{B}_{\hat{r}}\|_F^2 \le \min_{r \in [n\wedge p \wedge r_X]} \left\{ 9 \sum_{k=r+1}^{r^*} \sigma_k(A^*XB^*)^2 \cdot \mathbf{1}_{r<r^*} + 6\lambda r \right\},$$

其概率大于 $1 - 2\exp(-t^2(\sqrt{n} + \sqrt{p})^2)$。

注意，前一个公式右侧的最小值始终小于 $r = r^*$ 处的值，因此在定理 2.3 的假设下，

$$\|A^*XB^* - \hat{A}_{\hat{r}}X\hat{B}_{\hat{r}}\|_F^2 \leq 6r^*\lambda,$$

以大于 $1 - 2\exp(-t^2(\sqrt{n} + \sqrt{p})^2)$ 的概率。
我们的方法所达到的 $r^*(n+p)$ 阶界与 [19] 和 [9]中的低秩矩阵回归模型的阶界类似。事实上，2MR 模型比 MR 模型更难，（即 其中一个矩阵是已知的），为了得到正确的下界，我们将假设秩较大的矩阵是已知的。因此，低秩 MR 模型的预测下界对我们的模型同样有效。

2.3 一致秩选择

我们研究（8）中的秩选择器 $\hat{r}$ 的一致性，并考察其何时能以高概率恢复真实秩r^* 。首先，我们证明，对于适当选择的 λ，数据驱动的秩 $\hat{r}$ 实际上是唯一解，并且与 $\lambda-$ Y 的秩一致，即 $\hat{r} = r_Y(\lambda)$。

命题 2.4 如果 $\lambda > \sigma_{r_Y}(Y)^2$, 优化问题 (8) 存在唯一解 $\hat{r}$，且它实际上是 Y 的 λ–秩，即 $\hat{r} = r_Y(\lambda)$。

接下来，我们证明 $\hat{r}$ 以高概率恢复 A^*XB^* 的 λ–秩。

命题 2.5 设 $\lambda > 0$，并用 $r^*(\lambda)$ 表示 A^*XB^* 的 λ–秩。若对于某个 *(0,1)* 内的常数 c，$\sigma_{r^*(\lambda)}(A^*XB^*)^2 > (1+c)^2\lambda$ 且 $\sigma_{r^*(\lambda)+1}(A^*XB^*)^2 < (1-c)^2\lambda$，则

$$\mathbb{P}(\hat{r} = r^*(\lambda)) \geq \mathbb{P}(\|E\|_{op}^2 \leq c^2\lambda).$$

特别地，若 $\lambda \geq 2C(n+p)\sigma^2(1+t)^2/c^2$ 对于某个绝对常数 $C > 0$ 以及任意$t > 0$均成立，则 $\hat{r} = r^*(\lambda)$ 以大于 $1 - 2\exp(-t^2(\sqrt{n} + \sqrt{p})^2)$ 的概率成立。

最后，请记住，$r^*(\lambda)$与真实潜在秩 r^* 相一致，这等价于满足 $\sigma_{r^*}(A^*XB^*)^2 \geq \lambda > 0$。如果 λ也满足 $\sigma_1(E)^2 \leq c^2\lambda$，其中某个绝对常数为 $c > 0$，那么秩选择器就会与 r^* 相一致。因此，信噪比（此处由 $\sigma_{r^*}(A^*XB^*)^2/\sigma_1(E)^2$ 给出）必须足够显著，才能由 $\hat{r}$选出真实潜在秩。结合前面的命题，我们得到如下结果。

命题 2.6 设 $\lambda > 0$。若存在某个常数 c，使得 c属于 *(0,1)*，且 $\sigma_{r^*}(A^*XB^*)^2 > (1+c)^2\lambda$，则

$$\mathbb{P}(\hat{r} = r^*) \geq \mathbb{P}(\|E\|_{op}^2 \leq c^2\lambda).$$

特别地，如果 $\lambda \geq 2C(n+p)\sigma^2(1+t)^2/c^2$ 对某个绝对常数 $C > 0$ 以及任意$t > 0$成立，那么$\hat{r} = r^*$ 以大于 $1 - 2\exp(-t^2(\sqrt{n} + \sqrt{p})^2)$的概率成立。

2.4 数据驱动的秩自适应预测

在条件 $\lambda \geq 4C(1+t)^2\sigma^2(n+p)$ 成立时，(8) 中的秩选择器 $\hat{r}$ 用于构造一致预测器，详见定理 2.3。然而，在一般情形下，噪声参数σ 未知。因此，在这些情况下，需要一种数据依赖的秩选择器来构造一致预测器。受前一种情形的启发（其中假设 σ^2 已知），我们按如下方式进行。首先，将惩罚项改为 $\lambda \cdot r\widehat{\sigma}_r^2$，其中

$$\widehat{\sigma}_r^2 = \frac{1}{np}\|Y - \hat{A}_r X \hat{B}_r\|_F^2.$$

请注意，在高斯噪声的特殊情况下，$\widehat{\sigma}_r^2$ 估计噪声的方差 σ^2 。接下来，给定真实秩 $r_{max} \leq n \wedge p \wedge r_X$的最大可能值，我们定义数据驱动的秩选择器

$$\bar{r} := \arg \min_{r \in [r_{max}]} \left\{ \|Y - \hat{A}_r X \hat{B}_r\|_F^2 + \lambda \cdot r\widehat{\sigma}_r^2 \right\}. \tag{9}$$

最后，我们使用预测器（$A_{\bar{r}}$， $B_{\bar{r}}$）。下一个定理将定理 2.3 的上界推广到这些数据驱动的秩自适应预测器。

定理 2.7 与 $\bar{r}$ 在(9) 中且满足$r_{max} \leq n \wedge p \wedge r_X$的条件相关联的数据驱动的秩自适应预测器$(A_{\bar{r}}, B_{\bar{r}})$，以及与 $\lambda = (1+\varepsilon)np/(r_{max} \vee r_Y)$ 相关联的数据驱动的秩自适应预测器（其中某个 $\varepsilon > 0$），对于某个绝对常数 $C > 0$ 以及任意 $t > 0$，满足预言机不等式

$$\begin{aligned}\|A^*XB^* - \hat{A}_{\bar{r}} X \hat{B}_{\bar{r}}\|_F^2 \leq \min_{r \in [r_{max}]} \left\{ 9\|A^*XB^* - \hat{A}_r X \hat{B}_r\|_F^2 + 6(1+\varepsilon) \cdot r\sigma_{r+1}(A^*XB^*)^2 \right\} \\ + 12C(2+\varepsilon)(1+t)^2 \cdot \sigma^2 r_{max}(n+p),\end{aligned}$$

其概率大于 $1 - 2\exp(-t^2(\sqrt{n} + \sqrt{p})^2)$。

应用推论 2.2，在定理 2.7 的假设下可得

$$\begin{aligned}\|A^*XB^* - \hat{A}_{\bar{r}} X \hat{B}_{\bar{r}}\|_F^2 \leq \min_{r \in [r_{max}]} \left\{ 9^2 \inf_{\substack{A,B: \\ r_A \wedge r_B \leq r}} \|A^*XB^* - A_r X B_r\|_F^2 + 6(1+\varepsilon) \cdot r\sigma_{r+1}(A^*XB^*)^2 \right\} \\ + 12(20+\varepsilon)C(1+t)^2 \cdot \sigma^2 r_{max}(n+p),\end{aligned}$$

其概率大于 $1 - 2\exp(-t^2(\sqrt{n} + \sqrt{p})^2)$。

注意，右侧前一个式子中的最小值总是小于其在 $r = r^*$ if r_{max} is larger than r^* 时的取值，因此在定理 2.7 的假设下可得

$$\|A^*XB^* - \hat{A}_{\bar{r}} X \hat{B}_{\bar{r}}\|_F^2 \leq 12(20+\varepsilon)C(1+t)^2 \cdot \sigma^2 r_{max}(n+p).$$

为了与先前的结果进行比较，请注意，定理 2.3 推导出的关于值 $r = r^*$和最小值 $\lambda = 4C(1+t)^2\sigma^2(n+p)$ 的上界非常相似

$$\|A^*XB^* - \hat{A}_{\hat{r}} X \hat{B}_{\hat{r}}\|_F^2 \leq 24C(1+t)^2 \cdot \sigma^2 r^*(n+p).$$

从计算角度来看，在某些情况下最好更改 $\widehat{\sigma}_r^2$ 。例如，在我们的数值模拟中，我们使用

$$\widehat{\sigma}_r^2 = \frac{1}{np-(m\wedge q)r_X}\|Y-\hat{A}_r X\hat{B}_r\|_F^2$$

当 $n\geq m$、 $p\geq q$，因此 $np>(m\wedge q)r_X$。考虑 $\lambda=(1+\varepsilon)(np-(m\wedge q)r_X)/(r_{max}\vee r_Y)$，即可直接证明定理 2.7 的对应结论。

3 核范数惩罚学习

众所周知，核范数惩罚最小二乘具有良好的性质，参见 [1] 或 [16]。因此，它可能优于秩惩罚方法。我们将核范数惩罚（NNP）优化问题定义如下:

$$\min_{A,B}\|Y-AXB\|_F^2+2\lambda\cdot\|AXB\|_*, \tag{10}$$

其中某个 $\lambda>0$。该优化问题的目标函数关于 A 和 B 并非联合凸函数。注意，在矩阵回归中（当 A^* 为单位矩阵时）， XB 的核范数已被采用，参见 [14]，或其他依赖于特征矩阵 X、 [15] 的自适应形式。不过，我们给出了属于该问题解集的显式预测器，并证明了它们满足一个预言机不等式。

定理 3.1 由(A,B) 定义的预测器

$$\bar{A}=U_Y I_{n,m}U_X^\top \quad and \quad \bar{B}=V_X\cdot Diag_{q,p}\left(\frac{(\sigma_k(Y)-\lambda)_+}{\sigma_k(X)},1\leq k\leq r_Y\wedge r_X\right)V_Y^\top \tag{11}$$

是问题(10)的解。此外，若 λ 满足：存在某个绝对常数$C>0$，且对于任意 $t>0$, $\lambda\geq 2C(1+t)^2\sigma^2(n+p)$，它们满足预言机不等式

$$\|A^*XB^*-\bar{A}X\bar{B}\|_F^2\leq 9\min_{r\in[n\wedge p\wedge r_X]}\left\{\sum_{k=r+1}^{r^*}\sigma_k(A^*XB^*)^2\cdot\mathbf{1}_{r<r^*}+16\lambda r\right\},$$

其成立概率大于 $1-2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$。

证明见第 5 节。

备注。 另一种方法是考虑向量化形式 (2) 下的模型，并求解问题

$$\min_{A,B}\|vec(Y)^\top-vec(X)^\top\cdot A^\top\otimes B\|_2^2+2\lambda\|A^\top\otimes B\|_*,$$

对于某个 $\lambda>0$。回想一下， $A^\top\otimes B$ 表示矩阵 $A^\top$ 和 B 的张量积，我们可以写作 $\|A^\top\otimes B\|_*=\sum_{k,\ j\geq 1}\sigma_k(A)\sigma_j(B)$。然而，特征是一维的

并且我们丢失了原始矩阵 X中包含的结构化信息。对于 t in $[T]$中的重复观测（Y_t, X_t），这种方法或许更有意义：将行 $vec(Y_t)^\top$ 和 $vec(X_t^\top)$分别堆叠成矩阵 $\mathbb{Y}$ 和 $\mathbb{X}$，然后进行经典矩阵回归。即便如此，在此情境下，为了获得良好的预测而对特征矩阵 $\mathbb{X}$ 作出的通常假设并不合理，因为这些假设与 $[T]$中的原始矩阵数据集 X_t, t 并无太大关联。

备注（可识别性的充分条件）我们曾多次指出，对于给定的矩阵M，许多矩阵对（A, B）都能满足方程 $M = AXB$。给定矩阵 M的奇异值分解后，我们可以通过选择由 $A = U_M A_0 U_X^\top$和$B = V_X B_0 V_M^\top$给出的解（A, B）来降低问题的维数，其中 A_0和 B_0为对角矩阵，使得

$$\sigma_k(A)\sigma_k(X)\sigma_k(B) = \sigma_k(M), \quad \text{for all } k \le r_X \wedge r_A \wedge r_B.$$

因此，即使采用对角形式，我们也只能识别 A 和 B各自奇异值的乘积。只有在非常严格的条件下，我们才有望识别矩阵 A 和 B： $X^\top X$ 为满秩，且假设矩阵 A 或矩阵 B 的奇异值已知，例如，它是一个奇异值为 1 或 0 的投影算子。矩阵分解领域中，文献还报道了少数其他可识别的设定，例如非负矩阵分解（NMF），参见 [7], 主题模型的非负矩阵分解 [11], [2], [13] 或协方差矩阵分解 [8]。

4 数值结果

设观测矩阵 Y的维度为 $n = 100$ 和 $p = 300$，设计矩阵 X 的维度为 $m = 50$ 和 $q = 60$。我们随机生成三个矩阵： A^*、B^*和 X，其元素相互独立且服从均值为 0、方差为 1 的高斯分布。随后，将这些矩阵投影到最佳低秩矩阵近似上，其中矩阵A^* 的秩为 $r_A^* = 16$，矩阵 B^* 的秩为 $r_B^* = 12$，矩阵 X 的秩为 $r_X = 25$。信号矩阵定义为 A^*XB^*，在所有实验中的秩均为 12。我们还为高斯噪声 E的方差 σ^2 设定了不同取值，使得信噪比 $SNR := \sigma_{r^*}(A^*XB^*)^2/\sigma_1(E)^2$ 大致在 $[0.5, 2]$范围内变化。

图 1：其实，预测器的预测性能

对于 $\hat{X}_r$ 的不同的 r。当 $\sigma < 8$ 使 SNR 约为 1 以上时，预测风险随秩增加而降低，在秩达到 12 之前始终不超过 12；秩超过 12 后，预测风险则随秩增加而上升。当 $\sigma \ge 8$ 使 SNR 低于 1 时，预测风险随秩增加而降低，在秩达到 11 之前始终不超过 11；秩超过 11 后，预测风险则随秩增加而上升。这表明，当噪声方差较低（即$\sigma < 8$）时，$r = r^* = 12$ 对应的预测器效果最佳；而当噪声方差较高（即 $\sigma \ge 8$）时， $r = 11$ 对应的预测器效果最佳。这说明在噪声较强时存在明显的过拟合现象，因此在这种情况下，适当低估秩会更好。

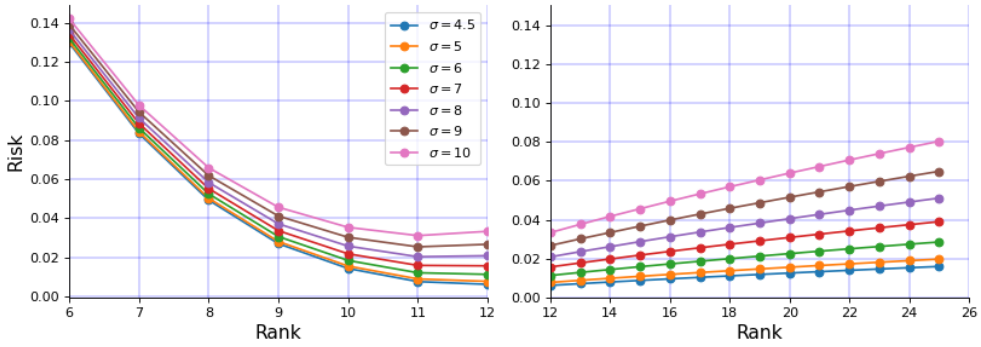

图1：不同σ取值下，风险$\frac{\|\hat{A}_r X \hat{B}_r - A^* X B^*\|_F^2}{\|A^* X B^*\|_F^2}$随 r 的变化

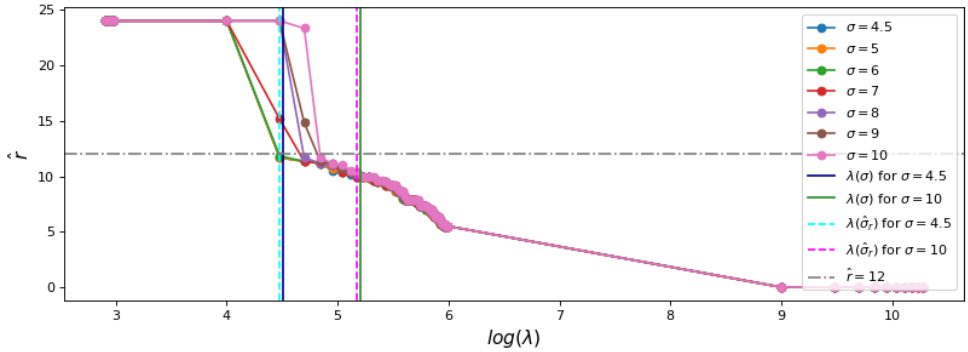

图 2：不同取值下，估计的$\hat{r}$ 随 log(λ) 的变化 σ

图2展示了对于不同的 λ 值，按式 (8) 定义的预测值 $\hat{r}$。无论噪声方差水平如何，当 λ 较小时，估计值 $\hat{r}$ 最大，并且存在 $\hat{r} = r_X = 25$。这说明了前文所述的过拟合现象，即秩 r 越高，误差 $\|Y - \hat{A}_r X \hat{B}_r\|_F^2$ 越低。随着 λ 增大，最小化过程中对秩 r 的惩罚变得更加重要，$\hat{r}$ 随之减小。然而，对于中等大小的 λ （即 约为 $\log(\lambda) \leq 5$），噪声方差水平 σ 越低，$\hat{r}$ 下降得越快。最终，对于较大的 λ （即 约为 $\log(\lambda) > 5$），$\hat{r}$ 随 λ 变化的衰减速率不再取决于 σ。

λ 的数值是一个重要问题。我们给出了用于选择该调节参数的明确（计算快速）过程。在已知噪声方差的情况下， [4] 提出的经验法则建议选择

$$\lambda(\sigma) = 2C(n+p)\sigma^2(1+t)^2$$

在定理 2.3 中， $t = 0$，且 $C = 2$。两条实线表示 λ(4.5)（蓝色）和 λ(10)（绿色）。采用这些调节参数，我们成功估计了信号的潜在秩 $\hat{r} \approx 12 = r^*$。需要强调的是，在低噪声情形下，秩会略微高估；在高噪声情形下，秩则会略微低估。这一现象与图1所得结果完全吻合：在低噪声情形下，高估秩不会影响性能，而在高噪声情形下，略微低估秩则能提升性能。

然而，在实际应用中，噪声的方差未知。这就引出了一个问题：在这种情况下，如何选择数据驱动的 λ ，同时又不损害预测性能？这一情形更具挑战性，因为首先需要估计 σ^2 ，然后才能使用前面介绍的经验法则。我们选择 r 的初始值为 $r_X \wedge n \wedge p$，并提出依赖于 r 的估计量 $\hat{\sigma}_r^2 := \frac{\|Y - \hat{A}_r X \hat{B}_r\|_F^2}{np - (m \wedge q)\, r_X \hat{}\,(r)}$。它可以计算先前定义的 $\lambda(\hat{\sigma}_r)$；利用这一数据驱动的调节参数，我们得到秩估计量 $\bar{r}$。该过程以 r 为参数，并返回 $\lambda(\hat{\sigma}_r)$和 $\bar{r}$。然而，当$\hat{}\,\hat{}$ r 明显大于 r^*时， $A_r X B_r$ 会对 Y 进行过拟合，仅执行一次该过程无法得到令人满意的输出 $\bar{r}$。因此，我们在 $\bar{r} < r$条件成立时迭代。我们将该过程的最终输出记为 $\lambda(\hat{\sigma}_{\bar{r}})$和 $\bar{r}$ 。两条虚线表示当 $\sigma = 4.5$分别取$\hat{\sigma}_{\bar{r}}$（青色）和 $\sigma = 10$ （洋红色）时的 $\lambda(\hat{\sigma}_{\bar{r}})$。所提出的过程具有优异的数值性质。

最后，在相同背景下针对不同真实潜在秩值开展的数值模拟，显示出类似的出色预测界，并且秩选择正确。结合当前 $\min(r_A^*,\ r_X,\ r_B^*) = r_B^*$ 的情况，我们还成功探讨了 $\min(r_A^*,\ r_X,\ r_B^*) = r_A^*$、$\min(r_A^*,\ r_X,\ r_B^*) = r_X$和 $\min(r_A^*,\ r_X,\ r_B^*) = r_A^* = r_X = r_B^*$ 的情况。

5 证明

基本事实对于任意矩阵 $M \in \mathbb{R}^{n\times m}$, $\|M\|_*^2 \leq r_M \|M\|_F^2$。此外，对于 $\mathbb{R}^{n\times m}$ 中的任意矩阵 M_1 和 M_2 ，以下不等式成立： $\langle M_1, M_2\rangle_F \leq \|M_1\|_* \|M_2\|_{op}$ 和 $\|M_1 +$

$M_2\|_F \leq \|M_1\|_F + \|M_2\|_F$。此外，若设 $a =$秩 $M_1 \wedge$秩 M_2，则 $\langle M_1, M_2\rangle_F \leq \|M_1\|_{(2,a)}\|M_2\|_{(2,a)}$。

引理 5.1 设 E 为一个 $n \times p$ 随机矩阵，其元素相互独立且服从高斯分布 $\mathcal{N}(0, \sigma^2)$。若 U 和 V分别属于 $\mathcal{O}_n$ 和 $\mathcal{O}_p$，则 $E_0 := U^\top EV$的元素相互独立，且服从高斯分布 $\mathcal{N}(0, \sigma^2)$。

引理 5.1 的证明。注意，我们可以将矩阵 E_0 向量化，从而得到

$$vec(E_0) = (V^\top \otimes U^\top) \cdot vec(E),$$

其中 $vec(E)$ 是一个维数为 np、均值为零、方差为 $\sigma^2 I_{np}$的高斯向量。此外，张量积 $V^\top \otimes U^\top$ 属于 $\mathcal{O}_{np}$，因此 $vec(E_0)$ 仍是一个服从分布 $\mathcal{N}_{np}(0, \sigma^2 I_{np})$ 的高斯向量。 ■

回忆一下，对于任意矩阵 M，我们用 $U_M \Sigma_M V_M^\top$表示它的奇异值分解。

引理 5.2 若 M^* 是一个秩为 r^*的 $n \times p$ 矩阵，则对于任意 $r \leq n \wedge p$，有

$$\inf_{M:\text{rank}\, M \leq r} \|M - M^*\|_F^2 = \sum_{k=r+1}^{r^*} \sigma_k(M^*)^2 \cdot \mathbf{1}_{r<r^*},$$

且下确界由空间中秩为 r 的矩阵 $n \times p$ 上的 M^* $[M^*]_r$ 投影取得，该投影由矩阵给出

$$[M^*]_r = U_{M^*} \cdot Diag_{n,p}(\sigma_1(M^*), ..., \sigma_{r\wedge r^*}(M^*)) \cdot V_{M^*}^\top.$$

5.1 定理 2.1 的证明

^^ 令 $r \in [n \wedge p \wedge r_X]$ 和 (A_{0r}, B_{0r}) 如式 (6) 所定义。这里记 $M_0^* = A_0^* \Sigma_X B_0^*$ 和 ^^^^ $M_0 = A_{0r}\Sigma_X B_{0r}$。根据构造，$M_0$ 是 Σ_Y 在秩小于或等于 r 的矩阵集合上的 $[\Sigma_Y]_r$ 投影，其含义见引理 5.2。因此，

$$\|\Sigma_Y - \hat{M}_0\|_F^2 \leq \|\Sigma_Y - [M_0^*]_r\|_F^2$$

回顾一下，在我们的模型中 $\Sigma_Y = M_0^* + E_0$，这将导出

$$\|M_0^* - \hat{M}_0 + E_0\|_F^2 \leq \|M_0^* - [M_0^*]_r + E_0\|_F^2.$$

展开平方项并整理各项，得到

$$\|M_0^* - \hat{M}_0\|_F^2 \leq \|M_0^* - [M_0^*]_r\|_F^2 + 2\langle \hat{M}_0 - [M_0^*]_r, E_0\rangle_F.$$

^^ 现在，由于秩$(M_0) = r$且秩$([M_0^*]_r) \leq r$，我们得到秩$(M_0 - [M_0^*]_r) \leq 2r$。这个不等式给出

$$\begin{aligned}
\|M_0^* - \hat{M}_0\|_F^2 &\leq \|M_0^* - [M_0^*]_r\|_F^2 + 2\|E_0\|_{(2,2r)} \cdot \|\hat{M}_0 - [M_0^*]_r\|_{(2,2r)} \\
&\leq \|M_0^* - [M_0^*]_r\|_F^2 + 2\|E_0\|_{(2,2r)} \cdot \|\hat{M}_0 - [M_0^*]_r\|_F \\
&\leq \|M_0^* - [M_0^*]_r\|_F^2 + 2\|E_0\|_{(2,2r)} \cdot \left(\|\hat{M}_0 - M_0^*\|_F + \|M_0^* - [M_0^*]_r\|_F\right).
\end{aligned}$$

我们应用不等式 $2xy \le \alpha x^2 + \alpha^{-1}y^2$，其中 $x, y \ge 0$ 且 $\alpha > 0$。对于实数 $\alpha > 1$ 和 $\beta > 0$, 我们得到:

$$(1-\alpha^{-1}) \cdot \|M_0^* - \hat{M}_0\|_F^2 \le (1+\beta^{-1}) \cdot \|M_0^* - [M_0^*]_r\|_F^2 + (\alpha+\beta) \cdot \|E_0\|_{(2,2r)}^2.$$

利用 $\|E_0\|_{(2,2r)}^2 \le 2r \cdot \|E_0\|_{op}^2$ 和引理 5.2，我们可以进一步得到

$$\|M_0^* - \hat{M}_0\|_F^2 \le \frac{1+\beta^{-1}}{1-\alpha^{-1}} \cdot \inf_{M:\operatorname{rank} M \le r} \|M_0^* - M\|_F^2 + \frac{\alpha+\beta}{1-\alpha^{-1}} \cdot 2r\|E_0\|_{op}^2. \tag{12}$$

注意，对于任意秩小于或等于 r 的矩阵 A_0, B_0，都有 $\operatorname{rank}(A_0\Sigma_X B_0) \le r_{A_0} \wedge r_X \wedge r_{B_0} \le r$，由此可得

$$\inf_{M:\operatorname{rank} M \le r} \|M_0^* - M\|_F^2 \le \inf_{\substack{A_0, B_0: \\ \operatorname{rank} A_0 \wedge \operatorname{rank} B_0 \le r}} \|M_0^* - A_0\Sigma_X B_0\|_F^2.$$

事实上，第二个下确界是在一个可能更小的矩阵集合上取的。我们实际上证明了前一个表达式中的等号成立。事实上，由引理 5.2 可知，下确界 $\inf_{M:\operatorname{rank} M \le r} \|M_0^* - M\|_F^2 = \sum_{k=r+1}^{r^*} \sigma_k(M_0^*)^2 \cdot \mathbf{1}_{r<r^*}$，其中 $r^* = \operatorname{rank}(M_0^*)$。回顾一下，$M_0^* = A_0^*\Sigma_X B_0^*$是对角矩阵的乘积，因此有 $r^* = \min(r_X, r_{A_0^*}, r_{B_0^*})$，并且$\sigma_k(M_0^*) = \sigma_k(A_0^*)\sigma_k(X)\sigma_k(B_0^*) \cdot \mathbf{1}_{k \le r^*}$。因此，特定的选择

$$A_{0r} := Diag_{n,m}(\sigma_1(A_0^*), \ldots, \sigma_{r \wedge r_{A_0^*}}(A_0^*)) \text{ and } B_{0r} := Diag_{q,p}(\sigma_1(B_0^*), \ldots, \sigma_{r \wedge r_{B_0^*}}(B_0^*))$$

恰好求解了给出 $M_0^* = A_{0r}\Sigma_X B_{0r}$ 的问题。最后，

$$\inf_{M:\operatorname{rank} M \le r} \|M_0^* - M\|_F^2 = \inf_{\substack{A_0, B_0: \\ \operatorname{rank} A_0 \wedge \operatorname{rank} B_0 \le r}} \|M_0^* - A_0\Sigma_X B_0\|_F^2. \tag{13}$$

将其代入 (12)，并考虑特定选择 $\alpha = 3/2$ 和 $\beta = 1/2$，即可得到定理:

$$\|A_0^*\Sigma_X B_0^* - \hat{A}_{0r}\Sigma_X \hat{B}_{0r}\|_F^2 \le 9 \inf_{\substack{A_0, B_0: \\ \operatorname{rank} A_0 \wedge \operatorname{rank} B_0 \le r}} \left(\|A_0^*\Sigma_X B_0^* - A_0\Sigma_X B_0\|_F^2\right) + 12r\|E_0\|_{op}^2.$$

最后一步是关于 $\|E_0\|_{op}$ 的高概率界。回顾一下，$E_0 = U_Y^\top E V_Y$ 在 $\mathcal{O}_n$ 中满足 U_Y，而 V_Y 在 $\mathcal{O}_p$ 中满足 E_0，因此 E 和 $\|E\|_{op} = \|E_0\|_{op}$ 具有相同的奇异值。因此 $\|E\|_{op} = \|E_0\|_{op}$。噪声矩阵 E 具有相互独立、均值为零的 $\sigma-$ 次高斯元素，其谱范数满足（见 [20]），其中 $C > 0$ 是某个绝对常数

$$\mathbb{P}\left(\|E\|_{op}^2 \le 2C\sigma^2 \cdot (1+t)^2(n+p)\right) \ge 1 - 2e^{-t^2(\sqrt{n}+\sqrt{p})^2}, \quad \text{for any } t > 0. \tag{14}$$

此外，$\mathbb{E}\left[\|E\|_{op}\right] \le \sqrt{C}\sigma(\sqrt{n} + \sqrt{p})$。

5.2 推论 2.2 的证明

^^^^^ 回顾记号 $M_0^* = A_0^* \Sigma_X B_0^*$ 和 $M_0 = A_{0r}\Sigma_X B_{0r}$，其中 A_{0r} 和 B_{0r} 由 (6) 给出；^^^^^ 并记 $M^* = A^* X B^*$ 和 $M = A_r X B_r$，其中 A_r 和 B_r 由 (7) 给出。注意，弗罗贝尼乌斯范数和秩在左乘或右乘正交矩阵时保持不变。因此，我们沿用定理 2.1 的证明思路，可得 $\|Y - \hat{M}\|_F^2 = \|\Sigma_Y - \hat{M}_0\|_F^2$，且 $\operatorname{rank} M^* = \operatorname{rank} M_0^* = r^*$。此外，$\hat{M}$ 是将 Y 投影到秩小于或等于 r 的矩阵空间上的投影 $[Y]_r$。最后，等式 (13) 可以进一步推出

$$\inf_{M:\operatorname{rank} M \le r} \|M_0^* - M\|_F^2 = \inf_{\substack{A_0, B_0: \\ \operatorname{rank} A_0 \wedge \operatorname{rank} B_0 \le r}} \|M_0^* - A_0 \Sigma_X B_0\|_F^2 = \inf_{\substack{A, B: \\ \operatorname{rank} A \wedge \operatorname{rank} B \le r}} \|M^* - AXB\|_F^2.$$

事实上，我们分别有从 A_0, B_0 到 A, B 的一一变换，且弗罗贝尼乌斯范数相等。证明至此完成。

5.3 定理 2.3 的证明

根据 $\hat{r} = \hat{r}(\lambda)$ 的定义，对于所有 $r \in [n \wedge p \wedge r_X]$，我们有：

$$\|Y - \hat{A}_{\hat{r}} X \hat{B}_{\hat{r}}\|_F^2 + \lambda \hat{r} \le \|Y - \hat{A}_r X \hat{B}_r\|_F^2 + \lambda r.$$

^^由于 $A_r X B_r$ 是 Y 在秩为 $M \le r$ 的矩阵空间 M 上的投影 $[Y]_r$，因此对于所有满足 $\operatorname{rank} A \wedge \operatorname{rank} B \le r$ 的矩阵 A 和 B，我们有：

$$\|Y - \hat{A}_r X \hat{B}_r\|_F^2 \le \|Y - AXB\|_F^2.$$

事实上，$\operatorname{rank}(AXB) \le r$，而勾股定理给出了前一个不等式。由此可得：

$$\|Y - \hat{A}_{\hat{r}} X \hat{B}_{\hat{r}}\|_F^2 + \lambda \hat{r} \le \|Y - AXB\|_F^2 + \lambda r.$$

接下来，代入 $Y = A^* X B^* + E$，展开平方项并重新整理各项，得到

$$\begin{aligned} \|A^* X B^* - \hat{A}_{\hat{r}} X \hat{B}_{\hat{r}}\|_F^2 \le{}& \|A^* X B^* - AXB\|_F^2 + \lambda(r - \hat{r}) \\ &+ 2\langle E, \hat{A}_{\hat{r}} X \hat{B}_{\hat{r}} - AXB \rangle. \end{aligned}$$

^^^^^ 记 $M(\hat{r}) = A_{\hat{r}} X B_{\hat{r}}$, $M(r) = AXB$，并注意到 $\operatorname{rank}(M(\hat{r}) - M(r)) \le \hat{r} + r$。我们有

$$\begin{aligned} \langle E, \hat{A}_{\hat{r}} X \hat{B}_{\hat{r}} - AXB \rangle &\le \|E\|_{op} \cdot \|\hat{M}(\hat{r}) - M(r)\|_* \\ &\le \|E\|_{op} \cdot \sqrt{\hat{r} + r} \|\hat{M}(\hat{r}) - M(r)\|_F \\ &\le \|E\|_{op} \cdot \sqrt{\hat{r} + r} (\|M^* - \hat{M}(\hat{r})\|_F + \|M^* - M(r)\|_F). \end{aligned}$$

然后，将不等式 $2xy \le \alpha x^2 + \alpha^{-1}y^2$ 分别与 $x, y \ge 0$ 和 $\alpha > 0$结合使用两次，对于任意实数 $\alpha > 1, \beta > 0$，可得:

$$(1-\alpha^{-1})\|M^* - \hat{M}(\hat{r})\|_F^2 \le (1+\beta^{-1})\|M^* - M(r)\|_F^2 + (\alpha+\beta)\|E\|_{op}^2(r+\hat{r}) + \lambda(r-\hat{r}).$$

因此，如果 $(\alpha+\beta)\|E\|_{op}^2 \le \lambda$:

$$(1-\alpha^{-1})\|M^* - \hat{M}(\hat{r})\|_F^2 \le (1+\beta^{-1})\|M^* - M(r)\|_F^2 + 2\lambda r,$$

对于所有 $r \in [n \wedge p \wedge r_X]$ 以及所有 $M(r) = AXB$，其秩为 $A \wedge \text{rank } B \le r$。再次替换 $\alpha = 3/2$ 和 $\beta = 1/2$ 即可得到结果。然后利用

$$\min_{\substack{A,B \\ \text{rank } A \wedge \text{rank } B \le r}} \|A^*XB^* - AXB\|_F^2 = \sum_{k=r+1}^{r^*} \sigma_k(A^*XB^*)^2$$

以及式 (14) 中的高概率界。

5.4 第 2.3 节结果的证明

ˆ ˆ命题 2.4 的证明。对于任意 $r \in [n \wedge p \wedge r_X]$，有 $A_r X B_r = [Y]_r$ 是 Y 在秩小于或等于 r 的矩阵空间上的投影。现在，记

$$\begin{aligned} F(r) :&= \|Y - \hat{A}_r X \hat{B}_r\|_F^2 + \lambda r \\ &= \sum_{k=r+1}^{r_Y} \sigma_k(Y)^2 \cdot \mathbf{1}_{r<r_Y} + \lambda r \\ &= \sum_{k=r+1}^{r_Y} (\sigma_k(Y)^2 - \lambda) \cdot \mathbf{1}_{r<r_Y} + \lambda r_Y. \end{aligned}$$

容易看出，将 F 视为 r 的函数时，当 $\lambda > \sigma_{r_Y}(Y)^2$,时，其在 $r_Y(\lambda)$ 处有唯一最小值；而当 $\lambda \le \sigma_{r_Y}(Y)^2$ 时，对于 $r = r_Y$、.....、$(n \wedge p \wedge r_X)$，它均达到最小值且为常数。 ■

命题 2.5 的证明。根据 ˆr 的定义，我们有 $k > \hat{r}$ 当且仅当 $\lambda > \sigma_k(Y)^2$，并且 $k < \hat{r}$ 当且仅当 $\lambda \le \sigma_{k+1}(Y)^2$。在我们的模型 $Y = A^*XB^* + E$ 中，Weyl不等式给出：对所有 k，$|\sigma_k(A^*XB^*) - \sigma_k(Y)| \le \sigma_1(E)$。关于 ˆ$r$ 的事件可以用 $\sigma_1(E) = \|E\|_{op}$ 表示如下。我们有

$$\begin{aligned} \{k > \hat{r}\} \quad &\text{implies} \quad \lambda > (\sigma_k(A^*XB^*) - \sigma_1(E))^2, \\ \{k < \hat{r}\} \quad &\text{implies} \quad \lambda \le (\sigma_{k+1}(A^*XB^*) + \sigma_1(E))^2. \end{aligned}$$

因此，$\{\hat{r} \ne k\}$ 意味着 $\sigma_1(E) > \sigma_k(A^*XB^*) - \sqrt{\lambda}$ 或 $\sigma_1(E) \ge \sqrt{\lambda} - \sigma_{k+1}(A^*XB^*)$。我们来考虑 $k = r^*(\lambda)$。那么，假设 $\sigma_{r^*(\lambda)}(A^*XB^*) > (1+c)\sqrt{\lambda}$ 可得 $\sigma_1(E) > c\sqrt{\lambda}$；而假设 $\sigma_{r^*(\lambda)+1}(A^*XB^*) < (1-c)\sqrt{\lambda}$ 也可得 $\sigma_1(E) > c\sqrt{\lambda}$。因此，

$$\mathbb{P}\left(\hat{r} \ne r^*(\lambda)\right) \le \mathbb{P}\left(\sigma_1(E) > c\sqrt{\lambda}\right).$$

利用不等式 (14) 即可完成证明。 ■

5.5 定理 2.7 的证明

将 $\widehat{\sigma}_r^2$ 替换后，优化问题 (9) 可写为：

$$\bar{r} \in \arg\min_{r\in[r_{max}]} \|Y - \hat{A}_r X \hat{B}_r\|_F^2 \left(1 + \frac{\lambda r}{np}\right).$$

我们用 $\bar{M} = A_{\bar{r}} X B_{\bar{r}}$、$M_r = A_r X B_r$ 和 $M^* = A^* X B^*$ 表示。采用这一记号可得，对于 $r \le r_{max}$，

$$\|Y - \bar{M}\|_F^2 \left(1 + \frac{\lambda \bar{r}}{np}\right) \le \|Y - \hat{M}_r\|_F^2 \left(1 + \frac{\lambda r}{np}\right).$$

展开平方并利用等式 $Y = M^* + E$，得到

$$\|M^* - \bar{M}\|_F^2 \le \|M^* - \hat{M}_r\|_F^2 + 2\langle E, \bar{M} - \hat{M}_r\rangle_F + \frac{\lambda r}{np}\|Y - \hat{M}_r\|_F^2 - \frac{\lambda \bar{r}}{np}\|Y - \bar{M}\|_F^2.$$

现在利用上界 $\langle E, M - M_r\rangle_F \le \|E\|_{op}\|M - M_r\|_*$ 以及 M 和 M_r 的定义，推导出

$$\|M^* - \bar{M}\|_F^2 \le \|M^* - \hat{M}_r\|_F^2 + 2\|E\|_{op}\|\bar{M} - \hat{M}_r\|_* + \frac{\lambda r}{np}\sum_{k>r}\sigma_k(Y)^2 - \frac{\lambda \bar{r}}{np}\sum_{k>\bar{r}}\sigma_k(Y)^2.$$

请注意，在 $k > r_Y$ 的情况下，我们使用 $\sigma_k(Y) = 0$。回想 $\|\bar{M} - \hat{M}_r\|_* \le \sqrt{r + \bar{r}}\, \|M - M_r\|_F$，进一步得到

$$\begin{aligned}\|M^* - \bar{M}\|_F^2 \le \|M^* - \hat{M}_r\|_F^2 &+ 2\|E\|_{op}\sqrt{r+\bar{r}}\left(\|M^* - \bar{M}\|_F + \|M^* - \hat{M}_r\|_F\right) \\ &+ \frac{\lambda r}{np}\sum_{k>r}\sigma_k(Y)^2 - \frac{\lambda \bar{r}}{np}\sum_{k>\bar{r}}\sigma_k(Y)^2.\end{aligned}$$

将不等式 $2ab \le \alpha a^2 + \alpha^{-1}b^2$ 分别用于 $a,\ b > 0$，第一次取 $\alpha > 1$，第二次取 $\beta > 0$，得到

$$\begin{aligned}(1 - \alpha^{-1})\|M^* - \bar{M}\|_F^2 \le (1 + \beta^{-1})\|M^* - \hat{M}_r\|_F^2 &+ (\alpha + \beta)\|E\|_{op}^2(r + \bar{r}) \\ &+ \frac{\lambda r}{np}\sum_{k>r}\sigma_k(Y)^2 - \frac{\lambda \bar{r}}{np}\sum_{k>\bar{r}}\sigma_k(Y)^2. \qquad (15)\end{aligned}$$

现在我们区分两种情况：$r \le \bar{r}$ 和 $r > \bar{r}$。对于第一种情况，即 $r \le \bar{r}$，我们给出如下上界：

$$\begin{aligned}\frac{\lambda r}{np}\sum_{k>r}\sigma_k(Y)^2 - \frac{\lambda \bar{r}}{np}\sum_{k>\bar{r}}\sigma_k(Y)^2 &= \frac{\lambda}{np}\left(r\sum_{k=r+1}^{\bar{r}}\sigma_k(Y)^2 + (r - \bar{r})\sum_{k>\bar{r}}\sigma_k(Y)^2\right) \\ &\le \frac{\lambda}{np} r(\bar{r} - r)\sigma_{r+1}(Y)^2 \\ &\le \frac{2\lambda r}{np}(\bar{r} - r)(\sigma_{r+1}(M^*)^2 + \|E\|_{op}^2) \\ &\le \frac{2\lambda r}{np} r_{max}\sigma_{r+1}(M^*)^2 + \frac{2\lambda r_{max}}{np}(\bar{r} - r)\|E\|_{op}^2,\end{aligned}$$

其中我们使用了 Weyl不等式 $\sigma_{r+1}(Y) \le \sigma_{r+1}(M^*) + \|E\|_{op}$ $\sigma_{r+1}(Y)^2 \le 2\|E\|_{op}^2 + 2\sigma_{r+1}(M^*)^2$ (15)，从而得到 (15)。将其代入 (15)，得到

$$\begin{aligned}(1-\alpha^{-1})\|M^* - \bar{M}\|_F^2 \le (1+\beta^{-1})\|M^* - \hat{M}_r\|_F^2 + \frac{2\lambda r_{max}}{np} r\sigma_{r+1}(M^*)^2 \\ + r\|E\|_{op}^2(\alpha+\beta-\frac{2\lambda r_{max}}{np}) \\ + \bar{r}\|E\|_{op}^2(\alpha+\beta+\frac{2\lambda r_{max}}{np}),\end{aligned}$$

对于所有属于 $[r_{max}]$ 的 $r \le \bar{r}$。因此，对于满足 $\frac{2\lambda\cdot(r_{max}\vee r_Y)}{np} = (1+\varepsilon)(\alpha+\beta)$ 的 λ，其中某个 $\varepsilon > 0$，我们得到

$$\begin{aligned}(1-\alpha^{-1})\|M^* - \bar{M}\|_F^2 \le \min_{r\in[\bar{r}]}\left\{(1+\beta^{-1})\|M^* - \hat{M}_r\|_F^2 + (1+\varepsilon)(\alpha+\beta)r\sigma_{r+1}(M^*)^2\right\} \\ + (2+\varepsilon)(\alpha+\beta)r_{max}\|E\|_{op}^2.\end{aligned}$$

现在我们关注第二种情形，即 $r > \bar{r}$。我们观察到，在这种情形下，

$$\begin{aligned}\frac{\lambda r}{np}\sum_{k>r}\sigma_k(Y)^2 - \frac{\lambda\bar{r}}{np}\sum_{k>\bar{r}}\sigma_k(Y)^2 &= \frac{\lambda}{np}\left((r-\bar{r})\sum_{k>r}\sigma_k(Y)^2 - \bar{r}\sum_{k=\bar{r}+1}^{r}\sigma_k(Y)^2\right) \\ &\le \frac{\lambda(r-\bar{r})}{np}(r_Y - r)\sigma_{r+1}(Y)^2 \\ &\le \frac{2\lambda r}{np} r_Y\cdot\sigma_{r+1}(M^*)^2 + \frac{2\lambda(r-\bar{r})}{np}\cdot(r_Y\vee r_{max})\|E\|_{op}^2,\end{aligned}$$

根据前一种情形中的类似推理。将其代入（15），得到

$$\begin{aligned}(1-\alpha^{-1})\|M^* - \bar{M}\|_F^2 \le (1+\beta^{-1})\|M^* - \hat{M}_r\|_F^2 + \frac{2\lambda\cdot r_{max}\vee r_Y}{np} r\sigma_{r+1}(M^*)^2 \\ + r\|E\|_{op}^2(\alpha+\beta+\frac{2\lambda\cdot r_{max}\vee r_Y}{np}) \\ + \bar{r}\|E\|_{op}^2(\alpha+\beta-\frac{2\lambda\cdot r_{max}\vee r_Y}{np}).\end{aligned}$$

采用同样的 λ 取值，使得对于某个 $\varepsilon > 0$，有 $\frac{2\lambda\cdot r_{max}\vee r_Y}{np} = (1+\varepsilon)(\alpha+\beta)$，在这种情形下我们也得到

$$\begin{aligned}(1-\alpha^{-1})\|M^* - \bar{M}\|_F^2 \le \min_{\bar{r}<r\le r_{max}}\left\{(1+\beta^{-1})\|M^* - \hat{M}_r\|_F^2 + (1+\varepsilon)(\alpha+\beta)r\sigma_{r+1}(M^*)^2\right\} \\ + (2+\varepsilon)(\alpha+\beta)r_{max}\|E\|_{op}^2.\end{aligned}$$

取 $\alpha = 3/2$ 和 $\beta = 1/2$，并结合这两种情形，可得以下结果

$$\|M^* - \bar{M}\|_F^2 \le \min_{r\in[r_{max}]}\left\{9\|M^* - \hat{M}_r\|_F^2 + 6(1+\epsilon)\cdot r\sigma_{r+1}(M^*)^2\right\} + 6(2+\epsilon)\cdot r_{max}\|E\|_{op}^2,$$

其中，我们选择 λ，使得 $\lambda\cdot r_{max}\vee r_Y = (1+\varepsilon)np$ 对某个 $\varepsilon > 0$ 成立。最后，我们使用不等式 (14) 得出结论。

5.6 定理 3.1 的证明

我们分两步求解该问题，以求解可等价地写为如下形式的优化问题 (10)

$$\min_{\substack{A,B\\M=AXB}} \min_{M} \|Y-M\|_F^2 + 2\lambda\cdot\|M\|_*,$$

对于 $\lambda>0$。 M 中问题的解有显式形式，且已知可通过对 Y 的特征值进行软阈值处理得到：$\bar{M}=U_Y Diag_{n,\ p}((\sigma_k(Y)\ -\lambda)+)V_Y^\top$，其中我们使用了 Y 的奇异值分解：$U_Y\Sigma_Y V_Y^\top$。接下来，我们将 $\bar{M}$ 投影到矩阵空间 AXB 上，使其在弗罗贝尼乌斯范数下满足 A 和 $B^{-}\ ^{-}$。容易验证，我们对 $A,\ B$ 的选择是精确解，即 $_{-}\ \ _{-}\ \ _M=AXB$。

$^{-}$与定理 2.3 的证明类似，根据 M 的定义，展开平方项并重新排列各项，可得对所有 M：

$$\begin{aligned}\|\bar{M}-M^*\|_F^2 &\le \|M^*-M\|_F^2 + 2\langle E,\bar{M}-M\rangle + 2\lambda(\|M\|_*-\|\bar{M}\|_*)\\ &\le \|M^*-M\|_F^2 + 2\sqrt{\lambda}(\|\bar{M}-M\|_*+\|M\|_*-\|\bar{M}\|_*),\end{aligned}$$

在事件 $\|E\|_{op}^2\le\lambda$ 成立的条件下。我们利用矩阵 $^{-}\ \ _{-}\ \ _{-}\ \ _{-}\ \ _{-}\ \ _{-}\ \ _{-}\ \ ^{-}$ 的核范数的可分解性，如 [5], 中所示，找到 M_1 和 M_2，使得 $M=M_1+M_2$、$\|M\|_*=\|M_1\|_*+\|M_2\|_*$ 且 $_{-}\ \ _{-}\ \ _{-}\ \ _{-}\|M-M\|_*=\|M_1-M\|_*+\|M_2\|_{*\circ}$ 此外，$\mathrm{rank}(M_1)\le 2\mathrm{rank}(M)$。这意味着

$$\begin{aligned}\|\bar{M}-M^*\|_F^2 &\le \|M^*-M\|_F^2 + 4\sqrt{\lambda}\|\bar{M}_1-M\|_*\\ &\le \|M^*-M\|_F^2 + 4\sqrt{\lambda}\sqrt{3\,\mathrm{rank}(M)}\cdot\|\bar{M}_1-M\|_F\\ &\le \|M^*-M\|_F^2 + 4\sqrt{\lambda}\sqrt{3\,\mathrm{rank}(M)}\cdot(\|\bar{M}-M^*\|_F+\|M-M^*\|_F).\end{aligned}$$

对于任意实数 $\alpha>1$ 和 $\beta>0$，对所有 M，我们得到

$$(1-\alpha^{-1})\|\bar{M}-M^*\|_F^2 \le (1+\beta^{-1})\|M^*-M\|_F^2 + 4(\alpha+\beta)\lambda\cdot 6\,\mathrm{rank}(M).$$

对于特定值 $\alpha=3/2$ 和 $\beta=1/2$，我们得到

$$\begin{aligned}\|\bar{M}-M^*\|_F^2 &\le \min_M\left\{9\|M^*-M\|_F^2 + 144\lambda\cdot\mathrm{rank}(M)\right\}\\ &\le 9\min_{r\in[n\wedge p\wedge r_X]}\left\{\min_{M:\mathrm{rank}\,M=r}\|M^*-M\|_F^2 + 16\lambda\cdot r\right\}.\end{aligned}$$

回想一下 $\min_{M:\mathrm{rank}\,M=r}\|M^*-M\|_F^2=\sum_{K=r+1}^{r^*}\sigma_K(M^*)^2\cdot\mathbf{1}_{r<r^*}$ 以获得最终结果 .

6 辅助结果

算法1 用于选择 $\bar{r}$ 和 λ的数据驱动程序 输入：数据 X, Y

要求：A_r**,**B_r**,**$(p \wedge q)$**,**$r_X > 0$**定义：** $\widehat{\sigma}_r^2 := \overline{\|Y - \hat{}$

$np - \ m \wedge q \ r_X \hat{}$ **( ) 定义：** $\lambda(\sigma) := 4(n+p)\sigma^2$**定义：** $\hat{}r_\lambda$ **：**

$= \textbf{arg min}_{r \in [n \wedge p \wedge r_X]} \left(\|Y - \hat{}A_r X \hat{}B_r\|_F^2 + \lambda \cdot r \right)$**初始化：**

$r \leftarrow r_X \wedge n \wedge p$**,** $\bar{r} \leftarrow \hat{}r_{\lambda(\widehat{\sigma}_r^2)}$ **当** $\bar{r} < r$ **时执行**$\_r \leftarrow r\bar{r} \leftarrow \hat{}$

$r_{\lambda(\widehat{\sigma}_r^2)}$ **结束循环 输出：** $\bar{r}$**,** $\lambda(\widehat{\sigma}_{\bar{r}}^2)$

致谢。作者感谢法国国家研究署（ANR）提供的 Labex Ecodec 项目资助（ANR-11-LABEX-0047）。

参考文献

[1] Francis R Bach. 迹范数最小化的一致性. 机器学习研究杂志, 9:1019–1048, 2008年.[2] Xin Bing, Florentina Bunea, and Marten Wegkamp. 稀疏主题模型的最优估计. 机器学习研究杂志, 21(1):7189–7233, 2020.[3] Xin Bing and Marten H. Wegkamp. 高维多元响应回归模型中系数矩阵秩的自适应估计. 《统计年刊》, 47(6):3157–3184, 2019年.[4] Lucien Birg´e and Pascal Massart. 高斯模型选择的最小惩罚. 概率论及相关领域, 138:33–73, 2007.[5] Florentina Bunea, Yiyuan She, and Marten H. Wegkamp. 高维矩阵降秩估计量的最优选择. 《统计年刊》, 39(2):1282–1309, 2011年.[6] Rong Chen, Han Xiao, and Dan Yang. 矩阵值时间序列的自回归模型. 《计量经济学杂志》, 222(1):539–560, 2021年。

[7] David Donoho and Victoria Stodden. 非负矩阵分解何时能够正确地分解为各个部分？神经信息处理系统进展, 16, 2003年.[8] Jianqing Fan, Yuan Liao, and Martina Mincheva. 近似因子模型中的高维协方差矩阵估计. 统计年鉴, 39(6):3320, 2011年.[9] Christophe Giraud. 低秩多元回归. 电子统计期刊, 5:775–799, 2011年.[10] Nan-Jung Hsu, Hsin-Cheng Huang, and Ruey S. Tsay. 矩阵自回归时空模型. 计算与图形统计学杂志, 30(4):1143–1155, 2021 年 .[11] Zheng Tracy Ke and Minzhe Wang. 使用奇异值分解进行主题建模. 美国统计学会会刊, pages 1–16, 2022年.[12] Olga Klopp, Yu Lu, Alexandre B. Tsybakov, and Harrison H. Zhou. 结构化矩阵估计与补全. *Bernoulli*, 25(4B):3883–3911, 2019年.[13] Olga Klopp, Maxim Panov, Suzanne Sigalla, and Alexandre Tsybakov. 通过连续投影为文档分配主题. *arXiv*预印本 *arXiv:2107.03684*, 2021年.[14]Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. 含噪低秩矩阵补全的核范数惩罚与最优速率. 《统计年刊》, 39(5):2302–2329, 2011年.[15] Chen Kun, Dong Hongbo, and Chan Kung-Sik. 通过自适应核范数惩罚实现降秩回归. 生物计量学, 100:901–920, 2013年.[16] Sahand Negahban and Martin J Wainwright. 含噪声和高维标度下的（近）低秩矩阵估计. 第*27*届国际机器学习会议论文集, 2011年.[17] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. 基于可分解正则化项的 M-估计量高维分析统一框架. 统计科学, 27(4):538–557, 2012年.[18] Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. 高维多元回归中的支撑集并集恢复. 《统计年刊》, 39(1):1–47, 2011年.[19] Angelika Rohde and Alexandre B Tsybakov. 高维低秩矩阵估计. 《统计年刊》, 39(2):887–930, 2011年.[20] Roman Vershynin. 高维概率, 剑桥统计与概率数学丛书第47卷。剑桥大学出版社，剑桥，2018年。数据科学应用导论，Sara van de Geer 作序。