Two-sided Matrix Regression

Nayel Bettache, Cristina Butucea

Two-sided Matrix Regression

Nayel Bettache[1] and Cristina Butucea[1]

[1]CREST, ENSAE, Institut Polytechnique de Paris, 5 avenue Henry Le Chatelier, 91120 Palaiseau, France

March 9, 2023

Abstract

The two-sided matrix regression model $Y = A^*XB^* + E$ aims at predicting Y by taking into account both linear links between column features of X, via the unknown matrix B^*, and also among the row features of X, via the matrix A^*. We propose low-rank predictors in this high-dimensional matrix regression model via rank-penalized and nuclear norm-penalized least squares. Both criteria are non jointly convex; however, we propose explicit predictors based on SVD and show optimal prediction bounds. We give sufficient conditions for consistent rank selector. We also propose a fully data-driven rank-adaptive procedure. Simulation results confirm the good prediction and the rank-consistency results under data-driven explicit choices of the tuning parameters and the scaling parameter of the noise.

Key Words: Matrix regression, Multivariate response regression, Nuclear norm penalized, Oracle inequality, Rank penalized, Rank selection, Two-sided matrix regression.

1 Introduction

Supervised learning is often performed on large data bases. Matrix regression assumes that the data Y can be well explained by a set of features given by the columns of the matrix X and linear combinations of these columns. It is often the case in real-life that the rows of Y can be explained by linear combinations of the rows of X.

For example, economic data store economic indicators as column features and countries as rows. Such a matrix is usually explained by a smaller matrix roughly containing a smaller number of countries (representatives of groups of geographically or economically close countries) and a few economic features or some factors produced out of all these indicators. We would like to predict a larger number of indicators for a larger number of countries, *i.e.* Y a $n \times p$ matrix, using the features X a $m \times q$ matrix.

Recommendation systems want to predict the opinion of n clients concerning p items. We can use publicly available data on a number m of different groups of clients and their affinity to a number q of large categories of items in order to predict by evaluating the client's

correlation to the prescribed groups in the population and the item's weight in its category. We may include a multiple-label situation where the items belonging to a main category are also related to other categories.
Other examples can be given for meteorological data, medical or pharmaceutical data and so on.

Model. We observe the matrix $Y \in \mathbb{R}^{n\times p}$ and a design matrix $X \in \mathbb{R}^{m\times q}$ related via the **two-sided matrix regression (2MR)** model involving two parameter matrices $A^* \in \mathbb{R}^{n\times m}$ and $B^* \in \mathbb{R}^{q\times p}$:

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

where the noise matrix E is assumed to have independent centered $\sigma-$sub-Gaussian entries.

The 2MR model encompasses known models like, *e.g.* matrix regression and matrix factorisation. Indeed, if $n = m$ and A^* is the identity, the matrix model (1) becomes the (one-sided) *matrix regression* (MR) model $Y = XB^* + E$, see [18], [5], [17].
Assume now that $m = q$ and that the design matrix X is the identity matrix of rank m smaller than both n and p. Our model becomes a *factorisation model* of the signal $M^* = A^*B^*$ observed with noise. The idea is to recover a low-rank structure generating the observed data. In [12] the authors have considered structured factorisation of the signal under assumptions that the rows of A^* and the columns of B^* have a common sparsity parameter and X, which they do not observe, has a much smaller dimension than Y.

The 2MR model (1) is strongly related to other models, but we argue that it cannot be reduced to these other models of a different nature. Indeed, note that the entry Y_{ij} of the matrix Y can be written

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

for any i in $[n]$, where $[n] = \{1, \dots, n\}$, and for any j in $[p]$. Thus every entry Y_{ij} brings information through the same design matrix X on the rank 1 matrix $B^*_{\cdot,j} A^*_{i,\cdot}$. This is unlike *the trace-regression model* or the more general *matrix completion* studied by [19], [14], where a different design matrix brings information on the parameter matrix B^*A^*.
Another way of writing model (1) is in the form of *vector regression model*, by stacking the columns of matrices Y, X and E into $vec(Y)$, $vec(X)$ and $vec(E)$, respectively, to get

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

where $\otimes$ denotes the tensor product of two matrices. Under this relation, we predict a row vector of size np using a row vector of size mq (the matrix of features has rank 1) via a parameter of size $(mq) \times (np)$ which cannot go well unless the structure of A and B is trivial. This approach cannot take into account the matrix structure of the features, of the matrices A^*, B^*, and it gives poor results on that account.

This model has been introduced in time series by [6] as the *auto-regressive matrix-valued model* of order 1, MAR(1), $Y_t = A^* Y_{t-1} B^* + E_t$, observed at times t in $[T]$. In this case A^* and B^* are squared matrices with spectral radii strictly less than 1 in order to ensure stability of the time series (X_t is thus stationary and causal). The authors propose three estimation methods: first, they use the vector form analogous to (2), stack the T lines of $vec(Y_t)^\top$ and

they use the nearest Kronecker product (NKP) problem to give estimators of A^* and B^* out of the global least squares estimator of ${A^*}^\top \otimes B^*$; then, their next method minimizes the least squares over A and B

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

by a sequential procedure minimizing over A for fixed given B, then over B for fixed A, and iterating; finally, they give an MLE procedure over A and B under a particular structure of the covariance matrix of E and proceed also sequentially. Theoretical results state the asymptotic normality as T tends to infinity, for fixed dimensions. However, the first procedure is cumbersome as the estimated matrix is very large, while the other two procedures are based on non-convex minimization without theoretical guarantees as to the limit points of the algorithm.

Least squares and MLE estimators with AIC and BIC penalties have been numerically studied by [10] of a more general time series model

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

which is treated as $Y_t = A^* X_t B^* + E_t$, where X_t is the block diagonal matrix containing the $L-$past observed matrices $Y_{t-1}, \dots, Y_{t-L}$ and $A^* = (A_1, \dots, A_L)$ and $B^* = (B_1^\top, \dots, B_L^\top)^\top$ are the concatenated matrices in the previous equation.

Thus, our paper is motivated by the need to deal with high-dimensional data and finite (non-asymptotic) time (say $T = 1$) in order to provide theoretical guarantees for prediction.

Contributions. We show in Section 2 that by using the SVD of matrices $Y = U_Y \Sigma_Y V_Y^T$ and $X = U_X \Sigma_X V_X^T$, the least squares procedure can be reduced to fitting predictors of the form $A_0 \Sigma_X B_0$ to the diagonal matrix Σ_Y with explicit relations between A_0, B_0 and A, B. There is a natural choice of predictors of A_0 and of B_0 under diagonal form. We study these predictors for given ranks r and that we transform back into the original space of Y without loss of prediction rate. Then we give a data-dependent rank selector and show that the predictors associated to it attain optimal bounds. We give sufficient conditions so that the rank selector is consistent. Finally, we slightly modify the procedure to be free of the parameter σ of the noise and show new upper bounds in this case. In Section 3, we study the nuclear norm penalized least squares and show it attains the optimal bounds too. All proofs are in a dedicated section in the Appendix. Finally, we illustrate in Section 4 via numerical simulations the excellent prediction results of these fast running, explicit predictors.

Notations. For any integers n and m we denote $n \wedge m$ for the minimum between n and m and $n \vee m$ for the maximum between n and m. For any matrix M of size $n \times m$ and rank r_M, we denote its singular value decomposition (SVD) by $M = U_M \Sigma_M V_M^\top$, where U_M belongs to $\mathcal{O}_n$ - the set of orthogonal matrices of size $n \times n$, V_M belongs to $\mathcal{O}_m$ and

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

Note that $\sigma_1(M), \dots, \sigma_{r_M}(M)$ are the positive singular values of M listed in decreasing order, and the $n \times m$ diagonal matrix $Diag_{n,m}(\sigma_k(M),\, 1 \le k \le r_M)$ has diagonal entries in the list

and 0 elsewhere. Furthermore, denote $\|M\|_F^2 = \sum_{k=1}^{n\wedge m} \sigma_k(M)^2$ its Frobenius norm, $\|M\|_{(2,q)}^2 = \sum_{k=1}^{q} \sigma_k(M)^2$ its Ky-Fan $(2,q)$ norm, $\|M\|_{op} = \sigma_1(M)$ its operator norm, $\|M\|_* = \sum_{k=1}^{n\wedge m} \sigma_k(M)$ its nuclear norm, $M^\dagger$ its Moore-Penrose inverse, r_M its rank and M^T its transpose. For any matrices M_1 and M_2 in $\mathbb{R}^{n\times m}$, $\langle M_1, M_2\rangle_F$ denotes the canonical scalar product, *i.e.* $\langle M_1, M_2\rangle_F = \mathrm{Tr}(M_1^T M_2)$. For any $r \in [r_M]$, we denote $[M]_r$ the best rank r approximation of M for the Frobenius norm. In the model (1), let us denote by r^* the rank of A^*XB^*.

2 Rank penalized learning

In this section we propose rank adaptive predictors and provide theoretical guarantees for their error. First we give explicit predictors under the assumption that the ranks of the parameter matrices are known, then a selection procedure will allow to provide a data-dependent rank selector and the associated rank-adaptive predictor. Even though we follow classical results for rank penalized (one-sided) matrix regression, e.g. [5], [9] and [3], we give details for the fixed rank two-sided matrix regression which is novel to the best of our knowledge. Surprisingly, explicit predictors can be proposed despite the identifiability issues of this model. Only after this, we proceed to rank selection and rank-adaptive learning.

2.1 Prediction for given ranks

Let r belong to $[n \wedge p \wedge r_X]$. Let us build explicit predictors $(\hat{A}_r, \hat{B}_r)$ solutions to the non-convex minimization problem

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

Notice that the rank constraints on A and B use the same value r. Indeed the objective is to build a predictor for the signal A^*XB^* which satisfies $\mathrm{rank}(A^*XB^*) \le \min(r_{A^*}, r_X, r_{B^*})$. In the steps of the proof of our results, we see that the upper bound of the risk depends on the ranks of A^* and of B^* only through their least value and no information can be recovered on the largest rank of the two. Hence it makes sense to look for A and B sharing the same rank as a dimension reduction technique without any impact on the final results.

The model (1) can be rewritten using the SVD of the observed matrix Y and of the design matrix X as

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

where $A_0^* = U_Y^T A^* U_X$, $B_0^* = V_X^T B^* V_Y$ and $E_0 := U_Y^T \cdot E \cdot V_Y$. In the particular case where E has independent entries with distribution $\mathcal{N}(0,\sigma^2)$ than so does E_0, see Lemma 5.1. Now, Σ_Y and Σ_X are diagonal matrices, not necessarily squared, not necessarily full rank. Given the invariance of the Frobenius norm by left or right multiplication with orthogonal matrices, we get that for any matrices $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{q\times p}$ we have

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

where $A_0 = U_Y^T A U_X$ and $B_0 = V_X^\top B V_Y$ are obtained via analogous transformations to those relating the true underlying parameters.

Obviously, matrices A and A_0 have the same rank, and the same holds for B and B_0. Therefore, solving (3) is equivalent to solving for $\hat{A}_{0r}$ and $\hat{B}_{0r}$ solutions of

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

Theorem 2.1 *Let us define for* $r \in [n \wedge p \wedge r_X]$

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

Then, $(\hat{A}_{0r}, \hat{B}_{0r})$ *belong to the set of solutions of problem* (5) *and the predictor* $\hat{A}_{0r}\Sigma_X \hat{B}_{0r}$ *satisfies for an absolute constant* $C > 0$ *and for any* $t > 0$*, the oracle inequality*

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

with probability larger than $1 - 2\exp(-t^2(\sqrt{n} + \sqrt{p})^2)$.

Next, from the explicit solutions $(\hat{A}_{0r}, \hat{B}_{0r})$ of (5) we deduce explicit solutions of (3).

Corollary 2.2 *Let us define for* $r \in [n \wedge p \wedge r_X]$

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

with $\hat{A}_{0r}$ *and* $\hat{B}_{0r}$ *defined in* (6)*. Then* $(\hat{A}_r, \hat{B}_r)$ *are solution to the problem* (3) *and the predictor* $\hat{A}_r X \hat{B}_r$ *satisfies for an absolute constant* $C > 0$ *and for any* $t > 0$*, the oracle inequality*

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

with probability larger than $1 - 2\exp(-t^2(\sqrt{n} + \sqrt{p})^2)$.

The proofs of Theorem 2.1 and of Corollary 2.2 can be found in Section 5. In the proofs we explicit the bias in terms of the unknown matrix parameters:

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

Note that our choice for the couple of predictors $(\hat{A}_{0r}, \hat{B}_{0r})$ is not unique and we can easily derive families of solutions to the problem (5). Each family of solutions can be turned

into a solution to the problem (3). Indeed, consider $(\alpha\hat{A}_{0r}, \frac{1}{\alpha}\hat{B}_{0r})$ with arbitrary $\alpha > 0$. Alternatively, let λ_i for all $i \leq m \wedge q$ be arbitrary positive numbers, then

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

give the same prediction. Let us see that the same transformations applied to the parameter matrices A_0^* and B_0^* also lead to the same signal matrix $A_0^*\Sigma_X B_0^*$. Indeed, the model is non-identifiable and so, without further strong assumptions, we can only hope to learn the global signal, and not the parameters of the model.

Alternative predictors. Let us define a second couple of predictors $(\tilde{A}, \tilde{B}_r)$ producing exactly the same prediction as $(\hat{A}_r, \hat{B}_r)$ with the same theoretical properties, but having the advantage that $\tilde{A}$ is full rank and does not depend on r. Define

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

where $I_{n,m}$ denotes the identity matrix of dimension $n \times m$, whereas $\tilde{B}_{0r}$ has rank $r \wedge r_Y$. Using the analogous transformations we obtain

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

It is easy to see that Theorem 2.1 is valid for $\tilde{A}_0$ and $\tilde{B}_{0r}$, and that Corollary 2.2 is valid for $\tilde{A}$ and $\tilde{B}_r$.

2.2 Rank-adaptive prediction

In this section, we propose rank-adaptive predictors $(\hat{A}_{\hat{r}}, \hat{B}_{\hat{r}})$ which are selected from the family $\{(\hat{A}_r, \hat{B}_r) : r \in [n \wedge p \wedge r_X]\}$ by a model selection procedure analogous to that of [5]. Let us first define, for a generic matrix M and any $\lambda > 0$, the $\lambda-$rank of M as

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

For given $\lambda > 0$, let

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

Consider the predictors introduced in (7) for the data-driven rank $\hat{r}$ as defined in (8). The next Theorem extends the oracle inequality to the rank-adaptive predictors $(\hat{A}_{\hat{r}}, \hat{B}_{\hat{r}})$ associated to the estimated rank $\hat{r}$ and to some $\lambda > 0$ large enough.

Theorem 2.3 *The rank-adaptive predictors $(\hat{A}_{\hat{r}}, \hat{B}_{\hat{r}})$ associated to $\hat{r}$ in (8) and to λ such that, for some absolute constant $C > 0$ and for any $t > 0$, $\lambda \geq 4C(1+t)^2\sigma^2(n+p)$, satisfy the oracle inequality*

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

with probability larger than $1 - 2\exp(-t^2(\sqrt{n} + \sqrt{p})^2)$.

Note that the minimum on the right-hand side of the previous display is always smaller than the value at $r = r^*$, giving under the assumptions of Theorem 2.3 that

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

with probability larger than $1 - 2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$.
The bounds of order $r^*(n+p)$ attained by our procedure are analogous to those for the low-rank matrix regression models in [19] and [9]. Indeed, the 2MR model is more difficult than the MR model, (*i.e.* one of the matrices is known) and we will suppose known the matrix with larger rank in order to achieve the correct lower bounds. Thus the lower bounds for prediction in the low-rank MR model will be valid for our model.

2.3 Consistent rank selection

We study the consistency of the rank selector $\hat{r}$ in (8) and see when it recovers the true rank r^* with high probability. First, we show that, for properly chosen λ, the data-driven rank $\hat{r}$ is actually the unique solution and coincides with the $\lambda-$rank of Y, $\hat{r} = r_Y(\lambda)$.

Proposition 2.4 *If $\lambda > \sigma_{r_Y}(Y)^2$, there is a unique solution $\hat{r}$ to the optimisation problem in* (8) *and it is actually the $\lambda-$rank of Y, i.e. $\hat{r} = r_Y(\lambda)$.*

Next, we prove that $\hat{r}$ recovers with high probability the $\lambda-$rank of A^*XB^*.

Proposition 2.5 *Let $\lambda > 0$ and denote by $r^*(\lambda)$ the $\lambda-$rank of A^*XB^* . If for some constant c in (0,1), $\sigma_{r^*(\lambda)}(A^*XB^*)^2 > (1+c)^2\lambda$ and $\sigma_{r^*(\lambda)+1}(A^*XB^*)^2 < (1-c)^2\lambda$, then*

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

In particular, if $\lambda \geq 2C(n+p)\sigma^2(1+t)^2/c^2$ for some absolute constant $C > 0$ and for any $t > 0$, then $\hat{r} = r^*(\lambda)$ with probability larger than $1 - 2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$.

Finally, remember that the fact that $r^*(\lambda)$ coincides with the true underlying rank r^* is equivalent to having $\sigma_{r^*}(A^*XB^*)^2 \geq \lambda > 0$. The rank selector will then coincide with r^* if λ also satisfies $\sigma_1(E)^2 \leq c^2\lambda$, for some absolute constant $c > 0$. It is therefore necessary that a signal-to-noise ratio, given here by $\sigma_{r^*}(A^*XB^*)^2/\sigma_1(E)^2$ be significant in order to have the true underlying rank selected by $\hat{r}$. By combining this with the previous Propositions we get the following.

Proposition 2.6 *Let $\lambda > 0$. If for some constant c in (0,1), $\sigma_{r^*}(A^*XB^*)^2 > (1+c)^2\lambda$, then*

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

In particular, if $\lambda \geq 2C(n+p)\sigma^2(1+t)^2/c^2$ for some absolute constant $C > 0$ and for any $t > 0$, then $\hat{r} = r^*$ with probability larger than $1 - 2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$.

2.4 Data-driven rank-adaptive prediction

The rank selector $\hat{r}$ in (8) is used for building consistent predictors as detailed in Theorem 2.3 provided that the condition $\lambda \geq 4C(1+t)^2\sigma^2(n+p)$ is satisfied. However the noise parameter σ is not known in general settings. Thus a data dependent rank selector is needed for building consistent predictors in those cases. Motivated by the previous case where σ^2 was supposed known, we proceed as follows. First, we change the penalty to $\lambda \cdot r\widehat{\sigma}_r^2$ with

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

Note that in the particular case of Gaussian noise $\widehat{\sigma}_r^2$ estimates the variance σ^2 of the noise. Next, given a largest possible value for the true rank $r_{max} \leq n \wedge p \wedge r_X$, we define the data-driven rank selector

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

Finally, we use the predictors $(\hat{A}_{\bar{r}}, \hat{B}_{\bar{r}})$. The next theorem extends the upper bounds of Theorem 2.3 to these data-driven rank-adaptive predictors.

Theorem 2.7 *The data-driven rank-adaptive predictors $(\hat{A}_{\bar{r}}, \hat{B}_{\bar{r}})$ associated to $\bar{r}$ in (9) with $r_{max} \leq n \wedge p \wedge r_X$, and to $\lambda = (1+\varepsilon)np/(r_{max} \vee r_Y)$ for some $\varepsilon > 0$, satisfy for some absolute constant $C > 0$ and for any $t > 0$ the oracle inequality*

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

with probability larger than $1 - 2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$.

Apply the Corollary 2.2, to get under the assumptions of Theorem 2.7 that

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

with probability larger than $1 - 2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$.

Note that the minimum on the right-hand side of the previous display is always smaller than its value at $r = r^*$ if r_{max} is larger than r^*, giving under the assumptions of Theorem 2.7 that

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

In order to compare to the previous results, note that the upper bound derived from Theorem 2.3 for the value $r = r^*$ and the least value $\lambda = 4C(1+t)^2\sigma^2(n+p)$ gives the very similar bound

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

From a computational point of view, it is preferable to change $\widehat{\sigma}_r^2$ in some cases. For example, we use in our numerical simulations

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

when $n \geq m$, $p \geq q$ and thus $np > (m \wedge q) r_X$. It is straightforward to prove the analogue of Theorem 2.7 by considering $\lambda = (1+\varepsilon)(np - (m \wedge q) r_X)/(r_{max} \vee r_Y)$.

3 Nuclear norm penalized learning

Nuclear norm penalized least squares is known to exhibit good properties, see [1] or [16]. Hence it may show advantages over rank-penalized methods. Let us define the nuclear norm penalized (NNP) optimisation problem

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

for some $\lambda > 0$. The objective of the optimization problem is non-jointly convex in A and B. Note that in matrix regression (when A^* is the identity matrix) the nuclear norm of XB has been used , see [14], or other adaptive forms depending on the feature matrix X, [15]. However, we exhibit explicit predictors belonging to the set of solutions of this problem and show an oracle inequality they satisfy.

Theorem 3.1 *The predictors $(\bar{A}, \bar{B})$ defined by*

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

are solutions to the problem in (10)*. Moreover, if λ is such that, for some absolute constant $C > 0$ and for any $t > 0$, $\lambda \geq 2C(1+t)^2\sigma^2(n+p)$, they satisfy the oracle inequality*

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

with probability larger than $1 - 2\exp(-t^2(\sqrt{n} + \sqrt{p})^2)$.

The proof can be found in Section 5.

Remark. Another approach could be to consider the model under the vectorized form (2) and solve the problem

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

for some $\lambda > 0$. Recall that $A^\top \otimes B$ denotes the tensor product of matrices $A^\top$ and B and that we can write $\|A^\top \otimes B\|_* = \sum_{k,j \geq 1} \sigma_k(A)\sigma_j(B)$. However, the features are 1-dimensional

and we loose the structured information contained in the original matrix X. This approach could make more sense in the case of repeated observation (Y_t, X_t) for t in $[T]$, by stacking the rows $vec(Y_t)^\top$ and $vec(X_t^\top)$ into matrices $\mathbb{Y}$ and $\mathbb{X}$, respectively, and do a classical matrix regression. Even so, the usual assumptions on the feature matrix $\mathbb{X}$ in order to achieve good prediction are not reasonable in this context as they are not much related to the original matrix data sets X_t, t in $[T]$.

Remark (Sufficient conditions for identifiability) We have indicated at several times that many couples of matrices (A, B) solve the equation $M = AXB$ for a given matrix M. Given the SVD of the matrix M, we may reduce the dimensionality of the problem by choosing the solution (A, B) given by $A = U_M A_0 U_X^\top$ and $B = V_X B_0 V_M^\top$, with A_0 and B_0 diagonal matrices such that

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

Thus, even under diagonal forms we can only identify the product of respective singular values of A and B. We can only hope to identify matrices A and B under very restrictive conditions where $X^\top X$ has full rank and either the matrix A or the matrix B is assumed to have known singular values, *e.g.* like a projector with singular values 1 or 0. Few other setups are known to be identifiable in the literature of factorisation of matrices, *e.g.* non-negative matrix factorisation (NMF), see [7], NMF for topic models [11], [2], [13] or covariance matrix factorization [8].

4 Numerical Results

Let us set the dimensions of the observed matrix Y to be $n = 100$ and $p = 300$, the dimensions of the design matrix X to be $m = 50$ and $q = 60$. We randomly generate three matrices: A^*, B^*, and X, with independent random gaussian entries with mean 0 and variance 1. These matrices are then projected onto the best low-rank matrix approximation, with the matrix A^* having a rank $r_A^* = 16$, the matrix B^* having a rank $r_B^* = 12$, and the matrix X having a rank $r_X = 25$. The signal matrix is defined as A^*XB^* and shows a rank of 12 in all experiments. We also define various settings for the variance σ^2 of the Gaussian noise E so that the signal-to-noise ratio $SNR := \sigma_{r^*}(A^*XB^*)^2/\sigma_1(E)^2$ varies approximately in the range $[0.5, 2]$.

Figure 1 illustrates the prediction performances of the predictor $\hat{A}_r X \hat{B}_r$, defined in (7), for different values of r. For $\sigma < 8$ giving the SNR approximately above the value 1, the prediction risk decreases when the rank increases while remaining bounded from above by 12 and then increases with the rank when the rank is above 12. For $\sigma \geq 8$ giving the SNR below the value 1, the prediction risk decreases when the rank increases while remaining bounded from above by 11 and then increases with the rank when the rank is above 11. It highlights that the best predictor is achieved when $r = r^* = 12$ for small noise variance levels (*i.e.* $\sigma < 8$) and when $r = 11$ for strong noise variance levels (*i.e.* $\sigma \geq 8$). This shows that there is a strong overfitting phenomenon in the case of strong noise and that it is therefore better to slightly underestimate the rank in these situations.

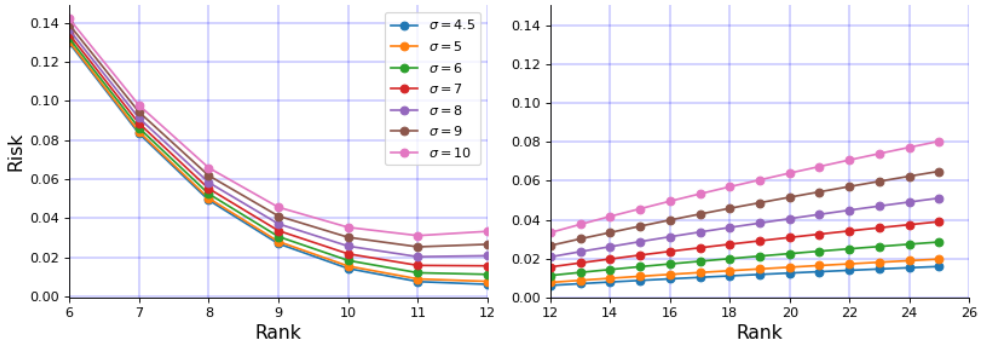

Figure 1: Evolution of the risk $\frac{\|\hat{A}_r X \hat{B}_r - A^* X B^*\|_F^2}{\|A^* X B^*\|_F^2}$ in function of r for different values of σ

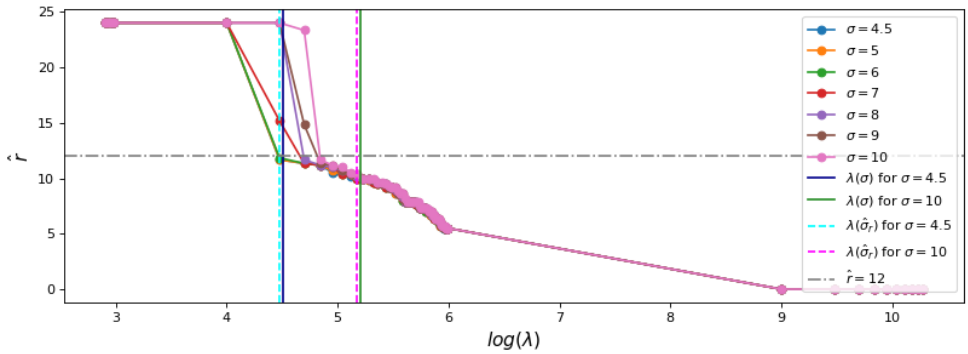

Figure 2: Evolution of the estimated $\hat{r}$ as a function of $\log(\lambda)$ for different values of σ

Figure 2 represents the predicted $\hat{r}$, defined in (8), for various values of λ. Independently of the noise variance level, for small values of λ the estimated $\hat{r}$ is maximal and there is $\hat{r} = r_X = 25$. This illustrates the previously exposed overfitting phenomenon, that is the higher the rank r, the lower the error $\|Y - \hat{A}_r X \hat{B}_r\|_F^2$. As λ increases the penalty on the rank r becomes more important in the minimization procedure and $\hat{r}$ decreases. However, for moderate values of λ (*i.e.* approximately $\log(\lambda) \leq 5$) the smaller the noise variance level σ, the faster $\hat{r}$ decreases. Ultimately, for large values of λ (*i.e.* approximately $\log(\lambda) > 5$) the rate of decay of $\hat{r}$ as a function of λ no longer depends on σ.

The numerical value of λ is an important issue. We exhibit explicit (fast to calculate) procedures for the choice of this tuning parameter. In the case of *known noise variance*, the rule of thumb suggested by [4] indicates to choose

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

in Theorem 2.3 with $t = 0$, and $C = 2$. The two solid vertical lines represent $\lambda(4.5)$ (blue) and $\lambda(10)$ (green). With these choices of the tuning parameter we get successful estimators of the underlying rank of the signal $\hat{r} \approx 12 = r^*$. We underline that in the small noise regime the rank is slightly overestimated and in the strong noise regime it is slightly underestimated. This behaviour perfectly matches the results drawn from Figure 1 showing that overestimating the rank in small noise regime does not impact the performances and slightly underestimating it in strong noise regime improves the performances.

However, in real world applications the noise has *unknown variance*. This raises the question of how to choose a data-driven λ in this case, without deteriorating the prediction. This situation is more challenging as it first requires an estimator of σ^2 before using the previously exposed rule of thumb. We choose the initial value of r equal to $r_X \wedge n \wedge p$ and propose the r-dependent estimator $\widehat{\sigma}_r^2 := \dfrac{\|Y - \hat{A}_r X \hat{B}_r\|_F^2}{np - (m \wedge q) r_X}$. It allows to compute the previously defined $\lambda(\widehat{\sigma}_r)$ and using this data-driven tuning parameter we produce the rank estimator $\bar{r}$. This procedure takes r as an argument and returns $\lambda(\widehat{\sigma}_r)$ and $\bar{r}$. However, when r is substantially larger than r^*, $\hat{A}_r X \hat{B}_r$ is overfitting Y and performing this procedure once will not lead to a satisfying output $\bar{r}$. Hence we iterate while $\bar{r} < r$. We note $\lambda(\widehat{\sigma}_{\bar{r}})$ and $\bar{r}$ the final outputs of the procedure. The two dashed vertical lines represent $\lambda(\widehat{\sigma}_{\bar{r}})$ when $\sigma = 4.5$ (cyan) and $\sigma = 10$ (magenta). The proposed procedure exhibits great numerical properties.

Finally, numerical simulations generated in the same context, with different values for the true underlying ranks, show similar excellent prediction bounds, combined with correct rank selection. Together with the current case where $\min(r_A^*, r_X, r_B^*) = r_B^*$, we have explored successfully the cases $\min(r_A^*, r_X, r_B^*) = r_A^*$, $\min(r_A^*, r_X, r_B^*) = r_X$ and $\min(r_A^*, r_X, r_B^*) = r_A^* = r_X = r_B^*$.

5 Proofs

Basic facts For any matrix $M \in \mathbb{R}^{n \times m}$, $\|M\|_*^2 \leq r_M \|M\|_F^2$. In addition, for any matrices M_1 and M_2 in $\mathbb{R}^{n \times m}$, the following inequalities hold $\langle M_1, M_2 \rangle_F \leq \|M_1\|_* \|M_2\|_{op}$ and $\|M_1 +$

$M_2\|_F \leq \|M_1\|_F + \|M_2\|_F$. Furthermore, if we set $a = \operatorname{rank} M_1 \wedge \operatorname{rank} M_2$ then $\langle M_1, M_2\rangle_F \leq \|M_1\|_{(2,a)}\|M_2\|_{(2,a)}$.

Lemma 5.1 *Let E be a $n \times p$ random matrix whose entries are independent and having Gaussian distribution $\mathcal{N}(0, \sigma^2)$. If U and V belong to $\mathcal{O}_n$ and $\mathcal{O}_p$ respectively, then $E_0 := U^\top EV$ has independent entries with Gaussian distribution $\mathcal{N}(0, \sigma^2)$.*

Proof of Lemma 5.1. Note that we can vectorize the matrix E_0 and get that

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

where $vec(E)$ is a Gaussian vector of dimension np, centered, with variance $\sigma^2 I_{np}$. Moreover, the tensor product $V^\top \otimes U^\top$ belongs to $\mathcal{O}_{np}$, thus $vec(E_0)$ is still a Gaussian vector with distribution $\mathcal{N}_{np}(0, \sigma^2 I_{np})$. ■

Recall that, for an arbitrary matrix M, we denote $U_M \Sigma_M V_M^\top$ its SVD.

Lemma 5.2 *If M^* is a $n \times p$ matrix of rank r^*, than for any $r \leq n \wedge p$, we have*

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

and the infimum is attained by the projection $[M^*]_r$ of M^* on the space of $n \times p$ matrices with rank r given by the matrix

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

5.1 Proof of Theorem 2.1

Let $r \in [n \wedge p \wedge r_X]$ and $(\hat{A}_{0r}, \hat{B}_{0r})$ defined in (6). Let us denote here $M_0^* = A_0^* \Sigma_X B_0^*$ and $\hat{M}_0 = \hat{A}_{0r} \Sigma_X \hat{B}_{0r}$. By construction, $\hat{M}_0$ is the projection $[\Sigma_Y]_r$ of Σ_Y onto the set of matrices with rank less than or equal to r, in the sense of Lemma 5.2. Therefore,

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

We recall that in our model $\Sigma_Y = M_0^* + E_0$ which leads to

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

We expand the squares and arrange terms to get

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

Now, since $\operatorname{rank}(\hat{M}_0) = r$ and $\operatorname{rank}([M_0^*]_r) \leq r$, we get that $\operatorname{rank}(\hat{M}_0 - [M_0^*]_r) \leq 2r$. This inequality gives

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

We apply the inequality $2xy \leq \alpha x^2 + \alpha^{-1} y^2$ with $x, y \geq 0$ and $\alpha > 0$. We obtain, for real numbers $\alpha > 1$ and $\beta > 0$,

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

Let us use that $\|E_0\|_{(2,2r)}^2 \leq 2r \cdot \|E_0\|_{op}^2$ and Lemma 5.2 to further get

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

Noticing that for any matrices A_0, B_0 having rank less than or equal to r, $\operatorname{rank}(A_0 \Sigma_X B_0) \leq r_{A_0} \wedge r_X \wedge r_{B_0} \leq r$, we deduce that

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

Indeed, the second inf is taken over a possibly smaller family of matrices. We actually show that equality holds in the previous display. Indeed, by Lemma 5.2 we have that $\inf_{M : \operatorname{rank} M \leq r} \|M_0^* - M\|_F^2 = \sum_{k=r+1}^{r^*} \sigma_k(M_0^*)^2 \cdot \mathbf{1}_{r < r^*}$, where $r^* = \operatorname{rank}(M_0^*)$. Recall that $M_0^* = A_0^* \Sigma_X B_0^*$ is a product of diagonal matrices, giving that $r^* = \min(r_X, r_{A_0^*}, r_{B_0^*})$ and $\sigma_k(M_0^*) = \sigma_k(A_0^*) \sigma_k(X) \sigma_k(B_0^*) \cdot \mathbf{1}_{k \leq r^*}$. Thus, the particular choice

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

solves exactly the problem giving $M_0^* = A_{0r} \Sigma_X B_{0r}$. Finally,

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

Plugging this into (12) and considering the particular choice $\alpha = 3/2$ and $\beta = 1/2$ give the theorem:

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

The last step is the high-probability bound on $\|E_0\|_{op}$. Recall that $E_0 = U_Y^\top E V_Y$ with U_Y in $\mathcal{O}_n$ and V_Y in $\mathcal{O}_p$ and therefore E_0 and E have the same singular values. Therefore $\|E\|_{op} = \|E_0\|_{op}$. The noise matrix E has independent, centered, $\sigma-$sub-Gaussian entries and its spectral norm verifies (see [20]) for some absolute constant $C > 0$

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

Moreover, $\mathbb{E}\left[\|E\|_{op} \right] \leq \sqrt{C} \sigma (\sqrt{n} + \sqrt{p})$.

5.2 Proof of Corollary 2.2

Recall the notation $M_0^* = A_0^* \Sigma_X B_0^*$ and $\hat{M}_0 = \hat{A}_{0r} \Sigma_X \hat{B}_{0r}$ with $\hat{A}_{0r}$ and $\hat{B}_{0r}$ given by (6) and let us denote $M^* = A^* X B^*$ and $\hat{M} = \hat{A}_r X \hat{B}_r$ with $\hat{A}_r$ and $\hat{B}_r$ given by (7). Notice that the Frobenius norm and the rank are invariant under left or right multiplication by orthogonal matrices. Therefore, we follow the lines of the proof of Theorem 2.1 and see that $\|Y - \hat{M}\|_F^2 = \|\Sigma_Y - \hat{M}_0\|_F^2$ and $\operatorname{rank} M^* = \operatorname{rank} M_0^* = r^*$. Also, $\hat{M}$ is the projection $[Y]_r$ of Y on the space of matrices with rank less than or equal to r. Finally, the equality (13) can be pushed forward

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

Indeed, we have one-to-one transformations of A_0, B_0 into A, B, respectively, and equality of the Frobenius norms. This finishes the proof.

5.3 Proof of Theorem 2.3

By definition of $\hat{r} = \hat{r}(\lambda)$, we have that, for all $r \in [n \wedge p \wedge r_X]$,

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

Since $\hat{A}_r X \hat{B}_r$ is the projection $[Y]_r$ of Y on the space of matrices M with $\operatorname{rank} M \le r$, we get that for all matrices A and B such that $\operatorname{rank} A \wedge \operatorname{rank} B \le r$

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

Indeed, $\operatorname{rank}(AXB) \le r$ and Pythagora's theorem gives the former inequality. We deduce that

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

Next, replace $Y = A^* X B^* + E$, expand the squares and rearrange terms to get

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

Let us denote by $\hat{M}(\hat{r}) = \hat{A}_{\hat{r}} X \hat{B}_{\hat{r}}$, $M(r) = AXB$ and see that $\operatorname{rank}(\hat{M}(\hat{r}) - M(r)) \le \hat{r} + r$. We have

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

Then, using twice the inequality $2xy \leq \alpha x^2 + \alpha^{-1} y^2$ with $x, y \geq 0$ and $\alpha > 0$, we obtain for arbitrary real numbers $\alpha > 1$, $\beta > 0$:

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

Consequently, if $(\alpha + \beta)\|E\|_{op}^2 \leq \lambda$:

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

for all r in $[n \wedge p \wedge r_X]$ and all $M(r) = AXB$ with $\operatorname{rank} A \wedge \operatorname{rank} B \leq r$. We get the result by replacing again $\alpha = 3/2$ and $\beta = 1/2$. Then we use that

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

and the high-probability bounds in (14).

5.4 Proofs of results in Section 2.3

Proof of Proposition 2.4. For any r in $[n \wedge p \wedge r_X]$, we have that $\hat{A}_r X \hat{B}_r = [Y]_r$ is the projection of Y on the space of matrices having rank smaller than or equal to r. Now, write

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

It is easy to see that F as a function of r has a unique minimum at $r_Y(\lambda)$ if $\lambda > \sigma_{r_Y}(Y)^2$, but is minimal and constant for $r = r_Y, \ldots, (n \wedge p \wedge r_X)$ whenever $\lambda \leq \sigma_{r_Y}(Y)^2$. ■

Proof of Proposition 2.5. By definition of $\hat{r}$, we have $k > \hat{r}$ if and only if $\lambda > \sigma_k(Y)^2$ and $k < \hat{r}$ if and only if $\lambda \leq \sigma_{k+1}(Y)^2$. In our model $Y = A^* X B^* + E$, the Weyl inequality gives $|\sigma_k(A^* X B^*) - \sigma_k(Y)| \leq \sigma_1(E)$ for all k. The events on $\hat{r}$ can be written in terms of $\sigma_1(E) = \|E\|_{op}$ as follows. We have

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

Thus $\{\hat{r} \neq k\}$ implies either $\sigma_1(E) > \sigma_k(A^* X B^*) - \sqrt{\lambda}$ or $\sigma_1(E) \geq \sqrt{\lambda} - \sigma_{k+1}(A^* X B^*)$. Let us take $k = r^*(\lambda)$. Then the assumption that $\sigma_{r^*(\lambda)}(A^* X B^*) > (1+c)\sqrt{\lambda}$ gives that $\sigma_1(E) > c\sqrt{\lambda}$ and the assumption that $\sigma_{r^*(\lambda)+1}(A^* X B^*) < (1-c)\sqrt{\lambda}$ gives also that $\sigma_1(E) > c\sqrt{\lambda}$. Thus,

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

The proof is finished using the inequality (14). ■

5.5 Proof of Theorem 2.7

The optimization problem (9) can be written, after replacing $\widehat{\sigma}_r^2$, as follows

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

We denote by $\bar{M} = \hat{A}_{\bar{r}} X \hat{B}_{\bar{r}}$, $\hat{M}_r = \hat{A}_r X \hat{B}_r$ and $M^* = A^* X B^*$. With this notation it follows that, for $r \le r_{max}$,

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

Developing the squares and using the equality $Y = M^* + E$, we get

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

We now use the upper bound $\langle E, \bar{M} - \hat{M}_r\rangle_F \le \|E\|_{op}\|\bar{M} - \hat{M}_r\|_*$ and the definition of $\bar{M}$ and $\hat{M}_r$ to derive

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

Let us note that we use $\sigma_k(Y) = 0$ in case $k > r_Y$. We recall that $\|\bar{M} - \hat{M}_r\|_* \le \sqrt{r + \bar{r}} \cdot \|\bar{M} - \hat{M}_r\|_F$ and further obtain

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

Using twice the inequality $2ab \le \alpha a^2 + \alpha^{-1}b^2$ for $a, b > 0$, with $\alpha > 1$ first and with $\beta > 0$ second, we get

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

We now distinguish the two cases: $r \le \bar{r}$ and $r > \bar{r}$. In the first case, namely $r \le \bar{r}$, we bound from above as follows:

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

where we used Weyl inequality $\sigma_{r+1}(Y) \leq \sigma_{r+1}(M^*) + \|E\|_{op}$ leading to $\sigma_{r+1}(Y)^2 \leq 2\|E\|_{op}^2 + 2\sigma_{r+1}(M^*)^2$. We plug this into (15) to get

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

for all $r \leq \bar{r}$ belonging to $[r_{max}]$. Thus, for λ such that $\frac{2\lambda\cdot(r_{max}\vee r_Y)}{np} = (1+\varepsilon)(\alpha+\beta)$ for some $\varepsilon > 0$ we get

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

We now focus on the second case, namely $r > \bar{r}$. We observe that in this case,

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

by a similar reasoning in the previous case. We plug this into (15) to get

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

With the same choice of λ such that $\frac{2\lambda\cdot r_{max}\vee r_Y}{np} = (1+\varepsilon)(\alpha+\beta)$ for some $\varepsilon > 0$ we get also in this case that

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

Taking $\alpha = 3/2$ and $\beta = 1/2$ and combining both cases leads to the following result

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

where we choose λ such that $\lambda\cdot r_{max}\vee r_Y = (1+\varepsilon)np$ for some $\varepsilon > 0$. We conclude by using the inequality (14).

5.6 Proof of Theorem 3.1

We proceed by solving the problem in two steps for solving the optimization problem (10) which can be equivalently written as

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

for $\lambda > 0$. The solution to the problem in M is explicit and it is known to be obtained from Y by soft-thresholding of its eigenvalues: $\bar{M} = U_Y Diag_{n,p}((\sigma_k(Y) - \lambda)_+)V_Y^\top$, where we used the SVD of Y: $U_Y \Sigma_Y V_Y^\top$. Next, we project $\bar{M}$ on the space of matrices AXB for A and B in Frobenius norm. It is easy to check that our choice of $\bar{A}, \bar{B}$ are exact solutions, that is $\bar{M} = \bar{A}X\bar{B}$.

Similarly to the proof of Theorem 2.3, by applying the definition of $\bar{M}$, expanding the squares and rearranging terms we get for all M:

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

under the event that $\|E\|_{op}^2 \leq \lambda$. We use the decomposability of the nuclear norm of matrices as in [5], to find $\bar{M}_1$ and $\bar{M}_2$ such that $\bar{M} = \bar{M}_1 + \bar{M}_2$, $\|\bar{M}\|_* = \|\bar{M}_1\|_* + \|\bar{M}_2\|_*$ and $\|\bar{M} - M\|_* = \|\bar{M}_1 - M\|_* + \|\bar{M}_2\|_*$. Moreover, $\operatorname{rank}(\bar{M}_1) \leq 2\operatorname{rank}(M)$. This implies

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

We obtain for arbitrary real numbers $\alpha > 1$ and $\beta > 0$, for all M,

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

For the particular values $\alpha = 3/2$ and $\beta = 1/2$, we get

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

Recall that $\min_{M:\operatorname{rank} M=r} \|M^* - M\|_F^2 = \sum_{K=r+1}^{r^*} \sigma_K(M^*)^2 \cdot \mathbf{1}_{r<r^*}$ to get the final result.

6 Auxiliary results

Algorithm 1 Data-driven procedure for selecting $\bar{r}$ and λ

Input: data X, Y
Require: $np \geq (m \wedge q) r_X > 0$
Define: $\widehat{\sigma}_r^2 := \dfrac{\|Y - \hat{A}_r X \hat{B}_r\|_F^2}{np - (m \wedge q) r_X}$
Define: $\lambda(\sigma) := 4(n+p)\sigma^2$
Define: $\hat{r}_\lambda := \arg\min_{r \in [n \wedge p \wedge r_X]} \left(\|Y - \hat{A}_r X \hat{B}_r\|_F^2 + \lambda \cdot r \right)$
Initialize: $r \leftarrow r_X \wedge n \wedge p$, $\bar{r} \leftarrow \hat{r}_{\lambda(\widehat{\sigma}_r^2)}$
  while $\bar{r} < r$ **do**
    $r \leftarrow \bar{r}$
    $\bar{r} \leftarrow \hat{r}_{\lambda(\widehat{\sigma}_r^2)}$
  end while
  Output: $\bar{r}, \lambda(\widehat{\sigma}_{\bar{r}}^2)$

Acknowledgment. The authors thank the French National Research Agency (ANR) under the grant Labex Ecodec (ANR-11-LABEX-0047).

References

[1] Francis R Bach. Consistency of trace norm minimization. *The Journal of Machine Learning Research*, 9:1019–1048, 2008.

[2] Xin Bing, Florentina Bunea, and Marten Wegkamp. Optimal estimation of sparse topic models. *The Journal of Machine Learning Research*, 21(1):7189–7233, 2020.

[3] Xin Bing and Marten H. Wegkamp. Adaptive estimation of the rank of the coefficient matrix in high-dimensional multivariate response regression models. *Ann. Statist.*, 47(6):3157–3184, 2019.

[4] Lucien Birgé and Pascal Massart. Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138:33–73, 2007.

[5] Florentina Bunea, Yiyuan She, and Marten H. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309, 2011.

[6] Rong Chen, Han Xiao, and Dan Yang. Autoregressive models for matrix-valued time series. *Journal of Econometrics*, 222(1):539–560, 2021.

[7] David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? *Advances in neural information processing systems*, 16, 2003.

[8] Jianqing Fan, Yuan Liao, and Martina Mincheva. High dimensional covariance matrix estimation in approximate factor models. *Annals of statistics*, 39(6):3320, 2011.

[9] Christophe Giraud. Low rank multivariate regression. *Electron. J. Stat.*, 5:775–799, 2011.

[10] Nan-Jung Hsu, Hsin-Cheng Huang, and Ruey S. Tsay. Matrix autoregressive spatio-temporal models. *J. Comput. Graph. Statist.*, 30(4):1143–1155, 2021.

[11] Zheng Tracy Ke and Minzhe Wang. Using svd for topic modeling. *Journal of the American Statistical Association*, pages 1–16, 2022.

[12] Olga Klopp, Yu Lu, Alexandre B. Tsybakov, and Harrison H. Zhou. Structured matrix estimation and completion. *Bernoulli*, 25(4B):3883–3911, 2019.

[13] Olga Klopp, Maxim Panov, Suzanne Sigalla, and Alexandre Tsybakov. Assigning topics to documents by successive projections. *arXiv preprint arXiv:2107.03684*, 2021.

[14] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011.

[15] Chen Kun, Dong Hongbo, and Chan Kung-Sik. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100:901–920, 2013.

[16] Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Proceedings of the 27 th International Conference on Machine Learning*, 2011.

[17] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.*, 27(4):538–557, 2012.

[18] Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *Ann. Statist.*, 39(1):1–47, 2011.

[19] Angelika Rohde and Alexandre B Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

[20] Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. An introduction with applications in data science, With a foreword by Sara van de Geer.