Two-sided Matrix Regression

N. Bettache[1] C. Butucea[1]

[1]CREST

2022

Table of Contents

Table of Contents

Multivariate Linear Regression

- Collect $(y_1, \ldots, y_n)$ and $(x_1, \ldots, x_n)$ with $y_i \in \mathbb{R}^p$ and $x_i \in \mathbb{R}^q$.

Multivariate Linear Regression

- Collect $(y_1, \dots, y_n)$ and $(x_1, \dots, x_n)$ with $y_i \in \mathbb{R}^p$ and $x_i \in \mathbb{R}^q$.
- Form $Y \in \mathbb{R}^{n \times p}$ and $X \in \mathbb{R}^{n \times q}$.

Multivariate Linear Regression

- Collect $(y_1, \ldots, y_n)$ and $(x_1, \ldots, x_n)$ with $y_i \in \mathbb{R}^p$ and $x_i \in \mathbb{R}^q$.
- Form $Y \in \mathbb{R}^{n \times p}$ and $X \in \mathbb{R}^{n \times q}$.
- Assume $\exists B^* \in \mathbb{R}^{q \times p}$ s.t $Y = XB^* + E$ where E is a noise matrix.

Multivariate Linear Regression

- Collect $(y_1, \dots, y_n)$ and $(x_1, \dots, x_n)$ with $y_i \in \mathbb{R}^p$ and $x_i \in \mathbb{R}^q$.
- Form $Y \in \mathbb{R}^{n \times p}$ and $X \in \mathbb{R}^{n \times q}$.
- Assume $\exists B^* \in \mathbb{R}^{q \times p}$ s.t $Y = XB^* + E$ where E is a noise matrix.

$$\begin{pmatrix} Y_{11} & \cdots & Y_{1j} & \cdots & Y_{1p} \\ \vdots & & \vdots & & \vdots \\ Y_{i1} & \cdots & Y_{ij} & \cdots & Y_{ip} \\ \vdots & & \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{nj} & \cdots & Y_{np} \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1q} \\ \vdots & & \vdots \\ X_{i1} & \cdots & X_{iq} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{nq} \end{pmatrix} \cdot \begin{pmatrix} B^*_{11} & \cdots & B^*_{1j} & \cdots & B^*_{1p} \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ B^*_{q1} & \cdots & B^*_{qj} & \cdots & B^*_{qp} \end{pmatrix} + E$$

Multivariate Linear Regression

- Collect $(y_1, \dots, y_n)$ and $(x_1, \dots, x_n)$ with $y_i \in \mathbb{R}^p$ and $x_i \in \mathbb{R}^q$.
- Form $Y \in \mathbb{R}^{n \times p}$ and $X \in \mathbb{R}^{n \times q}$.
- Assume $\exists B^* \in \mathbb{R}^{q \times p}$ s.t $Y = XB^* + E$ where E is a noise matrix.

$$\begin{pmatrix} Y_{11} & \cdots & Y_{1j} & \cdots & Y_{1p} \\ \vdots & & \vdots & & \vdots \\ Y_{i1} & \cdots & Y_{ij} & \cdots & Y_{ip} \\ \vdots & & \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{nj} & \cdots & Y_{np} \end{pmatrix} =$$

$$\begin{pmatrix} X_{11} & \cdots & X_{1q} \\ \vdots & & \vdots \\ X_{i1} & \cdots & X_{iq} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{nq} \end{pmatrix} \cdot \begin{pmatrix} B^*_{11} & \cdots & B^*_{1j} & \cdots & B^*_{1p} \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ B^*_{q1} & \cdots & B^*_{qj} & \cdots & B^*_{qp} \end{pmatrix} + E$$

Multivariate Linear Regression

- Collect $(y_1, \ldots, y_n)$ and $(x_1, \ldots, x_n)$ with $y_i \in \mathbb{R}^p$ and $x_i \in \mathbb{R}^q$.
- Form $Y \in \mathbb{R}^{n \times p}$ and $X \in \mathbb{R}^{n \times q}$.
- Assume $\exists B^* \in \mathbb{R}^{q \times p}$ s.t $Y = XB^* + E$ where E is a noise matrix.

$$\begin{pmatrix} Y_{11} & \cdots & Y_{1j} & \cdots & Y_{1p} \\ \vdots & & \vdots & & \vdots \\ Y_{i1} & \cdots & Y_{ij} & \cdots & Y_{ip} \\ \vdots & & \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{nj} & \cdots & Y_{np} \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1q} \\ \vdots & & \vdots \\ X_{i1} & \cdots & X_{iq} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{nq} \end{pmatrix} \cdot \begin{pmatrix} B^*_{11} & \cdots & B^*_{1j} & \cdots & B^*_{1p} \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ B^*_{q1} & \cdots & B^*_{qj} & \cdots & B^*_{qp} \end{pmatrix} + E$$

Multivariate Linear Regression

- Collect $(y_1, \ldots, y_n)$ and $(x_1, \ldots, x_n)$ with $y_i \in \mathbb{R}^p$ and $x_i \in \mathbb{R}^q$.
- Form $Y \in \mathbb{R}^{n \times p}$ and $X \in \mathbb{R}^{n \times q}$.
- Assume $\exists B^* \in \mathbb{R}^{q \times p}$ s.t $Y = XB^* + E$ where E is a noise matrix.

$$\begin{pmatrix} Y_{11} & \cdots & Y_{1j} & \cdots & Y_{1p} \\ \vdots & & \vdots & & \vdots \\ Y_{i1} & \cdots & Y_{ij} & \cdots & Y_{ip} \\ \vdots & & \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{nj} & \cdots & Y_{np} \end{pmatrix} =$$

$$\begin{pmatrix} X_{11} & \cdots & X_{1q} \\ \vdots & & \vdots \\ X_{i1} & \cdots & X_{iq} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{nq} \end{pmatrix} \cdot \begin{pmatrix} B^*_{11} & \cdots & B^*_{1j} & \cdots & B^*_{1p} \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ B^*_{q1} & \cdots & B^*_{qj} & \cdots & B^*_{qp} \end{pmatrix} + E$$

Multivariate Linear Regression

- Collect $(y_1, \ldots, y_n)$ and $(x_1, \ldots, x_n)$ with $y_i \in \mathbb{R}^p$ and $x_i \in \mathbb{R}^q$.
- Form $Y \in \mathbb{R}^{n \times p}$ and $X \in \mathbb{R}^{n \times q}$.
- Assume $\exists B^* \in \mathbb{R}^{q \times p}$ s.t $Y = XB^* + E$ where E is a noise matrix.

$$\begin{pmatrix} Y_{11} & \cdots & Y_{1j} & \cdots & Y_{1p} \\ \vdots & & \vdots & & \vdots \\ Y_{i1} & \cdots & Y_{ij} & \cdots & Y_{ip} \\ \vdots & & \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{nj} & \cdots & Y_{np} \end{pmatrix} =$$

$$\begin{pmatrix} X_{11} & \cdots & X_{1q} \\ \vdots & & \vdots \\ X_{i1} & \cdots & X_{iq} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{nq} \end{pmatrix} \cdot \begin{pmatrix} B^*_{11} & \cdots & B^*_{1j} & \cdots & B^*_{1p} \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ B^*_{q1} & \cdots & B^*_{qj} & \cdots & B^*_{qp} \end{pmatrix} + E$$

Multivariate Linear Regression

- Collect $(y_1, \ldots, y_n)$ and $(x_1, \ldots, x_n)$ with $y_i \in \mathbb{R}^p$ and $x_i \in \mathbb{R}^q$.
- Form $Y \in \mathbb{R}^{n \times p}$ and $X \in \mathbb{R}^{n \times q}$.
- Assume $\exists B^* \in \mathbb{R}^{q \times p}$ s.t $Y = XB^* + E$ where E is a noise matrix.

$$\begin{pmatrix} Y_{1j} \\ \vdots \\ Y_{ij} \\ \vdots \\ Y_{nj} \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1k} & \cdots & X_{1q} \\ \vdots & & \vdots & & \vdots \\ X_{i1} & \cdots & X_{ik} & \cdots & X_{iq} \\ \vdots & & \vdots & & \vdots \\ X_{n1} & \cdots & X_{nk} & \cdots & X_{nq} \end{pmatrix} \cdot \begin{pmatrix} B^*_{1j} \\ \vdots \\ B^*_{kj} \\ \vdots \\ B^*_{qj} \end{pmatrix} + E$$

Multivariate Linear Regression

- Collect $(y_1, \ldots, y_n)$ and $(x_1, \ldots, x_n)$ with $y_i \in \mathbb{R}^p$ and $x_i \in \mathbb{R}^q$.
- Form $Y \in \mathbb{R}^{n \times p}$ and $X \in \mathbb{R}^{n \times q}$.
- Assume $\exists B^* \in \mathbb{R}^{q \times p}$ s.t $Y = XB^* + E$ where E is a noise matrix.

$$\begin{pmatrix} Y_{1j} \\ \vdots \\ Y_{ij} \\ \vdots \\ Y_{nj} \end{pmatrix} = B^*_{1j} \cdot \begin{pmatrix} X_{11} \\ \vdots \\ X_{i1} \\ \vdots \\ X_{n1} \end{pmatrix} + \cdots + B^*_{kj} \cdot \begin{pmatrix} X_{1k} \\ \vdots \\ X_{ik} \\ \vdots \\ X_{nk} \end{pmatrix} + \cdots + B^*_{qj} \cdot \begin{pmatrix} X_{1q} \\ \vdots \\ X_{iq} \\ \vdots \\ X_{nq} \end{pmatrix} + E$$

Multivariate Linear Regression

- Collect $(y_1, \dots, y_n)$ and $(x_1, \dots, x_n)$ with $y_i \in \mathbb{R}^p$ and $x_i \in \mathbb{R}^q$.
- Form $Y \in \mathbb{R}^{n \times p}$ and $X \in \mathbb{R}^{n \times q}$.
- Assume $\exists B^* \in \mathbb{R}^{q \times p}$ s.t $Y = XB^* + E$ where E is a noise matrix.

$$\forall j \in [p], \quad Y_j = \sum_{i=1}^{q} B^*_{ij} X_i$$

Multivariate Linear Regression

- Collect $(y_1, \ldots, y_n)$ and $(x_1, \ldots, x_n)$ with $y_i \in \mathbb{R}^p$ and $x_i \in \mathbb{R}^q$.
- Form $Y \in \mathbb{R}^{n\times p}$ and $X \in \mathbb{R}^{n\times q}$.
- Assume $\exists B^* \in \mathbb{R}^{q\times p} \quad \text{s.t} \quad Y = XB^* + E$ where E is a noise matrix.

$$\forall j \in [p], \quad Y_j = \sum_{i=1}^{q} B^*_{ij} X_i$$

- The columns of Y can be well explained by linear combinations of the columns of X.

Low-rank structure on B^*.

- Without any constraint on the structure of B^* (full rank), this is equivalent to performing p independent linear regressions.

Low-rank structure on B^*.

- Without any constraint on the structure of B^* (full rank), this is equivalent to performing p independent linear regressions.
- The j^{th} column of Y only depends on the j^{th} column of B^*.

Low-rank structure on B^*.

- Without any constraint on the structure of B^* (full rank), this is equivalent to performing p independent linear regressions.
- The j^{th} column of Y only depends on the j^{th} column of B^*.

$$\begin{pmatrix} Y_{11} & \cdots & Y_{1j} & \cdots & Y_{1p} \\ \vdots & & \vdots & & \vdots \\ Y_{i1} & \cdots & Y_{ij} & \cdots & Y_{ip} \\ \vdots & & \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{nj} & \cdots & Y_{np} \end{pmatrix} =$$

$$\begin{pmatrix} X_{11} & \cdots & X_{1q} \\ \vdots & & \vdots \\ X_{i1} & \cdots & X_{iq} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{nq} \end{pmatrix} \cdot \begin{pmatrix} B^*_{11} & \cdots & B^*_{1j} & \cdots & B^*_{1p} \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ B^*_{q1} & \cdots & B^*_{qj} & \cdots & B^*_{qp} \end{pmatrix} + E$$

Low-rank structure on B^*.

- Without any constraint on the structure of B^* (full rank), this is equivalent to performing p independent linear regressions.
- The j^{th} column of Y only depends on the j^{th} column of B^*.
- It ignores the multivariate nature of the response !

Low-rank structure on B^*.

- Without any constraint on the structure of B^* (full rank), this is equivalent to performing p independent linear regressions.
- The j^{th} column of Y only depends on the j^{th} column of B^*.
- It ignores the multivariate nature of the response !
- The columns of Y may be (heavily) correlated and the Least Squares estimator will not consider these correlations.

Low-rank structure on B^*.

- Without any constraint on the structure of B^* (full rank), this is equivalent to performing p independent linear regressions.
- The j^{th} column of Y only depends on the j^{th} column of B^*.
- It ignores the multivariate nature of the response !
- The columns of Y may be (heavily) correlated and the Least Squares estimator will not consider these correlations.
- Solution: impose a low-rank structure on B^*.

Low-rank structure on B^*.

- Without any constraint on the structure of B^* (full rank), this is equivalent to performing p independent linear regressions.
- The j^{th} column of Y only depends on the j^{th} column of B^*.
- It ignores the multivariate nature of the response !
- The columns of Y may be (heavily) correlated and the Least Squares estimator will not consider these correlations.
- Solution: impose a low-rank structure on B^*.
- This is studied in the literature.

How Y depends on the signal XB^* ?

- The j^{th} column of Y only depends on the j^{th} column of B^*.

How Y depends on the signal XB^* ?

- The j^{th} column of Y only depends on the j^{th} column of B^*.

$$\begin{pmatrix} Y_{11} & \cdots & Y_{1j} & \cdots & Y_{1p} \\ \vdots & & \vdots & & \vdots \\ Y_{i1} & \cdots & Y_{ij} & \cdots & Y_{ip} \\ \vdots & & \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{nj} & \cdots & Y_{np} \end{pmatrix} =$$

$$\begin{pmatrix} X_{11} & \cdots & X_{1q} \\ \vdots & & \vdots \\ X_{i1} & \cdots & X_{iq} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{nq} \end{pmatrix} \cdot \begin{pmatrix} B^*_{11} & \cdots & B^*_{1j} & \cdots & B^*_{1p} \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ B^*_{q1} & \cdots & B^*_{qj} & \cdots & B^*_{qp} \end{pmatrix} + E$$

How Y depends on the signal XB^* ?

- The j^{th} column of Y only depends on the j^{th} column of B^*.
- The i^{th} row of Y only depends on the i^{th} row of X.

How Y depends on the signal XB^* ?

- The j^{th} column of Y only depends on the j^{th} column of B^*.
- The i^{th} row of Y only depends on the i^{th} row of X.

$$\begin{pmatrix} Y_{11} & \cdots & Y_{1j} & \cdots & Y_{1p} \\ \vdots & & \vdots & & \vdots \\ Y_{i1} & \cdots & Y_{ij} & \cdots & Y_{ip} \\ \vdots & & \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{nj} & \cdots & Y_{np} \end{pmatrix} =$$

$$\begin{pmatrix} X_{11} & \cdots & X_{1q} \\ \vdots & & \vdots \\ X_{i1} & \cdots & X_{iq} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{nq} \end{pmatrix} \cdot \begin{pmatrix} B^*_{11} & \cdots & B^*_{1j} & \cdots & B^*_{1p} \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ B^*_{q1} & \cdots & B^*_{qj} & \cdots & B^*_{qp} \end{pmatrix} + E$$

How Y depends on the signal XB^* ?

- The j^{th} column of Y only depends on the j^{th} column of B^*.
- The i^{th} row of Y only depends on the i^{th} row of X.
- If the columns of Y are correlated, we can impose a low rank structure on B^*.

How Y depends on the signal XB^* ?

- The j^{th} column of Y only depends on the j^{th} column of B^*.
- The i^{th} row of Y only depends on the i^{th} row of X.
- If the columns of Y are correlated, we can impose a low rank structure on B^*.
- What if the rows of Y are correlated ?

How Y depends on the signal XB^* ?

- The j^{th} column of Y only depends on the j^{th} column of B^*.
- The i^{th} row of Y only depends on the i^{th} row of X.
- If the columns of Y are correlated, we can impose a low rank structure on B^*.
- What if the rows of Y are correlated ?
- The design matrix X is fixed so we cannot impose anything on its structure.

Example

- Do we have examples where we want to regress a matrix Y with correlated rows and columns on a fixed design matrix X ?

Example

- Do we have examples where we want to regress a matrix Y with correlated rows and columns on a fixed design matrix X ?
- Economic data store economic indicators as column features and countries as rows.

Example

- Do we have examples where we want to regress a matrix Y with correlated rows and columns on a fixed design matrix X ?
- Economic data store economic indicators as column features and countries as rows.
- $$Y = \begin{array}{c} \\ Country_1 \\ \vdots \\ Country_n \end{array} \begin{array}{c} \begin{array}{ccc} Indicator_1 & \cdots & Indicator_p \end{array} \\ \left(\begin{array}{ccc} & & \\ & & \\ & & \end{array} \right) \end{array}$$

Example

- Do we have examples where we want to regress a matrix Y with correlated rows and columns on a fixed design matrix X ?
- Economic data store economic indicators as column features and countries as rows.
- It can be explained by a smaller matrix containing a smaller number of countries (geographical or economic representatives) and a few economic features (one representative for each category).

Example

- Do we have examples where we want to regress a matrix Y with correlated rows and columns on a fixed design matrix X ?
- Economic data store economic indicators as column features and countries as rows.
- It can be explained by a smaller matrix containing a smaller number of countries (geographical or economic representatives) and a few economic features (one representative for each category).
-

$$X = \begin{matrix} & GPD & UR & CPI & IR & GD & CR \\ USA & & & & & & \\ CAN & & & & & & \\ JPN & & & & & & \\ CHN & & & & & & \\ IND & & & & & & \\ FRA & & & & & & \\ GER & & & & & & \end{matrix}$$

Example

- Do we have examples where we want to regress a matrix Y with correlated rows and columns on a fixed design matrix X ?
- Economic data store economic indicators as column features and countries as rows.
- It can be explained by a smaller matrix containing a smaller number of countries (geographical or economic representatives) and a few economic features (one representative for each category).
- Other cases: meteorological data, medical or pharmaceutical data and so on.

Table of Contents

Two-Sided Matrix Regression

- Observe the matrix $Y \in \mathbb{R}^{n \times p}$ and a design matrix $X \in \mathbb{R}^{m \times q}$.

Two-Sided Matrix Regression

- Observe the matrix $Y \in \mathbb{R}^{n \times p}$ and a design matrix $X \in \mathbb{R}^{m \times q}$.
- They are related via the 2MR model

$$Y = A^* X B^* + E.$$

Two-Sided Matrix Regression

- Observe the matrix $Y \in \mathbb{R}^{n \times p}$ and a design matrix $X \in \mathbb{R}^{m \times q}$.
- They are related via the 2MR model

$$Y = A^* X B^* + E.$$

- Two parameter matrices $A^* \in \mathbb{R}^{n \times m}$ and $B^* \in \mathbb{R}^{q \times p}$:

Two-Sided Matrix Regression

- Observe the matrix $Y \in \mathbb{R}^{n \times p}$ and a design matrix $X \in \mathbb{R}^{m \times q}$.
- They are related via the 2MR model

$$Y = A^* X B^* + E.$$

- Two parameter matrices $A^* \in \mathbb{R}^{n \times m}$ and $B^* \in \mathbb{R}^{q \times p}$: low-rank.

Two-Sided Matrix Regression

- Observe the matrix $Y \in \mathbb{R}^{n \times p}$ and a design matrix $X \in \mathbb{R}^{m \times q}$.
- They are related via the 2MR model

$$Y = A^* X B^* + E.$$

- Two parameter matrices $A^* \in \mathbb{R}^{n \times m}$ and $B^* \in \mathbb{R}^{q \times p}$: low-rank.
- The noise matrix E is assumed to have independent centered $\sigma-$sub-Gaussian entries.

Two-Sided Matrix Regression

- Observe the matrix $Y \in \mathbb{R}^{n \times p}$ and a design matrix $X \in \mathbb{R}^{m \times q}$.
- They are related via the 2MR model

$$Y = A^* X B^* + E.$$

- Two parameter matrices $A^* \in \mathbb{R}^{n \times m}$ and $B^* \in \mathbb{R}^{q \times p}$: low-rank.
- The noise matrix E is assumed to have independent centered $\sigma-$sub-Gaussian entries.
- Objective: Retrieve the signal $A^* X B^*$.

Two-Sided Matrix Regression

- Observe the matrix $Y \in \mathbb{R}^{n \times p}$ and a design matrix $X \in \mathbb{R}^{m \times q}$.
- They are related via the 2MR model

$$Y = A^* X B^* + E.$$

- Two parameter matrices $A^* \in \mathbb{R}^{n \times m}$ and $B^* \in \mathbb{R}^{q \times p}$: low-rank.
- The noise matrix E is assumed to have independent centered $\sigma-$sub-Gaussian entries.
- Objective: Retrieve the signal $A^* X B^*$.
- ⚠: The problem is not convex anymore !

Related models

$$Y \in \mathbb{R}^{n \times p} \quad \text{and} \quad X \in \mathbb{R}^{m \times q},$$

$$Y = A^* X B^* + E.$$

The 2MR model encompasses known models:

Related models

$$Y \in \mathbb{R}^{n \times p} \quad \text{and} \quad X \in \mathbb{R}^{m \times q},$$

$$Y = A^* X B^* + E.$$

The 2MR model encompasses known models:

- If $n = m$ and A^* is the identity, the 2MR model becomes the (one-sided) *matrix regression* (MR) model $Y = XB^* + E$.

Related models

$$Y \in \mathbb{R}^{n\times p} \quad \text{and} \quad X \in \mathbb{R}^{m\times q},$$

$$Y = A^* X B^* + E.$$

The 2MR model encompasses known models:

- If $n = m$ and A^* is known to be the identity, the 2MR model becomes the (one-sided) *matrix regression* (MR) model $Y = XB^* + E$.
- If $m = q$ and X is the identity matrix, the 2MR model becomes a rank m *factorisation model* of the signal $M^* = A^* B^*$ observed with noise.

Related models

$$Y \in \mathbb{R}^{n\times p} \quad \text{and} \quad X \in \mathbb{R}^{m\times q},$$

$$Y = A^* X B^* + E.$$

The 2MR model encompasses known models:

- If $n = m$ and A^* is known to be the identity, the 2MR model becomes the (one-sided) *matrix regression* (MR) model $Y = XB^* + E$.
- If $m = q$ and X is the identity matrix, the 2MR model becomes a rank m *factorisation model* of the signal $M^* = A^* B^*$ observed with noise.

Unifies Low-rank Matrix Regression and Low-Rank Matrix Factorization under a same framework.

Table of Contents

Objective

If we know $r = \operatorname{rank} A^* X B^*$ we can exploit it.

Objective

- Let's fix $r \in [n \wedge p \wedge r_X]$ where $r_X = \operatorname{rank} X$.

Objective

- Let's fix $r \in [n \wedge p \wedge r_X]$ where $r_X = \operatorname{rank} X$.
- Let us build explicit predictors $(\hat{A}_r, \hat{B}_r)$ solutions to the non-convex constrained minimization problem:

Objective

- Let's fix $r \in [n \wedge p \wedge r_X]$ where $r_X = \operatorname{rank} X$.
- Let us build explicit predictors $(\hat{A}_r, \hat{B}_r)$ solutions to the non-convex constrained minimization problem:

$$\min_{\substack{A,B: \\ \operatorname{rank} A \wedge \operatorname{rank} B \leq r}} \quad \|Y - AXB\|_F^2.$$

Objective

- Let's fix $r \in [n \wedge p \wedge r_X]$ where $r_X = \operatorname{rank} X$.
- Let us build explicit predictors $(\hat{A}_r, \hat{B}_r)$ solutions to the non-convex constrained minimization problem:

$$\min_{\substack{A,B:\\ \operatorname{rank} A \wedge \operatorname{rank} B \leq r}} \|Y - AXB\|_F^2.$$

- Note: $\operatorname{rank} A^* X B^* \leq \min(\operatorname{rank} A^*, \operatorname{rank} X, \operatorname{rank} B^*)$.

Objective

- Let's fix $r \in [n \wedge p \wedge r_X]$ where $r_X = \operatorname{rank} X$.
- Let us build explicit predictors $(\hat{A}_r, \hat{B}_r)$ solutions to the non-convex constrained minimization problem:

$$\min_{\substack{A,B:\\ \operatorname{rank} A \wedge \operatorname{rank} B \leq r}} \|Y - AXB\|_F^2.$$

- Note: $\operatorname{rank} A^* X B^* \leq \min(\operatorname{rank} A^*, \operatorname{rank} X, \operatorname{rank} B^*)$.
- Intuition: There is lost information in the product and we can only hope to recover predictors $\hat{A}$ and $\hat{B}$ with respective ranks no more than r.

Objective

- Let's fix $r \in [n \wedge p \wedge r_X]$ where $r_X = \operatorname{rank} X$.
- Let us build explicit predictors $(\hat{A}_r, \hat{B}_r)$ solutions to the non-convex constrained minimization problem:

$$\min_{\substack{A,B:\\ \operatorname{rank} A \wedge \operatorname{rank} B \leq r}} \|Y - AXB\|_F^2.$$

- Note: $\operatorname{rank} A^* XB^* \leq \min(\operatorname{rank} A^*, \operatorname{rank} X, \operatorname{rank} B^*)$.
- Intuition: There is lost information in the product and we can only hope to recover predictors $\hat{A}$ and $\hat{B}$ with respective ranks no more than r.
- Global idea: $Y \longrightarrow Y_r \longrightarrow \hat{A} X \hat{B}$.

Rewriting of the model

The Frobenius norm is unitarily invariant and the SVD brings out unitary matrices.

Rewriting of the model

- The model can be re-written using the SVD of Y and X as follows:

Rewriting of the model

- The model can be re-written using the SVD of Y and X as follows:

$$Y = A^* X B^* + E$$

Rewriting of the model

- The model can be re-written using the SVD of Y and X as follows:

$$Y = A^* X B^* + E$$

$$U_Y \Sigma_Y V_Y^\top = A^* U_X \Sigma_X V_X^\top B^* + E$$

Rewriting of the model

- The model can be re-written using the SVD of Y and X as follows:

$$Y = A^* X B^* + E$$

$$U_Y \Sigma_Y V_Y^\top = A^* U_X \Sigma_X V_X^\top B^* + E$$

$$\Sigma_Y = U_Y^\top A^* U_X \Sigma_X V_X^\top B^* V_Y + U_Y^\top E V_Y$$

Rewriting of the model

- The model can be re-written using the SVD of Y and X as follows:

$$Y = A^* X B^* + E$$

$$U_Y \Sigma_Y V_Y^\top = A^* U_X \Sigma_X V_X^\top B^* + E$$

$$\Sigma_Y = \left(U_Y^\top A^* U_X \right) \Sigma_X \left(V_X^\top B^* V_Y \right) + U_Y^\top E V_Y$$

Rewriting of the model

- The model can be re-written using the SVD of Y and X as follows:

$$Y = A^* X B^* + E$$

$$U_Y \Sigma_Y V_Y^\top = A^* U_X \Sigma_X V_X^\top B^* + E$$

$$\Sigma_Y = \underbrace{\left(U_Y^\top A^* U_X\right)}_{A_0^*} \Sigma_X \underbrace{\left(V_X^\top B^* V_Y\right)}_{B_0^*} + U_Y^\top E V_Y$$

Rewriting of the model

- The model can be re-written using the SVD of Y and X as follows:

$$Y = A^* X B^* + E$$

$$U_Y \Sigma_Y V_Y^\top = A^* U_X \Sigma_X V_X^\top B^* + E$$

$$\Sigma_Y = \underbrace{\left(U_Y^\top A^* U_X\right)}_{A_0^*} \Sigma_X \underbrace{\left(V_X^\top B^* V_Y\right)}_{B_0^*} + U_Y^\top E V_Y$$

$$\Sigma_Y = A_0^* \Sigma_X B_0^* + E_0$$

Rewriting of the model

- The model can be re-written using the SVD of Y and X as follows:

$$Y = A^* X B^* + E$$

$$\Sigma_Y = A_0^* \Sigma_X B_0^* + E_0$$

Rewriting of the model

- The model can be re-written using the SVD of Y and X as follows:

$$Y = A^* X B^* + E$$

$$\Sigma_Y = A_0^* \Sigma_X B_0^* + E_0$$

- This leads, for any matrices A, B, to:

$$\|Y - AXB\|_F^2 = \|\Sigma_Y - U_Y^\top A U_X \Sigma_X V_X^T B V_Y\|_F^2,$$

Rewriting of the model

- The model can be re-written using the SVD of Y and X as follows:

$$Y = A^* X B^* + E$$

$$\Sigma_Y = A_0^* \Sigma_X B_0^* + E_0$$

- This leads, for any matrices A, B, to:

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

where $A_0 = U_Y^\top A U_X$ and $B_0 = V_X^\top B V_Y$.

Rewriting of the model

- The model can be re-written using the SVD of Y and X as follows:

$$Y = A^* X B^* + E$$

$$\Sigma_Y = A_0^* \Sigma_X B_0^* + E_0$$

- This leads, for any matrices A, B, to:

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

where $A_0 = U_Y^\top A U_X$ and $B_0 = V_X^\top B V_Y$.

- A and A_0 have the same rank, idem for B and B_0!

Rewriting of the model

- The model can be re-written using the SVD of Y and X as follows:

$$Y = A^* X B^* + E$$

$$\Sigma_Y = A_0^* \Sigma_X B_0^* + E_0$$

- This leads, for any matrices A, B, to:

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

where $A_0 = U_Y^\top A U_X$ and $B_0 = V_X^\top B V_Y$.

- The initial problem is equivalent to

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

Solution of the re-written problem

- We wish to solve

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

Solution of the re-written problem

- We wish to solve

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

The objective is

$$\left\| \underbrace{\begin{pmatrix} \sigma_1(Y) & & & \\ & \ddots & & \\ & & \sigma_{r_Y}(Y) & \\ & & & 0 \end{pmatrix}}_{n \times p} - A_0 \underbrace{\begin{pmatrix} \sigma_1(X) & & & \\ & \ddots & & \\ & & \sigma_{r_X}(X) & \\ & & & 0 \end{pmatrix}}_{m \times q} B_0 \right\|_F^2 .$$

Solution of the re-written problem

- We wish to solve

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

- A natural choice is

$$\hat{A}_{0r} = \underbrace{\begin{pmatrix} \sigma_1(Y) & & & \\ & \ddots & & \\ & & \sigma_{r \wedge r_Y}(Y) & \\ & & & 0 \end{pmatrix}}_{n \times m} = Diag_{n,m}(\sigma_k(Y), \, k \leq r \wedge r_Y)$$

$$\hat{B}_{0r} = \underbrace{\begin{pmatrix} \sigma_1(X)^{-1} & & & \\ & \ddots & & \\ & & \sigma_r(X)^{-1} & \\ & & & 0 \end{pmatrix}}_{q \times p} = Diag_{q,p}(\sigma_k(X)^{-1}, \, k \leq r)$$

Solution of the re-written problem

- We wish to solve

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

- $(\hat{A}_{0r}, \hat{B}_{0r})$ belongs to the set of solutions of the re-written problem.

Solution of the re-written problem

- We wish to solve

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

- $(\hat{A}_{0r}, \hat{B}_{0r})$ belongs to the set of solutions of the re-written problem.

$$\|\Sigma_Y - \hat{A}_{0r} \Sigma_X \hat{B}_{0r}\|_F^2 = \min_{\substack{A_0, B_0: \\ \text{rank } A_0 \wedge \text{rank } B_0 \leq r}} \|\Sigma_Y - A_0 \Sigma_X B_0\|_F^2.$$

Solution of the re-written problem

- We wish to solve

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

- $(\hat{A}_{0r}, \hat{B}_{0r})$ belongs to the set of solutions of the re-written problem.

$$\|\Sigma_Y - \hat{A}_{0r} \Sigma_X \hat{B}_{0r}\|_F^2 = \min_{\substack{A_0, B_0: \\ \text{rank}\, A_0 \wedge \text{rank}\, B_0 \leq r}} \|\Sigma_Y - A_0 \Sigma_X B_0\|_F^2.$$

- The predictor $\hat{A}_{0r} \Sigma_X \hat{B}_{0r}$ is the projection of Σ_Y onto the space of matrices with rank no more than r.

Solution of the re-written problem

- We wish to solve

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

- $(\hat{A}_{0r}, \hat{B}_{0r})$ belongs to the set of solutions of the re-written problem.

$$\|\Sigma_Y - \hat{A}_{0r} \Sigma_X \hat{B}_{0r}\|_F^2 = \min_{\substack{A_0, B_0: \\ \text{rank } A_0 \wedge \text{rank } B_0 \leq r}} \|\Sigma_Y - A_0 \Sigma_X B_0\|_F^2.$$

- We want to know how far the predictor $\hat{A}_{0r} \Sigma_X \hat{B}_{0r}$ is to the signal $A_0^* \Sigma_X B_0^*$.

Oracle inequality in the fixed rank case

- The predictor $\hat{A}_{0r}\Sigma_X\hat{B}_{0r}$ satisfies for $C > 0$ and for any $t > 0$:

Oracle inequality in the fixed rank case

- The predictor $\hat{A}_{0r}\Sigma_X\hat{B}_{0r}$ satisfies for $C > 0$ and for any $t > 0$:

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

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

Oracle inequality in the fixed rank case

- The predictor $\hat{A}_{0r}\Sigma_X\hat{B}_{0r}$ satisfies for $C > 0$ and for any $t > 0$:

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

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

- The value $\inf_{\substack{A_0,B_0: \\ \text{rank } A_0 \wedge \text{rank } B_0 \leq r}} \|A_0^*\Sigma_X B_0^* - A_0\Sigma_X B_0\|_F^2$ is know:

Oracle inequality in the fixed rank case

- The predictor $\hat{A}_{0r}\Sigma_X\hat{B}_{0r}$ satisfies for $C > 0$ and for any $t > 0$:

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

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

- The value $\inf_{\substack{A_0,B_0: \\ \text{rank } A_0 \wedge \text{rank } B_0 \leq r}} \|A_0^*\Sigma_X B_0^* - A_0\Sigma_X B_0\|_F^2$ is know:

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

Oracle inequality in the fixed rank case

- The predictor $\hat{A}_{0r}\Sigma_X\hat{B}_{0r}$ satisfies for $C > 0$ and for any $t > 0$:

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

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

- The value $\inf_{\substack{A_0,B_0:\\ \text{rank } A_0 \wedge \text{rank } B_0 \leq r}} \|A_0^*\Sigma_X B_0^* - A_0\Sigma_X B_0\|_F^2$ is know:

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

- $\mathcal{O}\left(r(n+p)\right)$ is the minimax optimal rate in the (one-sided) *matrix regression* (MR) model.

Solution of the initial problem

- From the explicit solutions $(\hat{A}_{0r}, \hat{B}_{0r})$ we deduce $(\hat{A}_r, \hat{B}_r)$ solution to the initial problem:

Solution of the initial problem

- From the explicit solutions $(\hat{A}_{0r}, \hat{B}_{0r})$ we deduce $(\hat{A}_r, \hat{B}_r)$ solution to the initial problem:

$$\hat{A}_r = U_Y \hat{A}_{0r} U_X^\top,$$
$$\hat{B}_r = V_X \hat{B}_{0r} V_Y^\top.$$

Solution of the initial problem

- From the explicit solutions $(\hat{A}_{0r}, \hat{B}_{0r})$ we deduce $(\hat{A}_r, \hat{B}_r)$ solution to the initial problem:

$$\hat{A}_r = U_Y \hat{A}_{0r} U_X^\top,$$
$$\hat{B}_r = V_X \hat{B}_{0r} V_Y^\top.$$

- They share the same ranks !

Solution of the initial problem

- From the explicit solutions $(\hat{A}_{0r}, \hat{B}_{0r})$ we deduce $(\hat{A}_r, \hat{B}_r)$ solution to the initial problem:

$$\hat{A}_r = U_Y \hat{A}_{0r} U_X^\top,$$
$$\hat{B}_r = V_X \hat{B}_{0r} V_Y^\top.$$

- They share the same ranks !
- The predictor $\hat{A}_r X \hat{B}_r$ satisfies for $C > 0$ and for any $t > 0$:

Solution of the initial problem

- From the explicit solutions $(\hat{A}_{0r}, \hat{B}_{0r})$ we deduce $(\hat{A}_r, \hat{B}_r)$ solution to the initial problem:

$$\hat{A}_r = U_Y \hat{A}_{0r} U_X^\top,$$

$$\hat{B}_r = V_X \hat{B}_{0r} V_Y^\top.$$

- They share the same ranks !
- The predictor $\hat{A}_r X \hat{B}_r$ satisfies for $C > 0$ and for any $t > 0$:

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

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

Alternative predictors

- There is an identifiability issue and the predictors are not uniquely defined in this setting.

Alternative predictors

- There is an identifiability issue and the predictors are not uniquely defined in this setting.
- Consider $(\alpha \hat{A}_{0r}, \frac{1}{\alpha}\hat{B}_{0r})$ with arbitrary $\alpha > 0$.

Alternative predictors

- There is an identifiability issue and the predictors are not uniquely defined in this setting.
- Consider $(\alpha \hat{A}_{0r}, \frac{1}{\alpha} \hat{B}_{0r})$ with arbitrary $\alpha > 0$.
- Let λ_i for all $i \leq m \wedge q$ be arbitrary positive numbers, then

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

Alternative predictors

- There is an identifiability issue and the predictors are not uniquely defined in this setting.
- Consider $(\alpha \hat{A}_{0r}, \frac{1}{\alpha}\hat{B}_{0r})$ with arbitrary $\alpha > 0$.
- Let λ_i for all $i \leq m \wedge q$ be arbitrary positive numbers, then

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

- Without further strong assumptions, we can only hope to learn the global signal, and not the parameters of the model.

Table of Contents

Rank-adaptive procedure

- How to derive a rank-adaptive procedure ?

Rank-adaptive procedure

- How to derive a rank-adaptive procedure ?
- For $\lambda \geq C_1(1+t)^2\sigma^2(n+p)$ with $C_1 > 0,\ t > 0$, consider

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

Rank-adaptive procedure

- How to derive a rank-adaptive procedure ?
- For $\lambda \geq C_1(1+t)^2\sigma^2(n+p)$ with $C_1 > 0,\ t > 0$, consider

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

Then,

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

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

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?
- If for some constant c in (0,1), $\sigma_{r^*}(A^*XB^*)^2 > (1+c)^2\lambda$, then

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

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?
- If for some constant c in (0,1), $\sigma_{r^*}(A^* X B^*)^2 > (1+c)^2 \lambda$, 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 $1 - 2\exp(-t^2(\sqrt{n} + \sqrt{p})^2)$.

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?
- 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 $1 - 2\exp(-t^2(\sqrt{n} + \sqrt{p})^2)$.
- The rank selector requires λ to be lower bounded by a function of σ^2.

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?
- If for some constant c in (0,1), $\sigma_{r^*}(A^* X B^*)^2 > (1+c)^2 \lambda$, then

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

- The rank selector requires λ to be lower bounded by a function of σ^2. What if we don't have access to σ^2 ?

Table of Contents

Unknown σ case

- In previous situations, λ needed to be lower bounded by a function of σ^2.

Unknown σ case

- In previous situations, λ needed to be lower bounded by a function of σ^2.
- What can we do if σ is unknown ?

Unknown σ case

- In previous situations, λ needed to be lower bounded by a function of σ^2.
- What can we do if σ is unknown ?
- Consider the following σ^2 estimator

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

Unknown σ case

- In previous situations, λ needed to be lower bounded by a function of σ^2.
- What can we do if σ is unknown ?
- Consider the following σ^2 estimator

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

- Consider the data-driven rank-adaptive procedure

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

Unknown σ case

- In previous situations, λ needed to be lower bounded by a function of σ^2.
- What can we do if σ is unknown ?
- Consider the following σ^2 estimator

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

- Consider the data-driven rank-adaptive procedure

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

- If $r_{max} \geq r^*$ and $\lambda = 2np/(r_{max} \vee r_Y)$, then for any $t > 0$:

$$\|A^* X B^* - \hat{A}_{\bar{r}} X \hat{B}_{\bar{r}}\|_F^2 \leq C_2(1+t)^2 \cdot \sigma^2 r_{max}(n+p),$$

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

Unknown σ case

- In previous situations, λ needed to be lower bounded by a function of σ^2.
- What can we do if σ is unknown ?
- Consider the following σ^2 estimator

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

- Consider the data-driven rank-adaptive procedure

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

- If $r_{max} \geq r^*$ and $\lambda = 2np/(r_{max} \vee r_Y)$, then for any $t > 0$:

$$\|A^* X B^* - \hat{A}_{\bar{r}} X \hat{B}_{\bar{r}}\|_F^2 \leq C_2(1+t)^2 \cdot \sigma^2 r_{max}(n+p),$$

with probability larger than $1 - 2\exp(-t^2(\sqrt{n} + \sqrt{p})^2)$.
- Similar as in the known σ case !

Table of Contents

Numerical simulations

- How to numerically choose λ ?

Numerical simulations

- How to numerically choose λ ?
- We derive explicit and fast to calculate procedures !

Numerical simulations

- How to numerically choose λ ?
- We derive explicit and fast to calculate procedures !
- Great numerical performances in various settings.

What's next ?

- What if we observe a collection of matrices (Y_i, X_i) ?

What's next ?

- What if we observe a collection of matrices (Y_i, X_i) ?
- What if we model a matrix autoregressive process with the 2MR model $Y_{t+1} = A^* Y_t B^* + E_t$?

What's next ?

- What if we observe a collection of matrices (Y_i, X_i) ?
- What if we model a matrix autoregressive process with the 2MR model $Y_{t+1} = A^* Y_t B^* + E_t$?
- What if we impose other sparsity assumptions on A^* and B^* ?

END

Thanks for listening !

Table of Contents

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?
- Consider the λ-rank of a matrix M, $r_M(\lambda)$, as the number of singular values above $\sqrt{\lambda}$.

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?
- Consider the λ-rank of a matrix M, $r_M(\lambda)$, as the number of singular values above $\sqrt{\lambda}$.

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

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?
- Consider the λ-rank of a matrix M, $r_M(\lambda)$, as the number of singular values above $\sqrt{\lambda}$.

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

It performs a hard thresholding of the singular values !

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?
- Consider the λ-rank of a matrix M, $r_M(\lambda)$, as the number of singular values above $\sqrt{\lambda}$.

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

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

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?
- Consider the λ-rank of a matrix M, $r_M(\lambda)$, as the number of singular values above $\sqrt{\lambda}$.

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

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

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

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?
- Consider the λ-rank of a matrix M, $r_M(\lambda)$, as the number of singular values above $\sqrt{\lambda}$.

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

- If the λ-rank of the signal $A^* X B^*$ is well separated, the procedure retrieves it with high probability.

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?

- 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

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

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?

- 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

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

- $r^*(\lambda)$ coincides with the true underlying rank r^* is equivalent to having $\sigma_{r^*}(A^*XB^*)^2 \geq \lambda > 0$.

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?
- 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

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

- $r^*(\lambda)$ coincides with the true underlying rank r^* is equivalent to having $\sigma_{r^*}(A^*XB^*)^2 \geq \lambda > 0$.
- It is 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 r^* selected by $\hat{r}$.

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?

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

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

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?

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

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?

- 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 $1 - 2\exp(-t^2(\sqrt{n} + \sqrt{p})^2)$.
- The rank selector requires λ to be lower bounded by a function of σ^2.

Consistent rank selection

- Can we retrieve the true rank of the signal with high probability ?

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

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

- The rank selector requires λ to be lower bounded by a function of σ^2. What if we don't have access to σ^2 ?

Simulation context

- Consider $n = 100$ and $p = 300$ with $Y \in \mathbb{R}^{n \times p}$ together with $m = 50$ and $q = 60$ with $X \in \mathbb{R}^{m \times q}$.

Simulation context

- Consider $n = 100$ and $p = 300$ with $Y \in \mathbb{R}^{n \times p}$ together with $m = 50$ and $q = 60$ with $X \in \mathbb{R}^{m \times q}$.
- We randomly generate three matrices: A^*, B^*, and X, with independent random gaussian entries with mean 0 and variance 1.

Simulation context

- Consider $n = 100$ and $p = 300$ with $Y \in \mathbb{R}^{n \times p}$ together with $m = 50$ and $q = 60$ with $X \in \mathbb{R}^{m \times q}$.
- 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 $r_X = 25$.

Simulation context

- Consider $n = 100$ and $p = 300$ with $Y \in \mathbb{R}^{n \times p}$ together with $m = 50$ and $q = 60$ with $X \in \mathbb{R}^{m \times q}$.
- 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 $A^* X B^*$ and shows a rank of 12 in all experiments.

Simulation context

- Consider $n = 100$ and $p = 300$ with $Y \in \mathbb{R}^{n \times p}$ together with $m = 50$ and $q = 60$ with $X \in \mathbb{R}^{m \times q}$.
- 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 $A^* X B^*$ and shows a rank of 12 in all experiments.
- We define various settings for the variance σ^2 of the Gaussian noise E so that the signal-to-noise ratio $SNR := \sigma_{r^*}(A^* X B^*)^2 / \sigma_1(E)^2$ varies approximately in the range $[0.5, 2]$.

Predictor performances

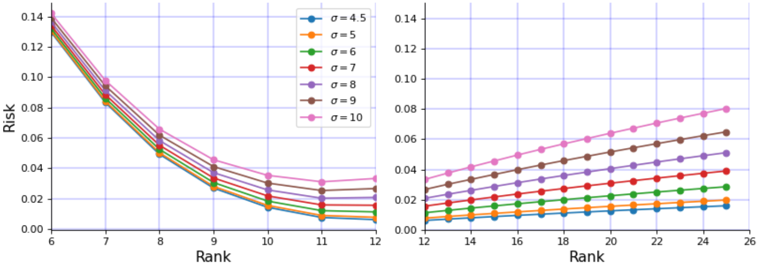

Figure: Evolution of the risk $\frac{\|\hat{A}_r X \hat{B}_r - A^* X B^*\|_F^2}{\|A^* X B^*\|_F^2}$ in function of r for different values of σ.

Rank recovering

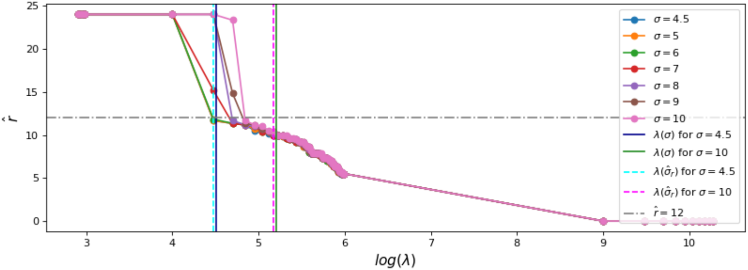

Figure: Evolution of the estimated $\hat{r}$ as a function of $\log(\lambda)$ for different values of σ.