Eliminating the Invariance on the Loss Landscape of Linear Autoencoders

Reza Oftadeh, Jiayi Shen, Atlas Wang, Dylan Shell

Texas A&M University
Department of Computer Sciense and Engineering

ICML 2020

Overview

- Linear Autoencoder (LAE) with Mean Square Error (MSE). The classical results:
  - Loss surface has been analytically characterized.
  - All local minima are global minima.
  - The columns of the optimal decoder does not identify the principal directions but only their low dimensional subspace (the so-called *invariance* problem).
- We present a new loss function for LAE:
  - Analytically characterize the loss landscape.
  - All local minima are global minima.
  - The columns of the optimal decoder span the *principal directions*.
  - Invariant local minima become saddle points.
  - Computational complexity is of the same order of MSE loss.

Overview

- Linear Autoencoder (LAE) with Mean Square Error (MSE). The classical results:
  - Loss surface has been analytically characterized.
  - All local minima are global minima.
  - The columns of the optimal decoder does not identify the principal directions but only their low dimensional subspace (the so-called *invariance* problem).
- We present a new loss function for LAE:
  - Analytically characterize the loss landscape.
  - All local minima are global minima.
  - The columns of the optimal decoder span the *principal directions*.
  - Invariant local minima become saddle points.
  - Computational complexity is of the same order of MSE loss.

Overview

- Linear Autoencoder (LAE) with Mean Square Error (MSE). The classical results:
  - Loss surface has been analytically characterized.
  - All local minima are global minima.
  - The columns of the optimal decoder does not identify the principal directions but only their low dimensional subspace (the so-called *invariance* problem).
- We present a new loss function for LAE:
  - Analytically characterize the loss landscape.
  - All local minima are global minima.
  - The columns of the optimal decoder span the *principal directions*.
  - Invariant local minima become saddle points.
  - Computational complexity is of the same order of MSE loss.

Overview

- Linear Autoencoder (LAE) with Mean Square Error (MSE). The classical results:
  - Loss surface has been analytically characterized.
  - All local minima are global minima.
  - The columns of the optimal decoder does not identify the principal directions but only their low dimensional subspace (the so-called *invariance* problem).
- We present a new loss function for LAE:
  - Analytically characterize the loss landscape.
  - All local minima are global minima.
  - The columns of the optimal decoder span the *principal directions*.
  - Invariant local minima become saddle points.
  - Computational complexity is of the same order of MSE loss.

Overview

- Linear Autoencoder (LAE) with Mean Square Error (MSE). The classical results:
  - Loss surface has been analytically characterized.
  - All local minima are global minima.
  - The columns of the optimal decoder does not identify the principal directions but only their low dimensional subspace (the so-called *invariance* problem).
- We present a new loss function for LAE:
  - Analytically characterize the loss landscape.
  - All local minima are global minima.
  - The columns of the optimal decoder span the *principal directions*.
  - Invariant local minima become saddle points.
  - Computational complexity is of the same order of MSE loss.

Overview

- Linear Autoencoder (LAE) with Mean Square Error (MSE). The classical results:
  - Loss surface has been analytically characterized.
  - All local minima are global minima.
  - The columns of the optimal decoder does not identify the principal directions but only their low dimensional subspace (the so-called *invariance* problem).
- We present a new loss function for LAE:
  - Analytically characterize the loss landscape.
  - All local minima are global minima.
  - The columns of the optimal decoder span the *principal directions*.
  - Invariant local minima become saddle points.
  - Computational complexity is of the same order of MSE loss.

Overview

- Linear Autoencoder (LAE) with Mean Square Error (MSE). The classical results:
  - Loss surface has been analytically characterized.
  - All local minima are global minima.
  - The columns of the optimal decoder does not identify the principal directions but only their low dimensional subspace (the so-called *invariance* problem).

- We present a new loss function for LAE:
  - Analytically characterize the loss landscape.
  - All local minima are global minima.
  - The columns of the optimal decoder span the *principal directions*.
  - Invariant local minima become saddle points.
  - Computational complexity is of the same order of MSE loss.

Overview

- Linear Autoencoder (LAE) with Mean Square Error (MSE). The classical results:
  - Loss surface has been analytically characterized.
  - All local minima are global minima.
  - The columns of the optimal decoder does not identify the principal directions but only their low dimensional subspace (the so-called *invariance* problem).
- We present a new loss function for LAE:
  - Analytically characterize the loss landscape.
  - All local minima are global minima.
  - The columns of the optimal decoder span the *principal directions*.
  - Invariant local minima become saddle points.
  - Computational complexity is of the same order of MSE loss.

Overview

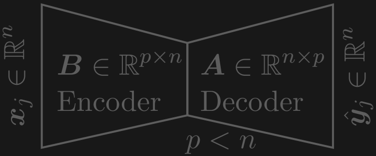
- Linear Autoencoder (LAE) with Mean Square Error (MSE). The classical results:
  - Loss surface has been analytically characterized.
  - All local minima are global minima.
  - The columns of the optimal decoder does not identify the principal directions but only their low dimensional subspace (the so-called *invariance* problem).
- We present a new loss function for LAE:
  - Analytically characterize the loss landscape.
  - All local minima are global minima.
  - The columns of the optimal decoder span the *principal directions*.
  - Invariant local minima become saddle points.
  - Computational complexity is of the same order of MSE loss.

Overview

- Linear Autoencoder (LAE) with Mean Square Error (MSE). The classical results:
  - Loss surface has been analytically characterized.
  - All local minima are global minima.
  - The columns of the optimal decoder does not identify the principal directions but only their low dimensional subspace (the so-called *invariance* problem).
- We present a new loss function for LAE:
  - Analytically characterize the loss landscape.
  - All local minima are global minima.
  - The columns of the optimal decoder span the *principal directions*.
  - Invariant local minima become saddle points.
  - Computational complexity is of the same order of MSE loss.

Setup

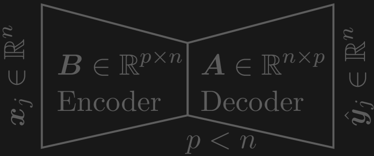
- Data: m sample points of dimension n:
  - Input: $\boldsymbol{x}_j \in \mathbb{R}^n$, Output: $\boldsymbol{y}_j \in \mathbb{R}^n$ for $j = 1, \ldots, m$.
  - In matrix form: $\boldsymbol{X} \in \mathbb{R}^{n \times m}$, $\boldsymbol{Y} \in \mathbb{R}^{n \times m}$.
- LAE: A neural network with linear activation functions and single hidden layer of width $p < n$.



  - The weights: The encoder matrix $\boldsymbol{B}$, and the decoder matrix $\boldsymbol{A}$.
  - The global map is $\hat{\boldsymbol{y}}_j = \boldsymbol{A}\boldsymbol{B}\boldsymbol{x}_j$ or $\hat{\boldsymbol{Y}} = \boldsymbol{A}\boldsymbol{B}\boldsymbol{X}$.

Setup

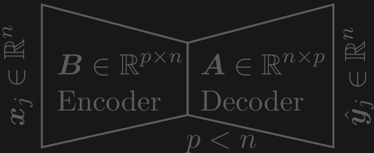
- Data: m sample points of dimension n:
  - Input: $\boldsymbol{x}_j \in \mathbb{R}^n$, Output: $\boldsymbol{y}_j \in \mathbb{R}^n$ for $j = 1, \ldots, m$.
  - In matrix form: $\boldsymbol{X} \in \mathbb{R}^{n \times m}$, $\boldsymbol{Y} \in \mathbb{R}^{n \times m}$.
- LAE: A neural network with linear activation functions and single hidden layer of width $p < n$.



  - The weights: The encoder matrix $\boldsymbol{B}$, and the decoder matrix $\boldsymbol{A}$.
  - The global map is $\hat{\boldsymbol{y}}_j = \boldsymbol{A}\boldsymbol{B}\boldsymbol{x}_j$ or $\hat{\boldsymbol{Y}} = \boldsymbol{A}\boldsymbol{B}\boldsymbol{X}$.

Setup

- Data: m sample points of dimension n:
  - Input: $\boldsymbol{x}_j \in \mathbb{R}^n$, Output: $\boldsymbol{y}_j \in \mathbb{R}^n$ for $j = 1, \ldots, m$.
  - In matrix form: $\boldsymbol{X} \in \mathbb{R}^{n \times m}$, $\boldsymbol{Y} \in \mathbb{R}^{n \times m}$.
- LAE: A neural network with linear activation functions and single hidden layer of width $p < n$.

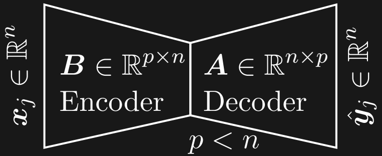


  - The weights: The encoder matrix $\boldsymbol{B}$, and the decoder matrix $\boldsymbol{A}$.
  - The global map is $\hat{\boldsymbol{y}}_j = \boldsymbol{A}\boldsymbol{B}\boldsymbol{x}_j$ or $\hat{\boldsymbol{Y}} = \boldsymbol{A}\boldsymbol{B}\boldsymbol{X}$.

Setup

- Data: m sample points of dimension n:
  - Input: $\boldsymbol{x}_j \in \mathbb{R}^n$, Output: $\boldsymbol{y}_j \in \mathbb{R}^n$ for $j = 1, \ldots, m$.
  - In matrix form: $\boldsymbol{X} \in \mathbb{R}^{n \times m}$, $\boldsymbol{Y} \in \mathbb{R}^{n \times m}$.

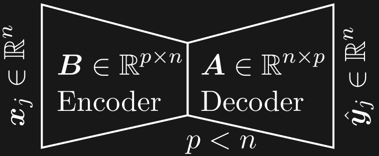
- LAE: A neural network with linear activation functions and single hidden layer of width $p < n$.



  - The weights: The encoder matrix $\boldsymbol{B}$, and the decoder matrix $\boldsymbol{A}$.
  - The global map is $\hat{\boldsymbol{y}}_j = \boldsymbol{A}\boldsymbol{B}\boldsymbol{x}_j$ or $\hat{\boldsymbol{Y}} = \boldsymbol{A}\boldsymbol{B}\boldsymbol{X}$.

Setup

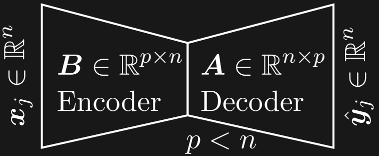
- Data: m sample points of dimension n:
  - Input: $\boldsymbol{x}_j \in \mathbb{R}^n$, Output: $\boldsymbol{y}_j \in \mathbb{R}^n$ for $j = 1, \ldots, m$.
  - In matrix form: $\boldsymbol{X} \in \mathbb{R}^{n \times m}$, $\boldsymbol{Y} \in \mathbb{R}^{n \times m}$.
- LAE: A neural network with linear activation functions and single hidden layer of width $p < n$.



  - The weights: The encoder matrix $\boldsymbol{B}$, and the decoder matrix $\boldsymbol{A}$.
  - The global map is $\hat{\boldsymbol{y}}_j = \boldsymbol{A}\boldsymbol{B}\boldsymbol{x}_j$ or $\hat{\boldsymbol{Y}} = \boldsymbol{A}\boldsymbol{B}\boldsymbol{X}$.

Setup

- Data: m sample points of dimension n:
  - Input: $\boldsymbol{x}_j \in \mathbb{R}^n$, Output: $\boldsymbol{y}_j \in \mathbb{R}^n$ for $j = 1, \ldots, m$.
  - In matrix form: $\boldsymbol{X} \in \mathbb{R}^{n\times m}$, $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$.
- LAE: A neural network with linear activation functions and single hidden layer of width $p < n$.



  - The weights: The encoder matrix $\boldsymbol{B}$, and the decoder matrix $\boldsymbol{A}$.
  - The global map is $\hat{\boldsymbol{y}}_j = \boldsymbol{A}\boldsymbol{B}\boldsymbol{x}_j$ or $\hat{\boldsymbol{Y}} = \boldsymbol{A}\boldsymbol{B}\boldsymbol{X}$.

The loss functions

- The MSE loss: $\tilde{L}(\boldsymbol{A}, \boldsymbol{B}) \coloneqq \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{B}\boldsymbol{X}\|_F^2$.
  - If $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ is a local minimum of $\tilde{L}$ then for any invertible $\boldsymbol{C} \in \mathbb{R}^{p \times p}$, $(\boldsymbol{A}^*\boldsymbol{C}, \boldsymbol{C}^{-1}\boldsymbol{B}^*)$ is another local minima:

$$\tilde{L}(\boldsymbol{A}^*\boldsymbol{C}, \boldsymbol{C}^{-1}\boldsymbol{B}^*) = \left\|\boldsymbol{Y} - \boldsymbol{A}^*\boldsymbol{C}\boldsymbol{C}^{-1}\boldsymbol{B}^*\boldsymbol{X}\right\|_F^2 = \tilde{L}(\boldsymbol{A}^*, \boldsymbol{B}^*).$$

- The proposed loss: $L(\boldsymbol{A}, \boldsymbol{B}) \coloneqq \sum_{i=1}^{p} \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\|_F^2$, where, $\boldsymbol{I}_{i;p} = \mathrm{diag}(\underbrace{1, \cdots, 1}_{i}, 0, \cdots, 0) \in \mathbb{R}^{p \times p}$.

The loss functions

- The MSE loss: $\tilde{L}(\boldsymbol{A}, \boldsymbol{B}) := \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{B}\boldsymbol{X}\|_F^2$.
  - If $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ is a local minimum of $\tilde{L}$ then for any invertible $\boldsymbol{C} \in \mathbb{R}^{p \times p}$, $(\boldsymbol{A}^*\boldsymbol{C}, \boldsymbol{C}^{-1}\boldsymbol{B}^*)$ is another local minima:

$$\tilde{L}(\boldsymbol{A}^*\boldsymbol{C}, \boldsymbol{C}^{-1}\boldsymbol{B}^*) = \left\|\boldsymbol{Y} - \boldsymbol{A}^*\boldsymbol{C}\boldsymbol{C}^{-1}\boldsymbol{B}^*\boldsymbol{X}\right\|_F^2 = \tilde{L}(\boldsymbol{A}^*, \boldsymbol{B}^*).$$

- The proposed loss: $L(\boldsymbol{A}, \boldsymbol{B}) := \sum_{i=1}^{p} \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\|_F^2$, where, $\boldsymbol{I}_{i;p} = \mathrm{diag}(\underbrace{1, \cdots, 1}_{i}, 0, \cdots, 0) \in \mathbb{R}^{p \times p}$.

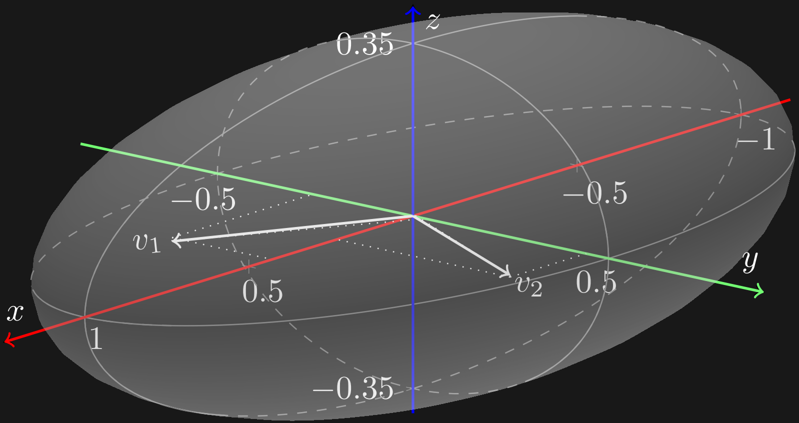
The loss functions

- The MSE loss: $\tilde{L}(\boldsymbol{A}, \boldsymbol{B}) := \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{B}\boldsymbol{X}\|_F^2$.
  - If $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ is a local minimum of $\tilde{L}$ then for any invertible $\boldsymbol{C} \in \mathbb{R}^{p \times p}$, $(\boldsymbol{A}^*\boldsymbol{C}, \boldsymbol{C}^{-1}\boldsymbol{B}^*)$ is another local minima:

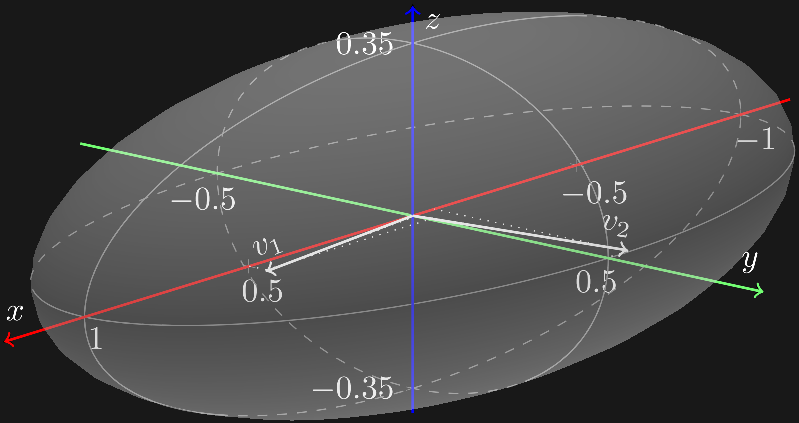
$$\tilde{L}(\boldsymbol{A}^*\boldsymbol{C}, \boldsymbol{C}^{-1}\boldsymbol{B}^*) = \left\|\boldsymbol{Y} - \boldsymbol{A}^*\boldsymbol{C}\boldsymbol{C}^{-1}\boldsymbol{B}^*\boldsymbol{X}\right\|_F^2 = \tilde{L}(\boldsymbol{A}^*, \boldsymbol{B}^*).$$

- The proposed loss: $L(\boldsymbol{A}, \boldsymbol{B}) := \sum_{i=1}^p \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\|_F^2$, where, $\boldsymbol{I}_{i;p} = \mathrm{diag}(\underbrace{1, \cdots, 1}_{i}, 0, \cdots, 0) \in \mathbb{R}^{p \times p}$.

A Visualization: MSE Loss



A Visualization: Proposed Loss



The loss functions

- The MSE loss: $\tilde{L}(\boldsymbol{A}, \boldsymbol{B}) := \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{B}\boldsymbol{X}\|_F^2$.
  - If $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ is a local minimum of $\tilde{L}$ then for any invertible $\boldsymbol{C} \in \mathbb{R}^{p \times p}$, $(\boldsymbol{A}^*\boldsymbol{C}, \boldsymbol{C}^{-1}\boldsymbol{B}^*)$ is another local minima:

$$\tilde{L}(\boldsymbol{A}^*\boldsymbol{C}, \boldsymbol{C}^{-1}\boldsymbol{B}^*) = \|\boldsymbol{Y} - \boldsymbol{A}^*\boldsymbol{C}\boldsymbol{C}^{-1}\boldsymbol{B}^*\boldsymbol{X}\|_F^2 = \tilde{L}(\boldsymbol{A}^*, \boldsymbol{B}^*).$$

- The proposed loss: $L(\boldsymbol{A}, \boldsymbol{B}) := \sum_{i=1}^p \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\|_F^2$, where, $\boldsymbol{I}_{i;p} = \mathrm{diag}(\underbrace{1, \cdots, 1}_{i}, 0, \cdots, 0) \in \mathbb{R}^{p \times p}$.
  - Intuition: (Sequential) As an example look at $p = 3$, where

$$\boldsymbol{I}_{1;3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \boldsymbol{I}_{2;3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \boldsymbol{I}_{3;3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

  - But does this work simultaneously? And is it computationally feasible (p can be large)?
  - Well, *it does and it is!* But before getting into details let's discuss some implications ...

The loss functions

- The MSE loss: $\tilde{L}(\boldsymbol{A}, \boldsymbol{B}) := \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{B}\boldsymbol{X}\|_F^2$.
  - If $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ is a local minimum of $\tilde{L}$ then for any invertible $\boldsymbol{C} \in \mathbb{R}^{p \times p}$, $(\boldsymbol{A}^*\boldsymbol{C}, \boldsymbol{C}^{-1}\boldsymbol{B}^*)$ is another local minima:

$$\tilde{L}(\boldsymbol{A}^*\boldsymbol{C}, \boldsymbol{C}^{-1}\boldsymbol{B}^*) = \left\|\boldsymbol{Y} - \boldsymbol{A}^*\boldsymbol{C}\boldsymbol{C}^{-1}\boldsymbol{B}^*\boldsymbol{X}\right\|_F^2 = \tilde{L}(\boldsymbol{A}^*, \boldsymbol{B}^*).$$

- The proposed loss: $L(\boldsymbol{A}, \boldsymbol{B}) := \sum_{i=1}^{p} \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\|_F^2$, where, $\boldsymbol{I}_{i;p} = \mathrm{diag}(\underbrace{1, \cdots, 1}_{i}, 0, \cdots, 0) \in \mathbb{R}^{p \times p}$.
  - Intuition: (Sequential) As an example look at $p = 3$, where

$$\boldsymbol{I}_{1;3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \boldsymbol{I}_{2;3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \boldsymbol{I}_{3;3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

  - But does this work simultaneously? And is it computationally feasible (p can be large)?
  - Well, *it does and it is!* But before getting into details let's discuss some implications ...

The loss functions

- The MSE loss: $\tilde{L}(\boldsymbol{A}, \boldsymbol{B}) := \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{B}\boldsymbol{X}\|_F^2$.
  - If $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ is a local minimum of $\tilde{L}$ then for any invertible $\boldsymbol{C} \in \mathbb{R}^{p\times p}$, $(\boldsymbol{A}^*\boldsymbol{C}, \boldsymbol{C}^{-1}\boldsymbol{B}^*)$ is another local minima:

$$\tilde{L}(\boldsymbol{A}^*\boldsymbol{C}, \boldsymbol{C}^{-1}\boldsymbol{B}^*) = \left\|\boldsymbol{Y} - \boldsymbol{A}^*\boldsymbol{C}\boldsymbol{C}^{-1}\boldsymbol{B}^*\boldsymbol{X}\right\|_F^2 = \tilde{L}(\boldsymbol{A}^*, \boldsymbol{B}^*).$$

- The proposed loss: $L(\boldsymbol{A}, \boldsymbol{B}) := \sum_{i=1}^p \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\|_F^2$, where, $\boldsymbol{I}_{i;p} = \mathrm{diag}(\underbrace{1, \cdots, 1}_{i}, 0, \cdots, 0) \in \mathbb{R}^{p\times p}$.
  - Intuition: (Sequential) As an example look at $p = 3$, where

$$\boldsymbol{I}_{1;3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \boldsymbol{I}_{2;3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \boldsymbol{I}_{3;3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

  - But does this work simultaneously? And is it computationally feasible (p can be large)?
  - Well, *it does and it is!* But before getting into details let's discuss some implications ...

Implications

- Let $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ be the local minimum of MSE loss, where the columns of $\boldsymbol{A}^*$ are the largest eigenvectors of the sample covariance matrix, then for any invertible $\boldsymbol{C} \in \mathbb{R}^{p \times p}$, $(\boldsymbol{A}^*\boldsymbol{C}, \boldsymbol{C}^{-1}\boldsymbol{B}^*)$ is another local minima.
  - Numerically, on the same dataset different runs with different initializations lead to different optimal points.
  - Almost surely none will represent the principal directions.
- *The only local minimum of the loss L is $(\boldsymbol{A}^*, \boldsymbol{B}^*)$, up to the normalization of the columns.*
  - The loss L enables low rank decomposition as a single optimization block that can be incorporated as part of a larger pipeline.
  - Potentially enabling LAEs to compete with other approaches for low rank decomposition.

Critical Points

- The critical point equations of $\tilde{L}$ and L.

For $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$:	For $L(\boldsymbol{A}, \boldsymbol{B})$:
$\boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx} = \boldsymbol{A}'\boldsymbol{\Sigma}_{yx},$	$(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))\boldsymbol{B}\boldsymbol{\Sigma}_{xx} = \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx},$
$\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}' = \boldsymbol{\Sigma}_{yx}\boldsymbol{B}',$	$\boldsymbol{A}(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')) = \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p,$

where,

- $\boldsymbol{A}'$ is the transpose of $\boldsymbol{A}$.
- $\boldsymbol{\Sigma}_{xx} = XX'$, $\boldsymbol{\Sigma}_{yx} = YX'$ are covariance matrices.
- $\circ$ is the (element-wise) Hadamard product, and
- $\boldsymbol{T}_p$, and $\boldsymbol{S}_p$ are

$$\boldsymbol{T}_p = \operatorname{diag}(p, p-1, \cdots, 1),$$

$$\boldsymbol{S}_p = \begin{bmatrix} p & p-1 & \cdots & 1 \\ p-1 & p-1 & \cdots & 1 \\ \vdots & \vdots & \ddots & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \text{ e.g. } S_4 = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 3 & 2 & 1 \\ 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Critical Points

- The critical point equations of $\tilde{L}$ and L.

$$
\begin{array}{c|c}
\text{For } \tilde{L}(\boldsymbol{A}, \boldsymbol{B}): & \text{For } L(\boldsymbol{A}, \boldsymbol{B}): \\
\boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx} = \boldsymbol{A}'\boldsymbol{\Sigma}_{yx}, & (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))\boldsymbol{B}\boldsymbol{\Sigma}_{xx} = \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx}, \\
\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}' = \boldsymbol{\Sigma}_{yx}\boldsymbol{B}', & \boldsymbol{A}(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')) = \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p,
\end{array}
$$

  where,

  - $\boldsymbol{A}'$ is the transpose of $\boldsymbol{A}$.
  - $\boldsymbol{\Sigma}_{xx} = XX'$, $\boldsymbol{\Sigma}_{yx} = YX'$ are covariance matrices.
  - $\circ$ is the (element-wise) Hadamard product, and
  - $\boldsymbol{T}_p$, and $\boldsymbol{S}_p$ are

$$
\boldsymbol{T}_p = \operatorname{diag}(p, p-1, \cdots, 1),
$$

$$
\boldsymbol{S}_p = \begin{bmatrix} p & p-1 & \cdots & 1 \\ p-1 & p-1 & \cdots & 1 \\ \vdots & \vdots & \ddots & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \text{ e.g. } S_4 = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 3 & 2 & 1 \\ 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.
$$

Critical Points

- The critical point equations of $\tilde{L}$ and L.

$$
\begin{array}{c|c}
\text{For } \tilde{L}(\boldsymbol{A},\boldsymbol{B}): & \text{For } L(\boldsymbol{A},\boldsymbol{B}): \\
\begin{aligned} \boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx} &= \boldsymbol{A}'\boldsymbol{\Sigma}_{yx}, \\ \boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}' &= \boldsymbol{\Sigma}_{yx}\boldsymbol{B}', \end{aligned} & \begin{aligned} (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))\boldsymbol{B}\boldsymbol{\Sigma}_{xx} &= \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx}, \\ \boldsymbol{A}(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')) &= \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p, \end{aligned}
\end{array}
$$

where,

  - $\boldsymbol{A}'$ is the transpose of $\boldsymbol{A}$.
  - $\boldsymbol{\Sigma}_{xx} = XX'$, $\boldsymbol{\Sigma}_{yx} = YX'$ are covariance matrices.
  - $\circ$ is the (element-wise) Hadamard product, and
  - $\boldsymbol{T}_p$, and $\boldsymbol{S}_p$ are

$$
\boldsymbol{T}_p = \operatorname{diag}(p, p-1, \cdots, 1),
$$

$$
\boldsymbol{S}_p = \begin{bmatrix} p & p-1 & \cdots & 1 \\ p-1 & p-1 & \cdots & 1 \\ \vdots & \vdots & \ddots & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad \text{e.g. } S_4 = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 3 & 2 & 1 \\ 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.
$$

Critical Points

- The critical point equations of $\tilde{L}$ and L.

For $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$:	For $L(\boldsymbol{A}, \boldsymbol{B})$:
$\boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx} = \boldsymbol{A}'\boldsymbol{\Sigma}_{yx},$	$(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))\boldsymbol{B}\boldsymbol{\Sigma}_{xx} = \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx},$
$\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}' = \boldsymbol{\Sigma}_{yx}\boldsymbol{B}',$	$\boldsymbol{A}(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')) = \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p,$

  where,

  - $\boldsymbol{A}'$ is the transpose of $\boldsymbol{A}$.
  - $\boldsymbol{\Sigma}_{xx} = XX'$, $\boldsymbol{\Sigma}_{yx} = YX'$ are covariance matrices.
  - $\circ$ is the (element-wise) Hadamard product, and
  - $\boldsymbol{T}_p$, and $\boldsymbol{S}_p$ are

$$\boldsymbol{T}_p = \operatorname{diag}(p, p-1, \cdots, 1),$$

$$\boldsymbol{S}_p = \begin{bmatrix} p & p-1 & \cdots & 1 \\ p-1 & p-1 & \cdots & 1 \\ \vdots & \vdots & \ddots & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \text{ e.g. } S_4 = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 3 & 2 & 1 \\ 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Critical Points

- The critical point equations of $\tilde{L}$ and L.

$$
\begin{array}{rcl|rcl}
\multicolumn{3}{c|}{\text{For } \tilde{L}(\boldsymbol{A},\boldsymbol{B}):} & \multicolumn{3}{c}{\text{For } L(\boldsymbol{A},\boldsymbol{B}):} \\
\boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx} & = & \boldsymbol{A}'\boldsymbol{\Sigma}_{yx}, & (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))\boldsymbol{B}\boldsymbol{\Sigma}_{xx} & = & \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx}, \\
\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}' & = & \boldsymbol{\Sigma}_{yx}\boldsymbol{B}', & \boldsymbol{A}(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')) & = & \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p,
\end{array}
$$

where,

  - $\boldsymbol{A}'$ is the transpose of $\boldsymbol{A}$.
  - $\boldsymbol{\Sigma}_{xx} = XX'$, $\boldsymbol{\Sigma}_{yx} = YX'$ are covariance matrices.
  - $\circ$ is the (element-wise) Hadamard product, and
  - $\boldsymbol{T}_p$, and $\boldsymbol{S}_p$ are

$$
\boldsymbol{T}_p = \operatorname{diag}(p, p-1, \cdots, 1),
$$

$$
\boldsymbol{S}_p = \begin{bmatrix} p & p-1 & \cdots & 1 \\ p-1 & p-1 & \cdots & 1 \\ \vdots & \vdots & \ddots & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \text{ e.g. } S_4 = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 3 & 2 & 1 \\ 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.
$$

Results I

- Every critical point of $L(\boldsymbol{A}, \boldsymbol{B})$ is a critical point of $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$, but not the other way around.
- Local minima of L, and $\tilde{L}$:

For $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$:	For $L(\boldsymbol{A}, \boldsymbol{B})$:
$\boldsymbol{A}^* = \boldsymbol{U}_{1:p}\boldsymbol{C}_p,$	$\boldsymbol{A}^* = \boldsymbol{U}_{1:p}\boldsymbol{D}_p,$
$\boldsymbol{B}^* = \boldsymbol{C}_p^{-1}\boldsymbol{U}_{1:p}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1},$	$\boldsymbol{B}^* = \boldsymbol{D}_p^{-1}\boldsymbol{U}_{1:p}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1},$

  - The i^{th} column of $\boldsymbol{U}_{1:p}$ is a unit eigenvector of $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ corresponding the i^{th} largest eigenvalue.
  - $\boldsymbol{D}_p$ is a diagonal matrix with nonzero diagonal elements, and $\boldsymbol{C}_p \in \mathrm{GL}_p(\mathbb{R})$.
- The characterization of the loss landscape:
  - The structure of full rank saddle points.
  - The structure of low rank saddle points (rather involved!).

Results I

- Every critical point of $L(\boldsymbol{A}, \boldsymbol{B})$ is a critical point of $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$, but not the other way around.

- Local minima of L, and $\tilde{L}$:

For $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$:	For $L(\boldsymbol{A}, \boldsymbol{B})$:
$\boldsymbol{A}^* = \boldsymbol{U}_{1:p}\boldsymbol{C}_p,$	$\boldsymbol{A}^* = \boldsymbol{U}_{1:p}\boldsymbol{D}_p,$
$\boldsymbol{B}^* = \boldsymbol{C}_p^{-1}\boldsymbol{U}_{1:p}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1},$	$\boldsymbol{B}^* = \boldsymbol{D}_p^{-1}\boldsymbol{U}_{1:p}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1},$

  - The i^{th} column of $\boldsymbol{U}_{1:p}$ is a unit eigenvector of $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ corresponding the i^{th} largest eigenvalue.
  - $\boldsymbol{D}_p$ is a diagonal matrix with nonzero diagonal elements, and $\boldsymbol{C}_p \in \mathrm{GL}_p(\mathbb{R})$.

- The characterization of the loss landscape:
  - The structure of full rank saddle points.
  - The structure of low rank saddle points (rather involved!).

Results I

- Every critical point of $L(\boldsymbol{A}, \boldsymbol{B})$ is a critical point of $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$, but not the other way around.
- Local minima of L, and $\tilde{L}$:

For $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$:	For $L(\boldsymbol{A}, \boldsymbol{B})$:
$\boldsymbol{A}^* = \boldsymbol{U}_{1:p}\boldsymbol{C}_p,$	$\boldsymbol{A}^* = \boldsymbol{U}_{1:p}\boldsymbol{D}_p,$
$\boldsymbol{B}^* = \boldsymbol{C}_p^{-1}\boldsymbol{U}_{1:p}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1},$	$\boldsymbol{B}^* = \boldsymbol{D}_p^{-1}\boldsymbol{U}_{1:p}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1},$

  - The i^{th} column of $\boldsymbol{U}_{1:p}$ is a unit eigenvector of $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ corresponding the i^{th} largest eigenvalue.
  - $\boldsymbol{D}_p$ is a diagonal matrix with nonzero diagonal elements, and $\boldsymbol{C}_p \in \mathrm{GL}_p(\mathbb{R})$.
- The characterization of the loss landscape:
  - The structure of full rank saddle points.
  - The structure of low rank saddle points (rather involved!).

Results I

- Every critical point of $L(\boldsymbol{A}, \boldsymbol{B})$ is a critical point of $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$, but not the other way around.

- Local minima of L, and $\tilde{L}$:

For $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$:	For $L(\boldsymbol{A}, \boldsymbol{B})$:
$\boldsymbol{A}^* = \boldsymbol{U}_{1:p}\boldsymbol{C}_p,$	$\boldsymbol{A}^* = \boldsymbol{U}_{1:p}\boldsymbol{D}_p,$
$\boldsymbol{B}^* = \boldsymbol{C}_p^{-1}\boldsymbol{U}_{1:p}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1},$	$\boldsymbol{B}^* = \boldsymbol{D}_p^{-1}\boldsymbol{U}_{1:p}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1},$

  - The i^{th} column of $\boldsymbol{U}_{1:p}$ is a unit eigenvector of $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ corresponding the i^{th} largest eigenvalue.
  - $\boldsymbol{D}_p$ is a diagonal matrix with nonzero diagonal elements, and $\boldsymbol{C}_p \in \mathrm{GL}_p(\mathbb{R})$.

- The characterization of the loss landscape:

  - The structure of full rank saddle points.
  - The structure of low rank saddle points (rather involved!).

Results I

- Every critical point of $L(\boldsymbol{A}, \boldsymbol{B})$ is a critical point of $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$, but not the other way around.
- Local minima of L, and $\tilde{L}$:

For $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$:	For $L(\boldsymbol{A}, \boldsymbol{B})$:
$\boldsymbol{A}^* = \boldsymbol{U}_{1:p}\boldsymbol{C}_p,$ $\boldsymbol{B}^* = \boldsymbol{C}_p^{-1}\boldsymbol{U}'_{1:p}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1},$	$\boldsymbol{A}^* = \boldsymbol{U}_{1:p}\boldsymbol{D}_p,$ $\boldsymbol{B}^* = \boldsymbol{D}_p^{-1}\boldsymbol{U}'_{1:p}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1},$

  - The i^{th} column of $\boldsymbol{U}_{1:p}$ is a unit eigenvector of $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ corresponding the i^{th} largest eigenvalue.
  - $\boldsymbol{D}_p$ is a diagonal matrix with nonzero diagonal elements, and $\boldsymbol{C}_p \in \mathrm{GL}_p(\mathbb{R})$.
- The characterization of the loss landscape:
  - The structure of full rank saddle points.
  - The structure of low rank saddle points (rather involved!).

Results II

- The MSE loss $\tilde{L}$ and our loss L can be written as

$$
\begin{aligned}
\tilde{L}(\boldsymbol{A}, \boldsymbol{B}) &= p \operatorname{Tr}(\boldsymbol{\Sigma}_{yy}) - 2 \operatorname{Tr}\left(\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xy}\right) + \operatorname{Tr}\left(\boldsymbol{B}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\right), \\
L(\boldsymbol{A}, \boldsymbol{B}) &= p \operatorname{Tr}(\boldsymbol{\Sigma}_{yy}) - 2 \operatorname{Tr}\left(\boldsymbol{A}\boldsymbol{T}_p\boldsymbol{B}\boldsymbol{\Sigma}_{xy}\right) \\
&\qquad + \operatorname{Tr}\left(\boldsymbol{B}'\left(\boldsymbol{S}_p \circ \left(\boldsymbol{A}'\boldsymbol{A}\right)\right)\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\right).
\end{aligned}
$$

- The analytical gradients are:

$$
\begin{aligned}
d_{\boldsymbol{B}}\tilde{L}(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{W} &= -2\langle \boldsymbol{A}'\boldsymbol{\Sigma}_{yx} - \boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx}, \boldsymbol{W}\rangle_F, \\
d_{\boldsymbol{B}}L(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{W} &= -2\langle \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx} - \left(\boldsymbol{S}_p \circ \left(\boldsymbol{A}'\boldsymbol{A}\right)\right)\boldsymbol{B}\boldsymbol{\Sigma}_{xx}, \boldsymbol{W}\rangle_F,
\end{aligned}
$$

in direction of $\boldsymbol{W} \in \mathbb{R}^{p \times n}$. The gradient for $\boldsymbol{A}$ is similar.

- Finally, since the loss function is explicitly provided, any optimization method that works with MSE loss is usable with the proposed loss.

Results II

- The MSE loss $\tilde{L}$ and our loss L can be written as

$$\tilde{L}(\boldsymbol{A}, \boldsymbol{B}) = p\,\mathrm{Tr}(\boldsymbol{\Sigma}_{yy}) - 2\,\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xy}\right) + \mathrm{Tr}\left(\boldsymbol{B}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\right),$$

$$L(\boldsymbol{A}, \boldsymbol{B}) = p\,\mathrm{Tr}(\boldsymbol{\Sigma}_{yy}) - 2\,\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{T}_p\boldsymbol{B}\boldsymbol{\Sigma}_{xy}\right) + \mathrm{Tr}\left(\boldsymbol{B}'\left(\boldsymbol{S}_p \circ \left(\boldsymbol{A}'\boldsymbol{A}\right)\right)\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\right).$$

- The analytical gradients are:

$$d_{\boldsymbol{B}}\tilde{L}(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{W} = -2\langle \boldsymbol{A}'\boldsymbol{\Sigma}_{yx} - \boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx}, \boldsymbol{W}\rangle_F,$$

$$d_{\boldsymbol{B}}L(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{W} = -2\langle \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx} - \left(\boldsymbol{S}_p \circ \left(\boldsymbol{A}'\boldsymbol{A}\right)\right)\boldsymbol{B}\boldsymbol{\Sigma}_{xx}, \boldsymbol{W}\rangle_F,$$

  in direction of $\boldsymbol{W} \in \mathbb{R}^{p \times n}$. The gradient for $\boldsymbol{A}$ is similar.

- Finally, since the loss function is explicitly provided, any optimization method that works with MSE loss is usable with the proposed loss.

Results II

- The MSE loss $\tilde{L}$ and our loss L can be written as

$$\tilde{L}(\boldsymbol{A},\boldsymbol{B}) = p\,\mathrm{Tr}(\boldsymbol{\Sigma}_{yy}) - 2\,\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xy}\right) + \mathrm{Tr}\left(\boldsymbol{B}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\right),$$

$$L(\boldsymbol{A},\boldsymbol{B}) = p\,\mathrm{Tr}(\boldsymbol{\Sigma}_{yy}) - 2\,\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{T}_p\boldsymbol{B}\boldsymbol{\Sigma}_{xy}\right) + \mathrm{Tr}\left(\boldsymbol{B}'\left(\boldsymbol{S}_p \circ \left(\boldsymbol{A}'\boldsymbol{A}\right)\right)\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\right).$$

- The analytical gradients are:

$$d_{\boldsymbol{B}}\tilde{L}(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W} = -2\langle \boldsymbol{A}'\boldsymbol{\Sigma}_{yx} - \boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx}, \boldsymbol{W}\rangle_F,$$

$$d_{\boldsymbol{B}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W} = -2\langle \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx} - \left(\boldsymbol{S}_p \circ \left(\boldsymbol{A}'\boldsymbol{A}\right)\right)\boldsymbol{B}\boldsymbol{\Sigma}_{xx}, \boldsymbol{W}\rangle_F,$$

  in direction of $\boldsymbol{W} \in \mathbb{R}^{p\times n}$. The gradient for $\boldsymbol{A}$ is similar.

- Finally, since the loss function is explicitly provided, any optimization method that works with MSE loss is usable with the proposed loss.