Kurtosis-based projection pursuit for matrix-valued data

J. Virta

University of Turku, Finland

KU Leuven, ESAT Seminar
21st of November 2025

Reference

This talk is based on the following article:

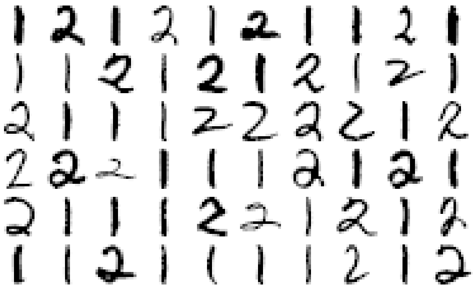
- **Radojičić, U., Nordhausen, K. and Virta, J.** (2025). *Kurtosis-based projection pursuit for matrix-valued data*. Accepted for publication in Annals of Statistics.

These slides are available at the speaker's website
`https://users.utu.fi/jomivi/talks/`

Table of Contents

Matrix data

- Many modern applications produce matrix-valued data $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^{p \times q}$.
- The following data contains $n = 1736$ images of size 16×16 of digits 1 and 2.



Matrix data, continued

- Matrix/image data is typically
  - Structured
  - High-dimensional
- A natural starting point to their analysis is structure-acknowledging dimension reduction.

Projection pursuit

In projection pursuit, one finds a projection $\mathbf{w}'\mathbf{x}$ which maximizes the "interestingness" g of a centered random vector $\mathbf{x} \in \mathbb{R}^p$,

$$\max_{\mathbf{w} \in \mathbb{S}^{p-1}} g(\mathbf{w}'\mathbf{x}).$$

- $g(z) = \mathrm{E}(z^2)$ gives PCA.

Projection pursuit and matrix data

- Projection pursuit can be applied to vectorized image data $\mathrm{vec}(\mathbf{X}_1), \ldots, \mathrm{vec}(\mathbf{X}_n) \in \mathbb{R}^{pq}$.

However, if $pq \geq n-1$ and the data have full rank, a projection corresponding to **any desired point configuration** can be found from the data!

Table of Contents

Our proposal

We "regularize" by projecting $\mathbf{X}$ as $\mathbf{u}'\mathbf{X}\mathbf{v}$ instead of as $\mathbf{w}'\mathrm{vec}(\mathbf{X})$.

Kurtosis-based matrix projection pursuit (MPP)

$$\max_{\mathbf{u}\in\mathbb{S}^{p-1},\mathbf{v}\in\mathbb{S}^{q-1}} \frac{\mathrm{E}\left\{(\mathbf{u}'\mathbf{X}\mathbf{v})^4\right\}}{\left[\mathrm{E}\left\{(\mathbf{u}'\mathbf{X}\mathbf{v})^2\right\}\right]^2}.$$

Since $\mathbf{u}'\mathbf{X}\mathbf{v} = (\mathbf{v}\otimes\mathbf{u})'\mathrm{vec}(\mathbf{X})$, we focus on a specific natural subset of projections.

Sequential solutions

MPP can be applied sequentially to produce solution pairs $(\mathbf{u}_1, \mathbf{v}_1), (\mathbf{u}_2, \mathbf{v}_2), \ldots, (\mathbf{u}_d, \mathbf{v}_d)$ which are required to be mutually orthogonal.

This gives us a vector of projections $(\mathbf{u}_1' \mathbf{X} \mathbf{v}_1, \ldots, \mathbf{u}_d' \mathbf{X} \mathbf{v}_d)$ which can be visualized to reveal hidden structures.

Table of Contents

Mixture model

Matrix normal location mixture

Assume that $\mathbf{X} \in \mathbb{R}^{p \times q}$ is generated as

$$\mathbf{X} \sim \alpha_1 \mathcal{N}_{p \times q}(\mathbf{T}_1, \mathbf{A}, \mathbf{B}) + \alpha_2 \mathcal{N}_{p \times q}(\mathbf{T}_2, \mathbf{A}, \mathbf{B}).$$

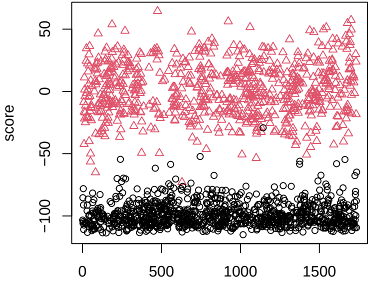
- The optimal likelihood-based classification rule depends on the data through the scores $\mathrm{tr}(\mathbf{W}_{\mathrm{LDA}} \mathbf{X}')$, where

$$\mathbf{W}_{\mathrm{LDA}} := \mathbf{A}^{-1}(\mathbf{T}_2 - \mathbf{T}_1)\mathbf{B}^{-1}.$$

- Given a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and their labels $y_1, \ldots, y_n$, $\mathbf{W}_{\mathrm{LDA}}$ is simple to estimate.

Separation of the digit data

- Given the digit data and their labels, we estimate $\hat{\mathbf{W}}_{\mathrm{LDA}}$ and visualize the scores $\mathrm{tr}(\hat{\mathbf{W}}_{\mathrm{LDA}} \mathbf{X}_i')$.



Result

Theorem 1 in [Radojičić et al., 2025] (simplified)

Let $(\mathbf{u}_1, \mathbf{v}_1), \ldots, (\mathbf{u}_d, \mathbf{v}_d)$ be mutually orthogonal, sequential MPP-solutions. If $|\alpha_1 - 1/2| \neq 1/\sqrt{12}$, then

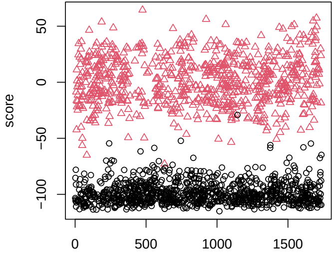
$$w_1 \cdot \mathbf{u}_1' \mathbf{X} \mathbf{v}_1 + \cdots + w_d \cdot \mathbf{u}_d' \mathbf{X} \mathbf{v}_d = \mathrm{tr}(\mathbf{W}_{\mathrm{LDA}} \mathbf{X}'),$$

for specific weights $w_1, \ldots, w_d$.

MPP uses only the data $\mathbf{X}_1, \ldots, \mathbf{X}_n$ but not the labels $(y_1, \ldots, y_n)$!

Separation of the digit data

The scores for the supervised estimator (left) and the unsupervised MPP (right).



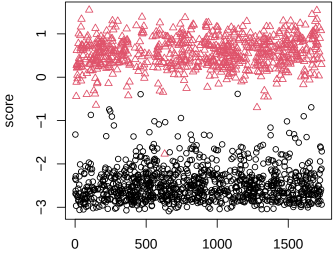
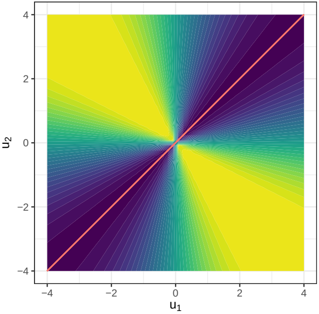
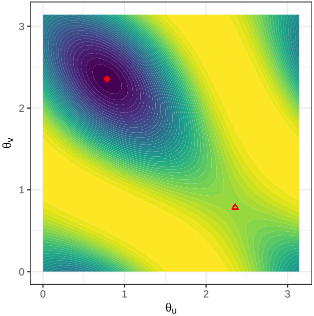


Table of Contents

Optimization

- The objective function surface is non-concave with local optima and saddle points, and we optimize it with informatively initialized ADAM.





Standard asymptotics

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from a matrix normal location mixture and assume the non-zero singular values of $\mathbf{W}_{\mathrm{LDA}}$ to be distinct.

Corollary 4 in [Radojičić et al., 2025]

$$\hat{\mathbf{W}}_{\mathrm{LDA}} \to_{\mathrm{a.s.}} \mathbf{W}_{\mathrm{LDA}}$$

Corollary 5 in [Radojičić et al., 2025]

$$\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{W}}_{\mathrm{LDA}} - \mathbf{W}_{\mathrm{LDA}}) \rightsquigarrow \mathcal{N}_{pq \times pq}(\mathbf{0}, \mathbf{\Theta})$$

High-dimensional asymptotics

Assume that the data dimensions grow, $p_n \to \infty$ and $q_n \to \infty$.

Theorem 9 in [Radojičić et al., 2025]

If $p_n + q_n = o(n^{1/4})$ and if certain technical conditions are satisfied, then

$$\|\mathbf{u}_{1n} - \mathbf{u}_1\| \to_p 0 \quad \text{and} \quad \|\mathbf{v}_{1n} - \mathbf{v}_1\| \to_p 0.$$

Table of Contents

Simulation setting

- We simulate data from the following models with specific choices of parameters.

$$\text{Model 1.2:} \quad \mathbf{X} \sim \alpha_1 \mathcal{N}_{5\times 3}(\mathbf{0}, \mathbf{A}_1, \mathbf{B}_1) + \alpha_2 \mathcal{N}_{5\times 3}(\mathbf{T}_{1.2}, \mathbf{A}_1, \mathbf{B}_1),$$
$$\text{Model 2.2:} \quad \mathbf{X} \sim \alpha_1 \mathcal{N}_{32\times 16}(\mathbf{0}, \mathbf{A}_2, \mathbf{B}_2) + \alpha_2 \mathcal{N}_{32\times 16}(\mathbf{T}_{2.2}, \mathbf{A}_2, \mathbf{B}_2),$$

- We estimate $\mathbf{W}_{\mathrm{LDA}}$ with MPP and NGPP (vectorized projection pursuit) and report median cosine similarities (higher is better) over 100 replicates.

Results

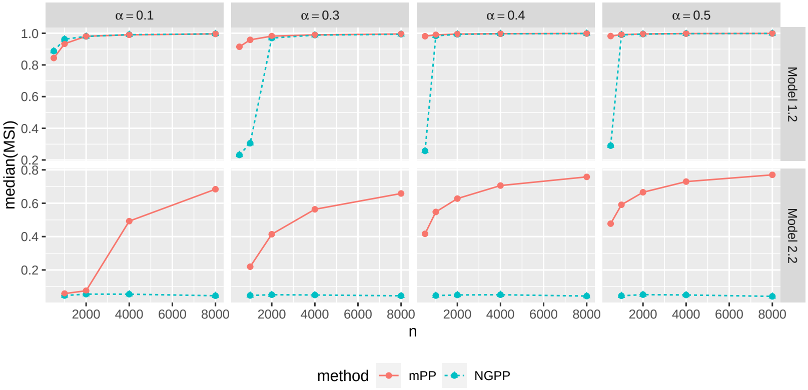

α = 0.1
α = 0.3
α = 0.4
α = 0.5
Model 1.2
Model 2.2
median(MSI)
1.0
0.8
0.6
0.4
0.2
2000
4000
6000
8000
n
method
mPP
NGPP

Table of Contents

Future work

- Skewness instead of kurtosis?
- Ellipticity instead of Gaussianity?
- Multiple groups instead of two?
- Tensors instead of matrices?

Thank you for your attention!

References I

Radojičić, U., Nordhausen, K., and Virta, J. (2025).
Kurtosis-based projection pursuit for matrix-valued data.
Annals of Statistics.