Interactive Dimensionality Reduction for Comparative Analysis

Sungmin Ji

Graduate Journal Discussion

April 6, 2026

Talk outline

Motivation

- Comparative analysis seeks to understand similarities and differences among groups in high-dimensional data.
- We need interpretable low-dimensional embeddings for this task: embedding can preserve information that characterizes each group
- LDA emphasizes group separation, while cPCA emphasizes target-group variance relative to background.
- **ULCA** integrates both goals into one comparative analysis framework.

Principal Component Analysis (PCA)

$$\max_{M^\top M = I_{d'}} \mathrm{tr}(M^\top C M) \tag{1}$$

- Purpose: preserve global variance
- Limitation: does not explicitly prioritize similarity/difference patterns across contrasting groups.

Contrastive PCA (cPCA)

$$\max_{M^\top M = I_{d'}} \operatorname{tr}\big(M^\top (C_{\mathrm{tg}} - \alpha C_{\mathrm{bg}}) M\big) \tag{2}$$

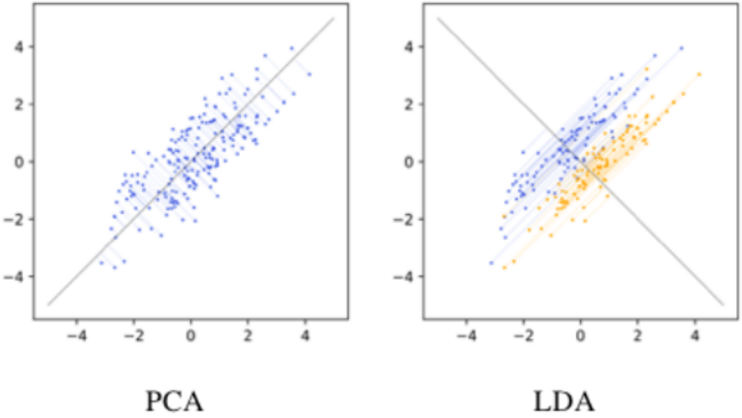
- Purpose: emphasize target variance, suppress background variance
- Strength: good for target-vs-background contrast
- Limitation: solution may not provide a clear interpretablity.

Linear Discriminant Analysis (LDA)

$$\max_{M^\top M = I_{d'}} \frac{\mathrm{tr}(M^\top C_{\mathrm{bw}} M)}{\mathrm{tr}(M^\top C_{\mathrm{wi}} M)} \tag{3}$$

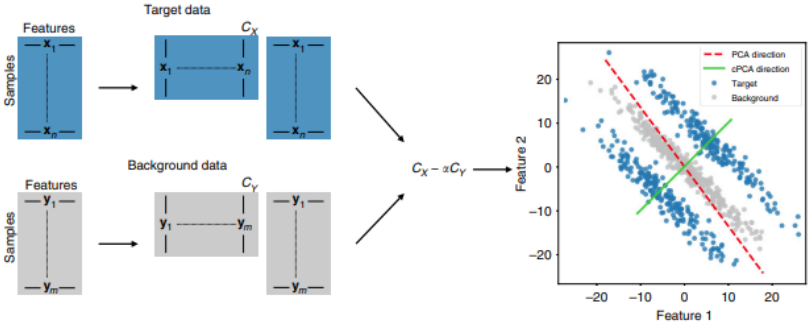
- Purpose: maximize class separation
- Strength: supervised separation
- Limitation: may suppress interesting within-group variation

Comparison of PCA and LDA



Source: Preprocessing Methods and Pipelines of Data Mining: An Overview

cPCA visualization



Source: Exploring patterns enriched in a dataset with contrastive principal component analysis

Generalization of cPCA: Eq. (4)

For a fixed $\alpha \geq 0$,

$$\max_{M^\top M = I_{d'}} \operatorname{tr}\left(M^\top \left(\sum_{j=1}^{c} w_{\mathrm{tg},j} C_{\mathrm{wi},j} - \alpha \sum_{j=1}^{c} w_{\mathrm{bg},j} C_{\mathrm{wi},j} \right) M \right) \tag{4}$$

When α is not fixed,

$$\text{Trace Ratio Form:} \quad \max_{M^\top M = I_{d'}} \frac{\operatorname{tr}\left(M^\top \left(\sum_{j=1}^{c} w_{\mathrm{tg},j} C_{\mathrm{wi},j} \right) M \right)}{\operatorname{tr}\left(M^\top \left(\sum_{j=1}^{c} w_{\mathrm{bg},j} C_{\mathrm{wi},j} \right) M \right)} \tag{5}$$

- $\alpha = 0$ reduces Eq. (4) to PCA
- Extends cPCA from two groups to many groups
- Weights allow partial target/background roles
- Uses *within-class variance*, whereas LDA uses *between-class variance* in its objective.

ULCA: Integration of gcPCA and LDA

Based on the analyst's goal, one can choose weights:

- $w_{\mathrm{tg},j}$: weight for emphasizing *within-class variance* of group j
- $w_{\mathrm{bg},j}$: weight for penalizing *within-class variance* of group j
- $w_{\mathrm{bw},j}$: weight for emphasizing *between-class separation* involving group j

$$\max_{M^\top M = I_{d'}} \frac{\operatorname{tr}(M^\top C_0 M)}{\operatorname{tr}(M^\top C_1 M)}$$

$$C_0 = \sum_{j=1}^{c} w_{\mathrm{tg},j} C_{\mathrm{wi},j} + \sum_{j=1}^{c} w_{\mathrm{bw},j} C_{\mathrm{bw},j} + \gamma_0 I$$

$$C_1 = \sum_{j=1}^{c} w_{\mathrm{bg},j} C_{\mathrm{wi},j} + \gamma_1 I$$

- If $w_{\mathrm{bw}} = \mathbf{0}$, ULCA reduces to a generalized cPCA-type objective, whereas with $w_{\mathrm{tg}} = \mathbf{0}$ and $w_{\mathrm{bg},j} = 1$ for all j, ULCA reduces to an LDA-type objective.
- $\gamma_1 \geq 0, \gamma_0 \geq 0$ are regularization purposes.

Relaxed ULCA: Using a user-specific $\alpha \geq 0$

If $\alpha > 0$ is fixed, ULCA problem can be reduced

$$\max_{M^\top M = I_{d'}} \operatorname{tr}\left(M^\top (C_0 - \alpha C_1) M\right) \tag{6}$$

- Convenient for optimization and interaction
- Increasing α more strongly suppresses background variance in the embedding

ULCA: iteratively updating α and M

$$\alpha_t \leftarrow \frac{\mathrm{tr}(M_t^\top C_0 M_t)}{\mathrm{tr}(M_t^\top C_1 M_t)} \tag{7}$$

$$M_{t+1} \leftarrow \arg \max_{M^\top M = I_{d'}} \mathrm{tr}\big(M^\top (C_0 - \alpha_t C_1) M\big) \tag{8}$$

- Iterative update of α_t
- Solve relaxed problem by eigenvalue decomposition (EVD).

EVD vs manifold optimization

	EVD	Manifold optimization
Idea	eigendecompose $C_0 - \alpha C_1$	direct manifold search
Strength	simple, explicit	flexible, general
Weakness	more tied to relaxed form	more algorithmically complex
Paper's stance	available option	default approach

Parameter selection

- w_{tg}: increase variance of selected groups
- w_{bg}: decrease variance of selected groups
- w_{bw}: increase between-group separation
- α: tradeoff / strength of background suppression

Backward parameter selection

- Analyst moves a centroid or scales a confidence ellipse
- System solves inverse optimization for parameter values
- Makes interaction more intuitive than direct tuning alone

Case Study 1: Political Groups

- PPIC Statewide Survey, October 2018
- Compare Democrat and Republican supporters
- ULCA reveals subgroup structure within Democrats
- Further split into Dem(+) and Dem(-)
- Dem(-) exhibits larger internal variation
- Interactive refinement highlights politically meaningful questions

Case Study 2: Handwritten Digits

- MNIST digits: compare 0, 6, and 9
- ULCA emphasizes variation in 6 and 9 relative to 0
- Reweighting makes 6 and 9 more comparable in the embedding
- Pixel-contribution heatmaps interpret the embedding axes
- Learned projection is reused to compare digit 7

Take-home message

- ULCA unifies PCA, cPCA, and LDA-style goals
- Main novelty: flexible comparative analysis with interaction
- Best viewed as an interpretable exploratory tool

References

- Li, C. Preprocessing methods and pipelines of data mining: An overview. *arXiv preprint arXiv:1906.08510*, 2019.
- Fujiwara, T., Wei, X., Zhao, J., and Ma, K.-L. Interactive dimensionality reduction for comparative analysis. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):758–768, 2021.
- Abid, A., Zhang, M. J., Bagaria, V. K., and Zou, J. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature Communications*, 9(1):2134, 2018.