Unsupervised linear discrimination using skewness

J. Virta

University of Turku

December 3rd 2024, RIKEN Center for Brain Science

Reference

This talk is based on the papers

- **Radojičić, U., Nordhausen, K. and Virta, J.** (2024). Unsupervised linear discrimination using skewness. *Submitted*.
- **Radojičić, U., Nordhausen, K. and Virta, J.** (2021). Large-sample properties of unsupervised estimation of the linear discriminant using projection pursuit. *Electronic Journal of Statistics*, 15(2), 6677-6739.

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

Table of Contents

1. Two-group separation
2. Unsupervised estimator
3. More estimators
4. Finite-sample behavior
5. Closing remarks

Normal mixture

Our model

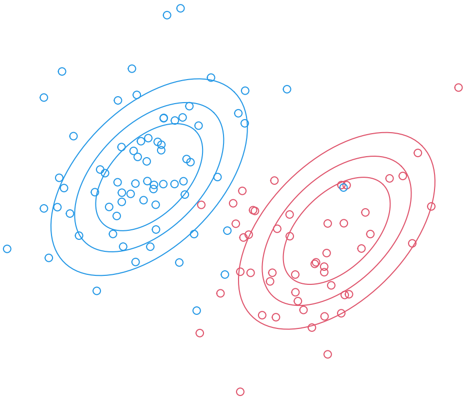
Let $X_1, \ldots, X_n \in \mathbb{R}^p$ be a random sample from the normal location mixture,

$$X \sim \alpha_1 \mathcal{N}(\mu_1, \Sigma) + \alpha_2 \mathcal{N}(\mu_2, \Sigma),$$

where

- $\alpha_1, \alpha_2 \in (0, 1)$, $\alpha_1 + \alpha_2 = 1$,
- $\mu_1 \neq \mu_2$,
- Σ is positive definite.

Illustration



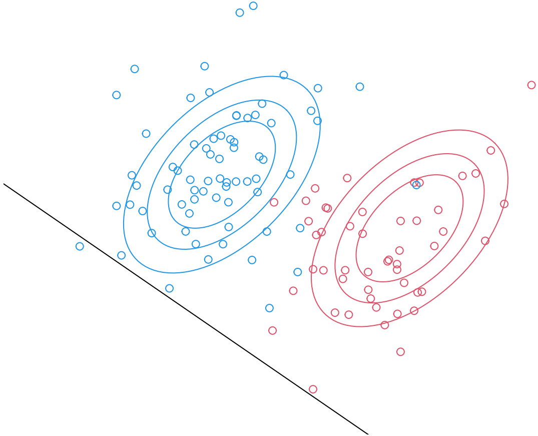
Separating projections

- We are interested in vectors $\theta \in \mathbb{R}^p$ such that the projection of the data

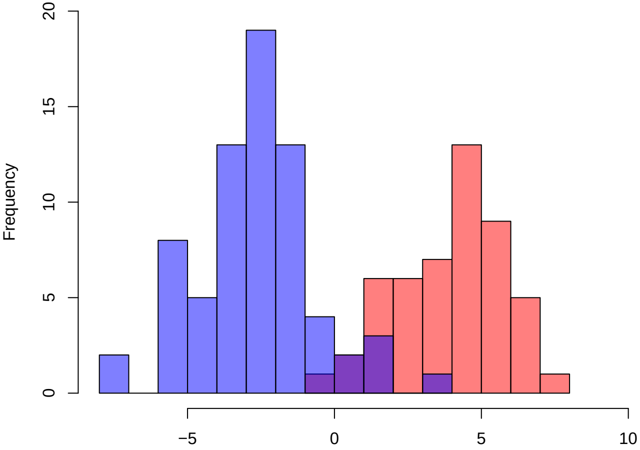
$$X \mapsto \theta' X$$

onto θ separates the two groups of the mixture

Illustration



The projection



Fisher's linear discriminant

- The projection direction on the previous slide is

$$\theta = \Sigma^{-1}(\mu_2 - \mu_1).$$

- This projection is used in linear discriminant analysis (LDA) and it leads to the Bayes optimal classifier under Gaussianity.

Sample estimator and asymptotic normality

- The sample LDA-estimator $\hat{\theta}$ is simple to compute given the sample $X_1, \ldots, X_n$ and the labels $Y_1, \ldots, Y_n$.
- $\hat{\theta}$ satisfies

$$\sqrt{n}\left(\frac{\hat{\theta}}{\|\hat{\theta}\|} - \frac{\theta}{\|\theta\|}\right) \rightsquigarrow \mathcal{N}_p(0, \Psi),$$

for a specific asymptotic covariance matrix Ψ.

Table of Contents

1 Two-group separation

2 Unsupervised estimator

3 More estimators

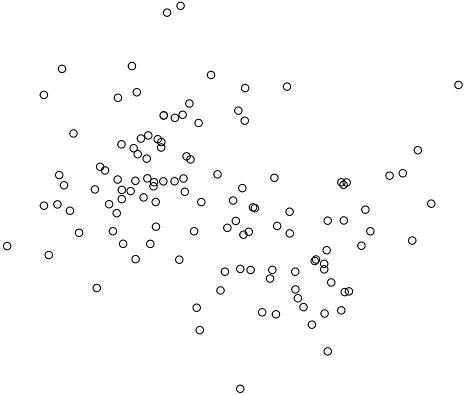
4 Finite-sample behavior

5 Closing remarks

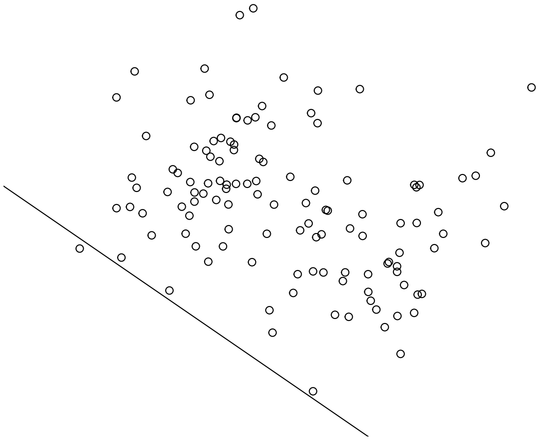
From classification to clustering

- The estimator $\hat{\theta}$ allows classification when (X_i, Y_i) are known.
- However, θ can be estimated also without the group labels Y_i! [Peña and Prieto, 2001]
- This lets us use the projection $\theta' X_i$ for clustering.

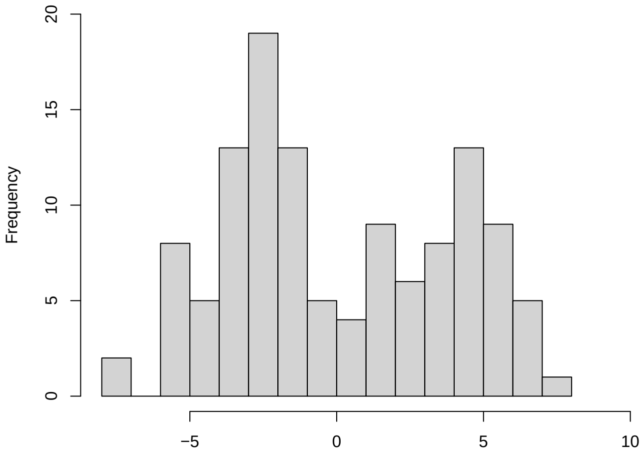
Unknown groups



Unknown groups



Histogram of the projection



Projection pursuit

- In projection pursuit [Huber, 1985], one chooses a projection index g which measures how "interesting" a random variable is and optimizes
$$\max_{\theta_0} g(\theta_0' X).$$
- One particular choice is squared Pearson's skewness, g_{skew}, which searches for maximally skewed projections.

Skewness and θ

Proposition 1 in [Loperfido, 2013]

For our normal mixture, if $\alpha_1 \neq \alpha_2$, the maximizer θ_{skew} of g_{skew} has

$$\frac{\theta_{\text{skew}}}{\|\theta_{\text{skew}}\|} = \frac{\theta}{\|\theta\|}.$$

Theorem 3 in [Radojičić et al., 2021]

Moreover,

$$\sqrt{n}\left(\frac{\hat{\theta}_{\text{skew}}}{\|\hat{\theta}_{\text{skew}}\|} - \frac{\theta}{\|\theta\|}\right) \rightsquigarrow \mathcal{N}_p(0, C\Psi),$$

for some constant $C \equiv C((\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 - \mu_1), \alpha_1\alpha_2)$.

Relative efficiencies vs. LDA

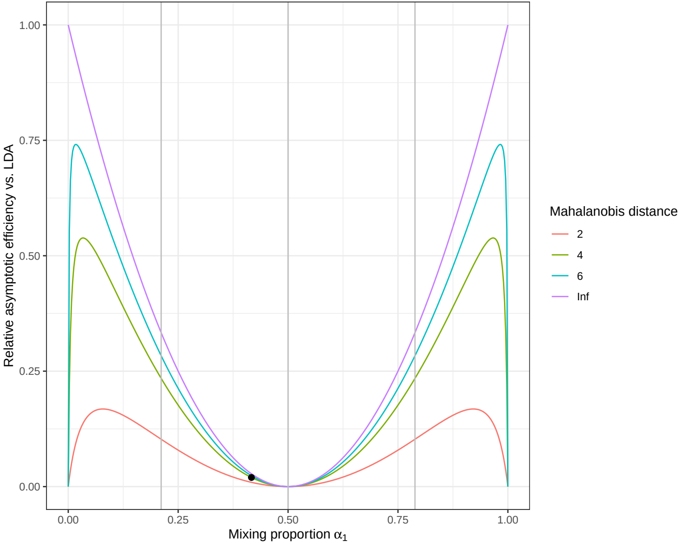


Table of Contents

1. Two-group separation
2. Unsupervised estimator
3. More estimators
4. Finite-sample behavior
5. Closing remarks

Affine equivariance

- An estimator $u(X_i) \in \mathbb{R}^p$ is affine equivariant if

$$u(A'X_i + b) = A^{-1}u(X_i)$$

  for all $b \in \mathbb{R}^p$ and all invertible $A \in \mathbb{R}^{p \times p}$.
- Projections onto affine equivariant directions are unaffected by the choice of the coordinate system of the data.

General result

Theorem 5 in [Radojičić et al., 2024]

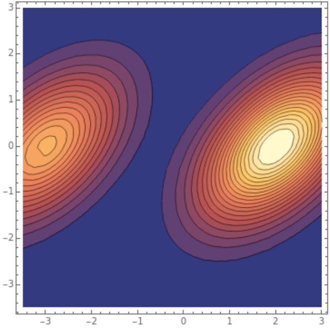
Any affine equivariant estimator of $\theta/\|\theta\|$ has asymptotic covariance matrix proportional to Ψ.

Standardization

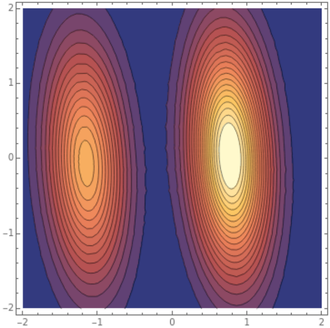
The standardized random vector X_{st} is obtained as

$$X_{\mathrm{st}} = \mathrm{Cov}(X)^{-1/2}\{X - \mathrm{E}(X)\}.$$

Illustration



Original data, X



Stndardized data, X_{st}

Estimator from Loperfido (2013)

- [Loperfido, 2013] defines the affine equivariant estimator $\theta_{\mathrm{L}} = \mathrm{Cov}(X)^{-1/2} u(X)$, where $u(X)$ is the leading unit-length eigenvector of the matrix

$$[\mathrm{E}\{(X_{\mathrm{st}} \otimes X_{\mathrm{st}}) X'_{\mathrm{st}}\}]' [\mathrm{E}\{(X_{\mathrm{st}} \otimes X_{\mathrm{st}}) X'_{\mathrm{st}}\}].$$

Proposition 3 in [Loperfido, 2013]

For our normal mixture, if $\alpha_1 \neq \alpha_2$, then

$$\frac{\theta_{\mathrm{L}}}{\|\theta_{\mathrm{L}}\|} = \frac{\theta}{\|\theta\|}.$$

A novel estimator

- [Radojičić et al., 2024] defines the affine equivariant estimator $\theta_{\mathrm{J}} = \mathrm{Cov}(X)^{-1/2} u(X)$, where $u(X)$ is the maximizer of

$$\max_{v \in \mathbb{R}^p, \|v\|=1} \sum_{k=1}^{p} \left\{ v^\top \mathrm{E}(X_{\mathrm{st}} X_{\mathrm{st}}' e_k X_{\mathrm{st}}') v \right\}^2 .$$

Lemma 7 in [Radojičić et al., 2024]

For our normal mixture, if $\alpha_1 \neq \alpha_2$, then

$$\frac{\theta_{\mathrm{J}}}{\|\theta_{\mathrm{J}}\|} = \frac{\theta}{\|\theta\|}.$$

Limiting covariance matrices

Theorems 6 and 7 in [Radojičić et al., 2024]

The asymptotic covariance matrices of $\hat{\theta}_{\mathrm{J}}/\|\hat{\theta}_{\mathrm{J}}\|$ and $\hat{\theta}_{\mathrm{L}}/\|\hat{\theta}_{\mathrm{L}}\|$ are exactly the same and equal to that of $\hat{\theta}_{\mathrm{skew}}/\|\hat{\theta}_{\mathrm{skew}}\|$.

Table of Contents

1. Two-group separation
2. Unsupervised estimator
3. More estimators
4. Finite-sample behavior
5. Closing remarks

Simulation comparison

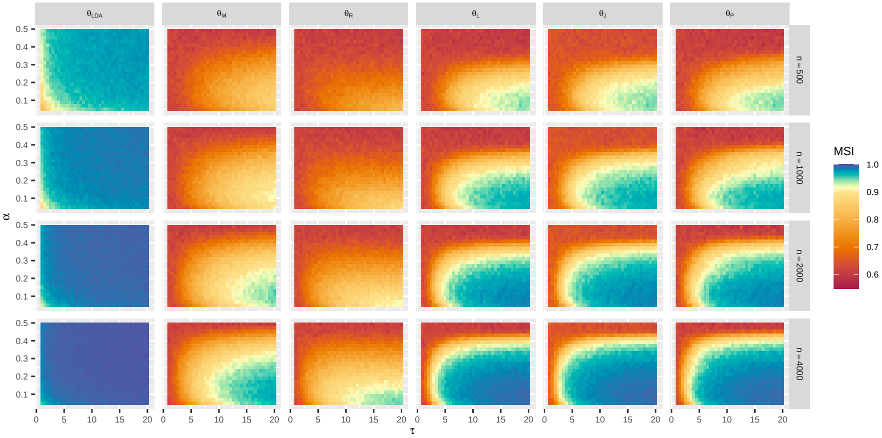


Table of Contents

1 Two-group separation

2 Unsupervised estimator

3 More estimators

4 Finite-sample behavior

5 Closing remarks

History of the problem

Kurtosis-based approaches:

- [Peña and Prieto, 2001]
- [Peña et al., 2010]
- [Peña et al., 2017]
- [Radojičić et al., 2021]

Skewness-based approaches:

- [Loperfido, 2013]
- [Loperfido, 2015]
- [Radojičić et al., 2021]
- [Radojičić et al., 2024]

Future directions

- Analogous study for kurtosis would allow discarding the assumption that $\alpha_1 \neq \alpha_2$.
- Going beyond Gaussianity?
- Unequal covariance matrices?
- Multiple groups?
- High-dimensional variants?

Thank you for your attention!

References I

Huber, P. J. (1985).
Projection pursuit.
Annals of Statistics, 13:435–475.

Loperfido, N. (2013).
Skewness and the linear discriminant function.
Statistics & Probability Letters, 83(1):93–99.

Loperfido, N. (2015).
Vector-valued skewness for model-based clustering.
Statistics & Probability Letters, 99:230–237.

Peña, D. and Prieto, F. J. (2001).
Cluster identification using projections.
Journal of the American Statistical Association, 96:1433–1445.

Peña, D., Prieto, F. J., and Viladomat, J. (2010).
Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure.
Journal of Multivariate Analysis, 101:1995–2007.

Peña, D., Prieto, J., and Rendon, C. (2017).
Clustering big data by extreme kurtosis projections.

References II

Radojičić, U., Nordhausen, K., and Virta, J. (2021).
Large-sample properties of blind estimation of the linear discriminant using projection pursuit.
Electronic Journal of Statistics, 15(2).

Radojičić, U., Nordhausen, K., and Virta, J. (2024).
Unsupervised linear discrimination using skewness.
Submitted.