MULTIVARIATE ANALYSIS

AS A NULL COMPARISON

WITH APPLICATIONS TO CLASSICAL, FUNCTIONAL AND TRANSFORMATIONAL DATA ANALYSIS

Andreas Buja
The Wharton School, University of Pennsylvania

Joint work with Deborah Donnell and Werner Stuetzle
(Fred Hutchinson Cancer Research, University of Washington)

What is Classical Multivariate Analysis?

Examples: (for random variables instead of data)

- **PCA**: Principal Component Analysis

  $$\mathbf{Var}(\sum_i a_i X_i) = \max$$

  subject to

  $$\sum a_i^2 = 1$$

- **CCA**: Canonical Correlation Analysis

  $$\mathbf{Cov}(\sum_i a_i X_i, \sum_j b_j Y_j) \ = \ \max$$

  subject to

  $$\mathbf{Var}(\sum_i a_i X_i) = \mathbf{Var}(\sum_j b_j Y_j) = 1$$

- **LDA**: Linear Discriminant Analysis

  For dimension reduction: LDA = CCA, where ...

  ... Y_j are dummy variables of a partition.

- **GCA**: Generalized Canonical Analysis

  = CCA with ...

  ... more than two variable blocks.

What is PCA?

$$\mathbf{Var}(\sum_i a_i X_i) = \max \quad \text{subject to} \quad \sum a_i^2 = 1$$

Question: **Why** $\sum a_i^2 = 1$ **?**

- Yes, we need a constraint, but why this one?
- Most common version of PCA:
  correlation-based, not covariance-based
  $\Rightarrow$ constraint $\sum a_i^2 \, \mathbf{Var}(X_i) = 1$
- When do we use cov-, when corr-based PCA?
- "Dirty little secret of machine learning:"
  How do we choose reference metrics?
  (Tom Dietterich, U. of Oregon)

Goal: An answer for multivariate analysis

Reference Metrics for PCA:

- **Idea**: Derive the reference metric from the criterion by evaluating the criterion under a **null hypothesis**.

- PCA should be driven by correlations.
  $\Rightarrow$ Null hypothesis = "no correlations"

$$H_0 : \quad \mathbf{Cov}(X_i, X_j) = 0 \quad (i \neq j)$$

What is the criterion under the null hypothesis?

$$\mathbf{Var}(\sum a_j X_j) \;=\; \sum a_j^2 \,\mathbf{Var}(X_j)$$

$\Rightarrow$ Constraint: $\sum_j a_j^2 \,\mathbf{Var}(X_j) = 1.$

$\Rightarrow$ Correlation-based PCA

- Add equal variances to the null hypothesis:

$$H_0 : \quad ... \text{ and } \ \mathbf{Var}(X_i) = V \quad (\forall i)$$

The criterion under the null hypothesis:

$$\mathbf{Var}(\sum_j a_j X_j) \ = \ \sum_j a_j^2 \, V$$

$\Rightarrow$ Constraint: $\quad \sum_j a_j^2 = 1/V$.

$\Rightarrow$ Covariance-based PCA

- Difference:
  - Correlation-based PCA responds only to correlations.
  - Covariance-based PCA responds also to non-constant variances.

Example: $\mathbf{Cov}(X_i, X_j) = 0$,
$\mathbf{Var}(X_1) = 10, \quad \mathbf{Var}(X_2) = ... = \mathbf{Var}(X_p) = 1$
When do you want which? ...

What is CCA?

- Usual definition:

$$\mathbf{Corr}(\sum a_i X_i, \sum b_j Y_j) \;=\; \max$$

- Equivalent:

$$\mathbf{Cov}(\sum a_i X_i, \sum b_j Y_j) \;=\; \max$$

subject to $\mathbf{Var}(\sum a_i X_i) = \mathbf{Var}(\sum b_j Y_j) = 1.$

- Equivalent:

$$\mathbf{Var}(\sum a_i X_i \pm \sum b_j Y_j) \;=\; \max / \min$$

subject to $\mathbf{Var}(\sum a_i X_i) + \mathbf{Var}(\sum b_j Y_j) = 2.$

CCA as a Null Comparison:

- Criterion:

$$\mathbf{Var}(\sum a_i X_i \pm \sum b_j Y_j) \;=\; \max / \min$$

- Null hypothesis of CCA:

$$\mathbf{Cov}(X_i, Y_j) = 0 \ (\forall i, j)$$

  (The two variable blocks are uncorrelated between but not within.)

- Constraint or "reference metric"
  = criterion evaluated under the null hypothesis:

$$\mathbf{Var}(\sum a_i X_i \pm \sum b_j Y_j) \;=\; \mathbf{Var}(\sum a_i X_i) + \mathbf{Var}(\sum b_j Y_j)$$

GCA as a Null Comparison:

- Three or more blocks of variables: X_i, Y_j, Z_k, ...
- Criterion:

  $$\mathbf{Var}(\sum a_i X_i + \sum b_j Y_j + \sum c_k Z_k + ...)$$

- Null hypothesis:

  Blocks are uncorrelated between, but not within.

  $$\mathbf{Cov}(X_i, Y_j) = \mathbf{Cov}(X_i, Z_k) = \mathbf{Cov}(Y_j, Z_k) = 0$$

- Evaluation of criterion under the null hypothesis:

  $$\begin{aligned} & \mathbf{Var}(\sum a_i X_i + \sum b_j Y_j + \sum c_k Z_k) \\ = & \ \mathbf{Var}(\sum a_i X_i) + \mathbf{Var}(\sum b_j Y_j) + \mathbf{Var}(\sum c_k Z_k) \\ = & \ \text{reference metric} \end{aligned}$$

- Uses: Multiple correspondence analysis

  X_i = dummy variables of partition 1

  Y_j = dummy variables of partition 2

  Z_k = dummy variables of partition 3

  $\sum a_i X_i$ = quantification of categories of partition 1

  ...

Summary of Classical MA Methods:

Rayleigh quotient of ...

- **PCA** (corr-based):

$$\frac{\mathbf{Var}(\sum a_i X_i)}{\sum a_i^2\, \mathbf{Var}(X_i)}$$

- **PCA** (cov-based):

$$\frac{\mathbf{Var}(\sum a_i X_i)}{\sum a_i^2\, V}$$

- **CCA**:

$$\frac{\mathbf{Var}(\sum a_i X_i + \sum b_j Y_j)}{\mathbf{Var}(\sum a_i X_i) + \mathbf{Var}(\sum b_j Y_j)}$$

- **GCA**:

$$\frac{\mathbf{Var}(\sum a_i X_i + \sum b_j Y_j + \sum c_k Z_k)}{\mathbf{Var}(\sum a_i X_i) + \mathbf{Var}(\sum b_j Y_j) + \mathbf{Var}(\sum c_k Z_k)}$$

Extensions of Classical MA:

- **CCA with equal covariance in the two blocks**:

  Assume correspondence $X_1 \sim Y_1$, ..., $X_p \sim Y_p$.

  H_0: $\mathbf{Cov}(X_i, X_j) = \mathbf{Cov}(Y_i, Y_j)$
  in addition to $\mathbf{Cov}(X_i, Y_j) = 0$.

  $$\text{Rayleigh} \;=\; \frac{\mathbf{Var}(\sum a_i X_i + \sum b_i Y_i)}{\left(\begin{array}{r} (\mathbf{Var}(\sum a_i X_i) + \mathbf{Var}(\sum a_i Y_i))/2 \\ +\,(\mathbf{Var}(\sum b_i Y_i) + \mathbf{Var}(\sum b_i X_i))/2 \end{array}\right)}$$

- **PCA with H_0 = exchangeable covariance**:

  H_0: $\mathbf{Cov}(X_i, X_j) = C \quad (\forall i \neq j)$
  in addition to $\mathbf{Var}(X_i) = V \quad (\forall i)$.

  $$\text{Rayleigh} \;=\; \frac{\mathbf{Var}(\sum a_i X_i)}{\sum_i a_i^2\, V \;+\; \sum_{i \neq j} a_i a_j\, C}$$

  where $V = \mathbf{ave}(\mathbf{Var}(X_i) | i)$
  and $\quad C = \mathbf{ave}(\mathbf{Cov}(X_i, X_j) | i \neq j)$.

Further Extensions:

- **"Chaining Variance" for Tri-Partite Graphs**:

  Criterion:

  $$\mathbf{Var}(\sum a_i X_i - \sum b_j Y_j) + \mathbf{Var}(\sum b_j Y_j - \sum c_k Z_k)$$

  H_0: Three blocks uncorrelated between, not within.

  Reference Metric:

  $$\mathbf{Var}(\sum a_i X_i) + 2\mathbf{Var}(\sum b_j Y_j) + \mathbf{Var}(\sum c_k Z_k)$$

  Rayleigh = Criterion / Reference Metric

- **Diagnostics for Any Covariance Model**

  H_0: Structural equation, Lisrel, graphical ... model

  $\mathbf{C}$ = unconstrained covariance matrix ($\mathbf{Cov}(X_i, X_j)$)

  $\mathbf{C}_0$ = covariance matrix fitted under H_0

  $$\text{Rayleigh} = \frac{\mathbf{a}^T \mathbf{C}\, \mathbf{a}}{\mathbf{a}^T \mathbf{C}_0\, \mathbf{a}}$$

Functional Multivariate Analysis:

- "Functional": Indexes have topological structure.
  - Example 1: $(X_1, ..., X_p)$ = time series
    X_i = observation at the i'th time point
  - Example 1: $(X_1, ..., X_p)$ = spatial random field
    X_i = observation at the i'th location
- Requirement: i and j "near" $\Rightarrow |a_i - a_j|$ small.
- Quadratic penalization:
  - Example 1: $(X_1, ..., X_p)$ = time series
    $\Rightarrow$ Naive penalty $= \sum_{i=2,...p}(a_i - a_{i-1})^2$
  - General: Penalty $= \mathbf{a}^T \Omega \, \mathbf{a}$,
    Ω is $p \times p$ non-negative definite
    Principled: smoothing splines, RKHS
- **Recipe** for penalization approaches to functional MA:
  Add quadratic penalties to variance criteria
  and derive reference metrics from H_0.

Functional PCA:

- Examples: Multiple time series, multiple random fields aligned on same grid

- Criterion: $\mathbf{Var}(\sum a_i X_i) + \mathbf{a}^T \Omega\, \mathbf{a}$

- H_0: $\mathbf{Cov}(X_i, X_j) = 0 \ (i \neq j)$, $\mathbf{Var}(X_i) = V$

- Reference metric: $\sum a_i^2 \, V + \mathbf{a}^T \Omega\, \mathbf{a}$

- Rayleigh:

$$\frac{\mathbf{Var}(\sum a_i X_i) + \mathbf{a}^T \Omega\, \mathbf{a}}{\sum a_i^2 \, V + \mathbf{a}^T \Omega\, \mathbf{a}}$$

- Rice & Silverman (1991):

$$\frac{\mathbf{Var}(\sum a_i X_i) - \mathbf{a}^T \Omega\, \mathbf{a}}{\sum a_i^2}$$

- Silverman (1996):

$$\frac{\mathbf{Var}(\sum a_i X_i)}{\sum a_i^2 + \mathbf{a}^T \Omega\, \mathbf{a}}$$

Functional CCA:

- Assume both $X_1, ..., X_p$ and $Y_1, ..., Y_q$ functional.

- Two penalties: $\mathbf{a}^T \Omega_X \mathbf{a}$ and $\mathbf{b}^T \Omega_Y \mathbf{b}$

- Criterion:

  $$\mathbf{Var}(\sum a_i X_i - \sum b_j Y_j) + \mathbf{a}^T \Omega_X \, \mathbf{a} + \mathbf{b}^T \Omega_Y \mathbf{b}$$

- Reference metric:

  $$\mathbf{Var}(\sum a_i X_i) + \mathbf{Var}(\sum b_j Y_j) + \mathbf{a}^T \Omega_X \, \mathbf{a} + \mathbf{b}^T \Omega_Y \mathbf{b}$$

- Rayleigh is equivalent to pen. can. corr.:

$$\frac{\mathbf{Cov}(\sum a_i X_i, \sum b_j Y_j)}{\left[(\mathbf{Var}(\sum a_i X_i) + \mathbf{a}^T \Omega_X \mathbf{a})(\mathbf{Var}(\sum b_j Y_j) + \mathbf{b}^T \Omega_Y \mathbf{b})\right]^{1/2}}$$

- CCA of Leurgans, Moyeed and Silverman (1993), Pen. LDA of Hastie, Buja, Tibshirani (1995)

- CCA needs penalization much more than PCA because of collinearity within blocks due to functional nature of the data.

Outlook on Transformational Data Analysis:

- Want to marginally transform variables: $f_i(X_i)$, before subjecting them to MA.

- PCA: criterion = $\mathbf{Var}(\sum a_i f_i(X_i))$
  (absorb a_i in f_i)
  $\Rightarrow$ criterion = $\mathbf{Var}(\sum f_i(X_i))$

- H_o: $X_1, ..., X_p$ are pairwise independent.

- Reference metric: $\sum \mathbf{Var}(f_i(X_i))$

- Underlying: infinite dimensional GCA
  Dauxois & Pousse (1976)
  Closed subspaces $V_1, ..., V_p$ in a Hilbert space
  Find $v_i \in V_i$ such that $\| \sum v_i\|^2 / \sum \|v_i\|^2 = \max$
  Reference metric: assume V_i are orthogonal.

- Donnell, Buja, Stuetzle (1994):
  Estimate implicit equations $\sum f_i(X_i) \approx 0$

Transformational CCA:

- Breiman & Friedman's ACE "regression":
  Criterion = $\mathbf{Var}(g(Y) - \sum f_i(X_i))$

- H_0: Y is independent of $(X_1, ..., X_p)$

- Reference metric: $\mathbf{Var}(g(Y)) + \mathbf{Var}(\sum f_i(X_i))$

- CCA form of ACE:
  $\mathbf{Corr}(g(Y), \sum f_i(X_i)) = \max$

- "Regression" form of ACE:

$$1 - R^2 \; = \; \frac{\mathbf{Var}(g(Y) - \sum f_i(X_i))}{\mathbf{Var}(g(Y))}$$

Note: Transformational MA can be easily done with spline transformations using spline penalties.
Criterion for ACE:

$$\mathbf{Var}(g(Y) - \sum f_i(X_i)) + J(g) + \sum J_i(f_i)$$

Conclusions:

- All of classical MA has variance as the criterion.
- The types of MA differ in the reference metric, not the criterion.
- Functional MA can be performed by adding quadratic penalties to the variance.
- Transformational MA can be performed using spaces of transformations instead of spaces of linear combinations of variables.
- Penalized transformational MA can be performed by adding spline penalties to the variance criterion.