Object Oriented Data Analysis

J. S. Marron

Dept. of Statistics and Operations Research,
University of North Carolina

January 16, 2014
An Aside on Current “Fashion”

Big Data

• Isn’t It Just Statistics?
An Aside on Current “Fashion”

Big Data

• Isn’t It Just Statistics?
• Yes, But We Need to Remind Folks
• Maybe Bigger Challenge:

Complex Data
What is the “atom” of a statistical analysis?

- 1st Course: Numbers
- Multivariate Analysis Course: Vectors
- Functional Data Analysis: Curves
- More generally: Data Objects
Object Oriented Data Analysis

Original Thought:
OODA = Mathematical Framework
(containing wide variety of interesting cases)
Object Oriented Data Analysis

Original Thought:
OODA = Mathematical Framework

Current View:
OODA = Focal Point
Object Oriented Data Analysis

Original Thought:
OODA = Mathematical Framework

Current View:
OODA = Focal Point

{For discussions (interdisciplinary) about tackling serious analyses}
Object Oriented Data Analysis

Original Thought:
OODA = Mathematical Framework

Current View:
OODA = Focal Point

What should be the Data Objects?
Curves as Data Objects

Important Duality:

Curve Space \leftrightarrow Point Cloud Space

Illustrate with Travis Gaydos Graphics

- 2 dim’al curves (easy to visualize)
Curves as Data Objects

Important Duality Concept:

Curve Space $\leftrightarrow$ Point Cloud Space

($=\text{Object Space}$) ($=\text{Feature Space}$)

Illustrate with Travis Gaydos Graphics

- 2 dim’al curves (easy to visualize)
Functional Data Analysis, Toy EG I
Functional Data Analysis, Toy EG II
Functional Data Analysis, Toy EG III
Functional Data Analysis, Toy EG IV

Centered Raw Data Curves

Centered Raw Curves as Point Cloud
Functional Data Analysis, Toy EG V

Centered Raw Data Curves

Centered Raw Curves as Point Cloud
Functional Data Analysis, Toy EG VI

Projection of Centered Curves on PC1

Projection of Points onto PC1
Functional Data Analysis, Toy EG VII

Centered Raw Data Curves

Centered Raw Curves as Point Cloud
Functional Data Analysis, Toy EG VIII

Projection of Centered Curves on PC2

Projection of Points onto PC2
Functional Data Analysis, Toy EG IX

Raw Data Curves

Mean Curve

Projection of Curves on PC1

Projection of Curves on PC2
Functional Data Analysis, Toy EG X
Functional Data Analysis, 10-d Toy EG 1
Functional Data Analysis, 10-d Toy EG 1

PC2 Proj.

Center -- PC2

10.42% of CR

Center -- PC3

0.7035% of CR

Center -- PC4

0.6682% of CR

Scores kde, PC2

Scores kde, PC3

Scores kde, PC4
Functional Data Analysis, 50-d Toy EG 2
Functional Data Analysis, 50-d Toy EG 2
Principal Component Analysis

More Than *Dimensionality Reduction*
More Than *Dimensionality Reduction*:

- **Visualization**
  - Relationships Between Objects (Scores)
  - Drivers of Relationships (Loadings)
More Than *Dimensionality Reduction*:

- **Visualization**
  - Relationships Between Objects (Scores)
  - Drivers of Relationships (Loadings)
- **Summarization**
  - Lower-d Representation
  - E.g. \( n \ll d \)
Principal Component Analysis

More Than *Dimensionality Reduction*:

- **Visualization**
  - Relationships Between Objects (Scores)
  - Drivers of Relationships (Loadings)
- **Summarization**
  - Lower-d Representation
  - E.g. $n << d$
- **Careful about Information Loss**
Interesting Data Set:
• Mortality Data
• For Spanish Males (thus can relate to history)
• Each curve is a single year
• x coordinate is age
• Mortality = # died / total # (for each age)
• Study on log scale
Mortality Time Series

Conventional Coloring:

Rotate Through (7) Colors

Hard to See Time Structure
Mortality Time Series

Improved Coloring:

Rainbow Representing Year:

Magenta = 1908

Red = 2002
Mortality Time Series

Find Population Center (Mean Vector)

Compute in Feature Space

Show in Object Space

Spanish Males Mortality Curves, 1908–2002

97.62% of Tol

mean \log(\text{Mortality})

Age
Mortality Time Series

Blips Appear At Decades Since Ages Not Precise (in Spain) Reported as “about 50”, Etc.
Mortality Time Series

Spanish Males Mortality Curves, 1908–2002

Mean Residual

Object Space View of Shifting Data To Origin In Feature Space
Mortality Time Series

Shows:

Main Age Effects in Mean, Not Variation About Mean
Mortality Time Series

Object Space
View of
Projections
Onto PC1
Direction

Main Mode
Of Variation:
Constant
Across Ages

Spanish Males Mortality Curves, 1908–2002

PC1 Loadings Plot
Mortality Time Series

Shows Major Improvement Over Time

(Pub. Health, Medicine,...)

Loadings: Biggest Benefit For the Young

PC1 Loadings Plot
Mortality Time Series

- Shows **Major** Improvement Over Time
- And Change In Age Rounding Blips

**Spanish Males Mortality Curves, 1908–2002**

**PC1 Loadings Plot**
Mortality Time Series

Corresponding PC 1 Scores Again Shows Overall Improvement

High Mortality Early
Lower Later
Mortality Time Series

Outliers

1918 Global Flu Pandemic

1936-1939 Spanish Civil War
Mortality Time Series

Object Space View of Projections Onto PC2 Direction

2nd Mode Of Variation: Difference Between 20-45 & Rest
Mortality Time Series

Explain Using PC 2 Scores

Early Improvement

Pandemic Hit Hardest
Mortality Time Series

Explain Using PC 2 Scores

Then better

Spanish Civil War Hit Hardest
Mortality Time Series

Explain Using PC 2 Scores

Steady Improvement To mid-50s

Increasing Automotive Death Rate
Mortality Time Series

Explain Using PC 2 Scores

Corner Finally Turned by Safer Cars & Roads
Mortality Time Series

Feature (Point Cloud) Space View

Connecting Lines Highlight Time Order

Good View of Historical Effects
Interesting Real Data Example

- Genetics (Cancer Research)
- RNAseq (Next Gener’n Sequen’g)
- Deep look at “gene components”

Microarrays: Single number (per gene)
RNAseq: Thousands of measurements
Functional Data Analysis

Interesting Real Data Example

- Genetics (Cancer Research)
- RNAseq (Next Gener’n Sequen’g)
- Deep look at “gene components”
Interesting Real Data Example
• Genetics (Cancer Research)
• RNAseq (Next Gener’n Sequen’g)
• Deep look at “gene components”

• Gene studied here: CDKN2A
• Goal: *Study Alternate Splicing*
• Sample Size, \( n = 180 \)
• Dimension, \( d = \sim 1700 \)
Simple 1st View: Curve Overlay (log scale)

Thanks to Matt Wilkerson
Functional Data Analysis

Often Useful Population View:

PCA Scores
Functional Data Analysis

Suggestion Of Clusters
Functional Data Analysis

Suggestion Of Clusters Which Are These?
Functional Data Analysis

Manually Brush Clusters
Functional Data Analysis

Manually Brush Clusters Clear Alternate Splicing
Functional Data Analysis

Important Points

✓ PCA found *Important Structure*

✓ In **High Dimensional** Data Analysis
Object Oriented Data Analysis

What is the “atom” of a statistical analysis?

- **1st Course:** Numbers
- Multivariate Analysis Course: Vectors
- Functional Data Analysis: Curves
- More generally: Data Objects
Object Oriented Data Analysis

Examples:

- Medical Image Analysis
  - Images as Data Objects?
  - Shape Representations as Objects

- Gene Expression (Microarrays – RNAseq)
  - Just multivariate analysis?
Object Oriented Data Analysis

Typical Goals:

- Understanding population variation
- Visualization
- Principal Component Analysis + Discrimination (a.k.a. Classification)
- “Vertical Integration” of Data Types
Major Statistical Challenge, I:

High Dimension Low Sample Size (HDLSS)

- Dimension $d \gg \text{sample size } n$
- “Multivariate Analysis” nearly useless
  - Can’t “normalize the data”
- Land of Opportunity for Statisticians
  - Need for “creative statisticians”
Aside on Terminology

Personal suggestion:

High Dimension Low Sample Size (HDLSS)

- Dimension: $d$
- Sample size $n$

Versus: “Small n, large p”

- Why p? (parameters??? predictors???)
- Only because of statistical tradition...
Major Statistical Challenge, II:

- Data may live in *non-Euclidean space*
  - Lie Group / Symmetric Spaces (manifold data)
  - Trees/Graphs as data objects

**Interesting Issues:**
- What is “the mean” (pop’n center)?
- How do we quantify “pop’n variation”?
First Generation Problems:

- Denoising
- Segmentation
- Registration

(all about single images)
Second Generation Problems:

- Populations of Images
  - Understanding Population Variation
  - Discrimination (a.k.a. Classification)
- Complex Data Structures (& Spaces)
- HDLSS Statistics
Why HDLSS (High Dim, Low Sample Size)?

- Complex 3-d Objects Hard to Represent
  - Often need $d = 100$’s of parameters

- Complex 3-d Objects Costly to Segment
  - Often have $n = 10$’s cases
Medical Imaging – A Challenging Example

- Male Pelvis
  - Bladder – Prostate – Rectum
  - How do they move over time (days)?
  - Critical to Radiation Treatment (cancer)

- Work with 3-d CT
- Very Challenging to Segment
  - Find boundary of each object?
  - Represent each Object?
Male Pelvis – Raw Data

One CT Slice
(in 3d image)

Coccyx
(Tail Bone)

Rectum

Bladder
Bladder:

manual segmentation

Slice by slice

Reassembled
Male Pelvis – Raw Data

Bladder:

Slices:
   Reassembled in 3d

How to represent?

Thanks: Ja-Yeon Jeong
Object Representation

- Landmarks (hard to find)
- Boundary Rep’ns (no correspondence)
- Medial representations
  - Find “skeleton”
  - Discretize as “atoms” called M-reps
3-d m-reps

Bladder – Prostate – Rectum  (multiple objects, J. Y. Jeong)

• Medial Atoms provide “skeleton”

• Implied Boundary from “spokes” → “surface”
3-d m-reps

M-rep model fitting

- Easy, when starting from binary (blue)
- But very expensive (30 – 40 minutes technician’s time)
- Want automatic approach
- Challenging, because of poor contrast, noise, ...
- Need to borrow information across training sample
- Use Bayes approach: prior & likelihood $\rightarrow$ posterior
- $\sim$Conjugate Gaussians, but there are issues:
  - Major HLDSS challenges
  - Manifold aspect of data
Object Space $\leftrightarrow$ Feature Space

Focus here on collection of *data objects*

Here conceptualize *population structure* via “point clouds”
Data Lying On a Manifold

Major issue: m-reps live in \( \mathbb{R}^3 \times \mathbb{R}^+ \times S^2 \times S^2 \)

(locations, radius and angles)

E.g. “average” of: \( 2^\circ, 3^\circ, 358^\circ, 359^\circ = ??? \)
Data Lying On a Manifold

Major issue: m-reps live in $\mathbb{R}^3 \times \mathbb{R}^+ \times S^2 \times S^2$

(locations, radius and angles)

E.g. “average” of: $2^\circ, 3^\circ, 358^\circ, 359^\circ = ?$

$\Sigma_i \theta_i / 4$  ?
Data Lying On a Manifold

Major issue: m-reps live in $\mathbb{R}^3 \times \mathbb{R}^+ \times S^2 \times S^2$
(locations, radius and angles)

E.g. “average” of: $2^\circ, 3^\circ, 358^\circ, 359^\circ = ???$

$\sum_i \theta_i / 4$
Data Lying On a Manifold

Major issue: m-reps live in $\mathbb{R}^3 \times \mathbb{R}^+ \times S^2 \times S^2$
(locations, radius and angles)

E.g. “average” of: $2^\circ, 3^\circ, 358^\circ, 359^\circ = ???$
Data Lying On a Manifold

Major issue: m-reps live in $\mathbb{R}^3 \times \mathbb{R}^+ \times S^2 \times S^2$

(locations, radius and angles)

E.g. “average” of: $2^\circ, 3^\circ, 358^\circ, 359^\circ = ???$

Natural Data Structure is:

Lie Groups $\sim$ Symmetric spaces
(smooth, curved manifolds)
Mildly Non-Euclidean Space

Useful View of Manifold Data: Tangent Space

Figure 2.2: The Riemannian exponential map.
Mildly Non-Euclidean Space

Useful View of Manifold Data: Tangent Space

At each point, $\exists$ Approximating Tangent Plane

Figure 2.2: The Riemannian exponential map.

Thanks to P. T. Fletcher
Mildly Non-Euclidean Space

Useful View of Manifold Data: Tangent Space

At each point, \( \exists \) Approximating Tangent Plane

Reason for terminology “mildly non Euclidean”
Mildly Non-Euclidean Space

Useful View of Manifold Data: Tangent Space

Useful Data

Center Point:

Geodesic Mean

(= Frèchet Mean)

(= Barycenter)

Figure 2.2: The Riemannian exponential map.

Thanks to P. T. Fletcher
Geodesic Mean

For $X_1, \ldots, X_n$ in any metric space:

$$\text{Mean} = \arg\min_x \sum_{i=1}^{n} d(x, X_i)^2$$

($x = \text{point with least square distance to data}$)
For $X_1, \ldots, X_n$ in any metric space:

$$\text{Mean} = \arg\min_x \sum_{i=1}^{n} d(x, X_i)^2$$

($x =$ point with least square distance to data)

Geodesic Mean (on Manifolds):

$$d = \text{Geodesic Distance}$$

(Along Manifold Surface)
PCA on non-Euclidean spaces?
(i.e. on Lie Groups / Symmetric Spaces)

T. Fletcher: Principal Geodesic Analysis

Idea: replace “linear summary of data”
With “geodesic summary of data”...
PGA for m-reps, Bladder-Prostate-Rectum

Bladder – Prostate – Rectum, 1 person, 17 days

PG 1               PG 2               PG 3

(analysis by Ja Yeon Jeong)
PGA for m-reps, Bladder-Prostate-Rectum

Bladder – Prostate – Rectum, 1 person, 17 days

PG 1                   PG 2                   PG 3

(analysis by Ja Yeon Jeong)
PGA for m-reps, Bladder-Prostate-Rectum

Bladder – Prostate – Rectum, 1 person, 17 days

PG 1  PG 2  PG 3

(analysis by Ja Yeon Jeong)
PCA Extensions for Data on Manifolds

• Fletcher (Principal Geodesic Anal.)
  • Best fit of geodesic to data
  • Constrained to go through geodesic mean
PCA Extensions for Data on Manifolds

- Fletcher (Principal Geodesic Anal.)
  - Best fit of geodesic to data
  - Constrained to go through geodesic mean

Counterexample:

Data on sphere, along equator
PCA Extensions for Data on Manifolds

- Fletcher (Principal Geodesic Anal.)
  - Best fit of geodesic to data
  - Constrained to go through geodesic mean
- Huckemann, Hotz & Munk (Geod. PCA)
  - Best fit of any geodesic to data
PCA Extensions for Data on Manifolds

- Fletcher (Principal Geodesic Anal.)
  - Best fit of geodesic to data
  - Constrained to go through geodesic mean
- Huckemann, Hotz & Munk (Geod. PCA)
  - Best fit of any geodesic to data

Counterexample:

Data follows Tropic of Capricorn

(thanks to Ja-Yeon Jeong)
PCA Extensions for Data on Manifolds

- Fletcher (Principal Geodesic Anal.)
  - Best fit of geodesic to data
  - Constrained to go through geodesic mean
- Huckemann, Hotz & Munk (Geod. PCA)
  - Best fit of any geodesic to data
- Jung, Foskey & Marron (Princ. Arc Anal.)
  - Best fit of any circle to data
  (motivated by conformal maps)
PCA Extensions for Data on Manifolds

Figure: Generalization of PCA on $S^2$. Yellow: fitted small circle, Green: great circle found by Geodesic PCA (Huckemann), Red: great circle found by PGA (Fletcher). $\mu$ (PC mean, or geodesic mean) is depicted as yellow (green, or red, respectively) diamond.)
Jung, Foskey & Marron

- Best fit of any circle to data
- Can give better fit than geodesics
- Observed for simulated m-rep example
Currently popular approaches to PCA on $S^k$:

- Early: PCA on projections
- Fletcher: Geodesics through mean

New Approach (Jung, Dryden, Marron): Principal Nested Sphere Analysis
Main Goal:

Extend Principal Arc Analysis \((S^2 \text{ to } S^k)\)
Main Goal:
Extend Principal Arc Analysis \((S^2 \text{ to } S^k)\)

Jung, Dryden & Marron (2012)
Principal Nested Spheres Analysis

Top Down Nested (small) spheres
Digit 3 data: Principal variations of the shape

Princ. geodesics by PNS

Principal arcs by PNS
Main Goal:

Extend Principal Arc Analysis ($S^2$ to $S^k$)

Jung, Dryden & Marron (2012)

Impact on Segmentation:

- PGA Segmentation: used $\sim$20 comp’s
- PNS Segmentation: only need $\sim$13
- Resulted in visually better fits to data
Main Goal:

Extend Principal Arc Analysis \((S^2 \text{ to } S^k)\)

Jung, Dryden & Marron (2012)

Important Landmark: This Motivated Backwards PCA
Key Idea:

Replace usual *forwards* view of PCA

With a *backwards* approach to PCA
Multiple linear regression:

\[ Y_i = \mu + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_k x_{ik} \]

Stepwise approaches:

- **Forwards**: Start small, iteratively add variables to model
- **Backwards**: Start with all, iteratively remove variables from model
Illust’n of PCA View: Recall Raw Data

"Point Cloud View" of Gene Expression
Illust’n of PCA View: PC1 Projections
Illust’n of PCA View:   PC2 Projections

Projections on PC 2 Direction

Gene 3

Gene 2

Gene 1
Illust’n of PCA View: Projections on PC1,2 plane
Replace usual *forwards* view of PCA

Data → PC1 (1-d approx)
→ PC2 (1-d approx of Data-PC1)
→ PC1 U PC2 (2-d approx)
  .
  .
→ PC1 U ... U PCr
(r-d approx)
**Principal Nested Spheres Analysis**

With a *backwards* approach to PCA

Data $\rightarrow$ PC1 $\cup$ ... $\cup$ PC$r$ $(r-d \text{ approx})$

$\rightarrow$ PC1 $\cup$ ... $\cup$ PC$(r-1)$

$\rightarrow$ PC1 $\cup$ PC2 $(2-d \text{ approx})$

$\rightarrow$ PC1 $(1-d \text{ approx})$
An Interesting Question

How generally applicable is Backwards approach to PCA?
An Interesting Question

How generally applicable is *Backwards* approach to PCA?

Where is this already being done???
An Interesting Question

How generally applicable is _Backwards_ approach to PCA?

Potential Application: **Principal Curves**

Hastie & Stuetzle, (1989) _JASA_

(Foundation of **Manifold Learning**)
1\textsuperscript{st} Principal Curve

Linear Reg’n

Usual Smooth
1st Principal Curve

Linear Reg’n

Proj’s Reg’n

Usual Smooth
$1^{\text{st}}$ Principal Curve

Linear Reg’n

Proj’s Reg’n

Usual Smooth

Princ’l Curve
An Interesting Question

How generally applicable is *Backwards* approach to PCA?

Potential Application: **Principal Curves**

Perceived Major Challenge: How to find 2\textsuperscript{nd} Principal Curve? *Backwards* approach???
How generally applicable is *Backwards* approach to PCA?

Another Potential Application: **Nonnegative Matrix Factorization**

= PCA in Positive Orthant

Current Approach, Lee et al (1999):

*Not Nested*, \( k = 3 \nlessapprox k = 4 \)
An Interesting Question

How generally applicable is *Backwards* approach to PCA?

Another Potential Application:

**Nonnegative Matrix Factorization**

= PCA in Positive Orthant

(Backwards **Nested** Approach: Lingsong Zhang)
An Interesting Question

How generally applicable is *Backwards* approach to PCA?

Another Potential Application: **Trees as Data**

(early days)
An Interesting Question

How generally applicable is *Backwards* approach to PCA?

An Attractive Answer
An Interesting Question

How generally applicable is Backwards approach to PCA?

An Attractive Answer:

James Damon, UNC Mathematics

Geometry

Singularity

Theory
How generally applicable is *Backwards* approach to PCA?

An Attractive Answer:
James Damon, UNC Mathematics

Key Idea: Express Backwards PCA as Nested Series of Constraints
Define *Nested Spaces* via **Constraints**

Satisfying More Constraints $\Rightarrow$

$\Rightarrow$ Smaller Subspaces
Define *Nested Spaces* via **Constraints**

E.g. SVD

(Singular Value Decomposition =

= Not Mean Centered PCA)

(notationally very clean)
Define *Nested Spaces* via *Constraints*

E.g. SVD

Have \( k \) Nested Subspaces:

\[
S_1 \subseteq S_2 \subseteq \cdots \subseteq S_d
\]
Define *Nested Spaces* via **Constraints**

E.g. SVD

\[ S_k = \{ x : x = \sum_{j=1}^{k} c_j u_j \} \]

\( k \)-th SVD Subspace
Scores
Loading Vectors
Define *Nested Spaces* via Constraints

E.g. SVD

\[ S_k = \{ x : x = \sum_{j=1}^{k} c_j \overrightarrow{u_j} \} \]

Now Define:

\[ S_{k-1} = \{ x \in S_k : \langle x, \overrightarrow{u_k} \rangle = 0 \} \]
Define *Nested Spaces* via Constraints

E.g. SVD  \[ S_k = \{x : x = \sum_{j=1}^{k} c_j \vec{u}_j \} \]

Now Define:
\[ S_{k-1} = \{x \in S_k : \langle x, \vec{u}_k \rangle = 0 \} \]

**Constraint** Gives *Nested* Reduction of Dim’n
General View of Backwards PCA

Define *Nested Spaces* via **Constraints**

- Backwards PCA

  Reduce Using Affine Constraints
General View of Backwards PCA

Define *Nested Spaces* via **Constraints**

- Backwards PCA
- Principal Nested Spheres

Use Affine Constraints (Planar Slices)

In Ambient Space
Define *Nested Spaces* via **Constraints**

- Backwards PCA
- Principal Nested Spheres
- Principal Surfaces

Spline Constraint Within Previous?
Define *Nested Spaces* via **Constraints**

- Backwards PCA
- Principal Nested Spheres
- Principal Surfaces

*Spline Constraint Within Previous???

{Been Done Already???}
Define *Nested Spaces* via **Constraints**

- Backwards PCA
- Principal Nested Spheres
- Principal Surfaces
- Other Manifold Data Spaces

Sub-Manifold Constraints??

(Algebraic Geometry)
Define *Nested Spaces* via *Constraints*

- Backwards PCA
- Principal Nested Spheres
- Principal Surfaces
- Other Manifold Data Spaces
- Tree Spaces

Suitable Constraints???
General View of Backwards PCA

Why does Backwards Work Better?
General View of Backwards PCA

Why does Backwards Work Better?

- Natural to *Sequentially Add Constraints*

  (I.e. Add Constraints, Using Information in Data)
Why does **Backwards** Work Better?

- Natural to *Sequentially Add Constraints*
- Hard to *Start With Complete Set*, And Sequentially Remove
OODA is more than a “framework”

It Provides a Focal Point

Highlights Pivotal Choices:

*What should be the Data Objects?*

*How should they be Represented?*