Three-dimensional Radial Visualization of High-dimensional Continuous or Discrete Datasets

Fan Dai, Yifan Zhu and Ranjan Maitra

Department of Statistics
Iowa State University
{fd43,yifanzhu,maitra}@iastate.edu

Motivation

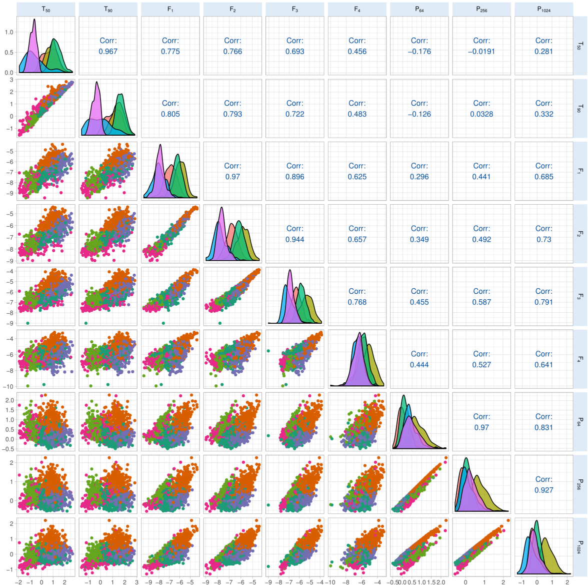
- Multivariate datasets
  - agriculture, engineering, genetics, social science...
- Complex data structure
  - datasets with many discrete, skewed or correlated features
    - image, voice, surveys...
    - need advanced methods for analysis and summaries
- Display distinct groups while also inherent variability

Example: Gamma Ray Bursts (GRBs)

- Extremely energetic explosions observed in distant galaxies.
  - data from NASA's Burst and Transient Source Experiment
  - 1,599 GRBs with complete information on 9 parameters
    - time for % flux to arrive, peak fluxes in different channels, time-integrated fluences over time-points
- Nine heavily-skewed "parameters" or attributes
  - use of logarithms to reduce skewness
- astrophysics community argued long over 2 or 3 types
  - analysis based on summary exclusion of some heavily-correlated attributes
  - recent analysis shows all 9 features important for clustering
    - actually 5 ellipsoidal groups, not 2 or 3
- smaller-dimensional 9D example used as a test case

- Visualization tools for continuous multivariate data
  - pairwise scatter plots

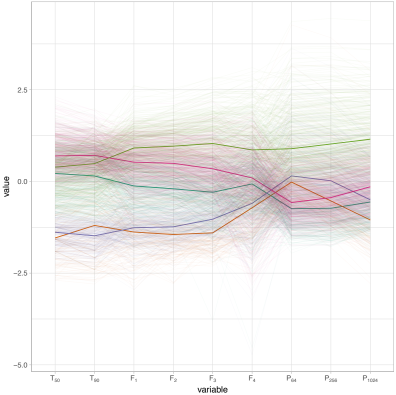
Pairwise Scatterplots: Gamma Ray Bursts



Background and Current Work

- Visualization tools for continuous multivariate data
  - pairwise scatter plots
    - limited in providing multivariate assessments
  - parallel coordinates plot (*Inselberg '85, Wegman '90*)

Parallel Coordinate Plots: Gamma Ray Bursts



- Represent multidimensional data using lines.
  - vertical line represents each dimension or attribute.
  - $p - 1$ lines connected at appropriate scaled dimensional value represent each observation
  - polar version provided by star plot

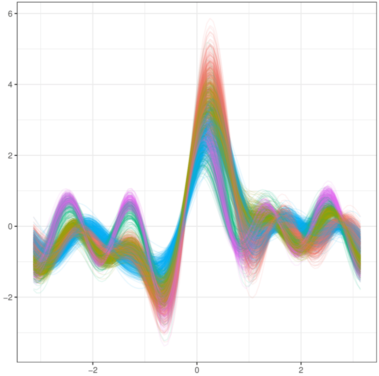
Background and Current Work

- Many approaches to display continuous multivariate data
  - pairwise scatter plots
    - limited in providing multivariate assessments
  - parallel coordinates plot (*Inselberg '85, Wegman '90*)
    - placement order matters, unclear for large n, p
    - hard to identify groups/patterns with even moderate n.
  - Andrews' curves represent each observation via trigonometric series

Andrews' Curves: Gamma Ray Bursts

- Plot each $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ as a curve:

$$f(t) = x_1 + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \ldots, \qquad t \in [-\pi, \pi]$$



- Entire curve displays one observation

Background and Current Work

- Many approaches to display continuous multivariate data
  - pairwise scatter plots
    - limited in providing multivariate assessments
  - parallel coordinates plot (*Inselberg '85, Wegman '90*)
    - placement order matters, unclear for large n, p
    - polar version provided by star plot
  - Andrews' curves
    - order in which coordinate enters series important
    - very computationally intensive for larger p
  - Star coordinates plot

    represents coordinate axes as equi-angled rays extending from center
    - order matters, optimized (*van Long & Linsen '11*)
- Use springs to display observation (radial visualization)

Two-dimensional radial visualization (RadViz2D)

- Uses Hooke's law to project data onto unit circle
  - place *p* springs (anchor points) on the rim
    - pull each spring by value relative to coordinate from center
    - observations w/ similar relative values in all attributes end up closer to center, others are closer to the anchor points
  - order of placement of springs affects display
    - refinements to improve RadViz2D exist (see later)

RadViz2D Illustration

$\boldsymbol{X} = (X_1, X_2, X_3, X_4, X_5) = (0.7, 0.5, 0.3, 0.2, 0.7)$

- Maps $\boldsymbol{X} \in \mathbb{R}^p$ to 2D point $\boldsymbol{\Psi}^{\bullet}(\boldsymbol{X}; \boldsymbol{U}) = \boldsymbol{U}\boldsymbol{X} / \mathbf{1}_p' \boldsymbol{X}$:
  $\boldsymbol{U}$ projection matrix, columns (anchor points) on $\mathbb{S}^1$

Two-dimensional radial visualization (RadViz2D)

- Uses Hooke's law to project data onto unit circle
  - place *p* springs (anchor points) on the rim
    - pull each spring by value relative to coordinate from center
    - observations w/ similar relative values in all attributes end up closer to center, others are closer to the anchor points
  - order of placement of springs affects display
    - refinements to improve RadViz2D exist (see later)
- Effective for sparse data, in evaluating distinct groups
  - Nonlinear map distorts, affects interpretability
  - High-dimensional observations more difficult to visualize
- Can fully 3D extension improve performance?
  - Viz3D provides third dimension, constant for all observations (*Artero & de Oliveira, '04*)

Generalizing Radial Visualization

- Allow anchor points in $\boldsymbol{U}$ on $\mathbb{S}^q$, $q > 1$, not necessarily equi-spaced
  - p springs at $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_p \in \mathbb{S}^q$, with spring constants $X_1, X_2, \ldots, X_p$.
  - equilibrium point $\boldsymbol{Y} \in \mathbb{R}^{q+1}$ of system satisfies

$$\sum_{j=1}^{p} X_j(\boldsymbol{Y} - \boldsymbol{u}_j) = 0,$$

    - $\boldsymbol{Y} = \boldsymbol{\Psi}(\boldsymbol{X}; \boldsymbol{U}) = \boldsymbol{U}\boldsymbol{X}/\boldsymbol{1}_p'\boldsymbol{X}$ solves the system.
  - is line-, point-ordering- and convexity-invariant.
  - scaling every coordinate to be in [0,1] allows for $\boldsymbol{Y} \in \mathbb{S}^q$.

Placement of Anchor Points

- Suppose: coordinates of $\boldsymbol{X}$ are uncorrelated.
- For $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^p$, let $\boldsymbol{Y}_i = \boldsymbol{\Psi}(\boldsymbol{X}_i; \boldsymbol{U}), i = 1, 2$.
  - Euclidean distance between $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ is

$$\|\boldsymbol{Y}_1 - \boldsymbol{Y}_2\|^2 = \left(\frac{\boldsymbol{X}_1}{\mathbf{1}_p'\boldsymbol{X}_1} - \frac{\boldsymbol{X}_2}{\mathbf{1}_p'\boldsymbol{X}_2}\right)' \boldsymbol{U}'\boldsymbol{U} \left(\frac{\boldsymbol{X}_1}{\mathbf{1}_p'\boldsymbol{X}_1} - \frac{\boldsymbol{X}_2}{\mathbf{1}_p'\boldsymbol{X}_2}\right),$$

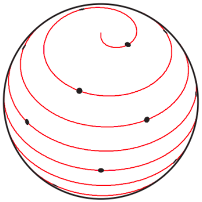
    - $\boldsymbol{X}_i$, $\boldsymbol{X}_j$ very dissimilar, with perfect negative correlation, should be placed as far away as possible (in opposite directions) in our radial visualization.
  - However, $\|\boldsymbol{Y}_i - \boldsymbol{Y}_j\|^2 \to 0$ as $\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle \to 0$.
    - may create artificial visual correlation between ith and jth coordinates if $\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle \to 0 < \pi/2$.
    - need $\boldsymbol{u}_j$s far from the other as possible; so evenly distributed.
    - $\mathbb{S}^q$: for larger q, can get larger angles between $\boldsymbol{u}_j$s
  - Also place positively correlated coordinates close together
    - $q > 1$ has advantage in placing multiple coordinates together

Three-dimensional Radial Visualization

- $q = 2$ in our generalization yields RadViz3D:
  - equi-spaced anchor points for 5 Platonic solids, $p = 4, 6, 8, 12, 20$.
    - closely related to Thomson problem in traditional molecular quantum chemistry (Atiyah & Sutcliffe '03).
  - for other p, approximate through Fibonacci grid, jth anchor point:

$$u_{j1} = \cos(2\pi j \varphi^{-1})\sqrt{1 - u_{j3}^2},$$
$$u_{j2} = \sin(2\pi j \varphi^{-1})\sqrt{1 - u_{j3}^2},$$
$$u_{j3} = \frac{2j-1}{p} - 1,$$

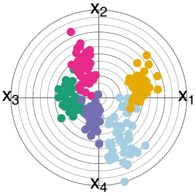
where $\varphi = (1 + \sqrt{5})/2$ is the golden ratio.



(González '10)

- distributes anchor points along generative spiral on $\boldsymbol{S}^2$, with consecutive points as separated as possible, satisfies "well-separation" property (Saff & Kuijlaars '97).

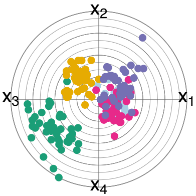
4D Examples simulated via *MixSim* package in R



RadViz2D, $\ddot{\omega} = 10^{-3}$

Viz3D

RadViz3D



RadViz2D,$\ddot{\omega} = 10^{-2}$

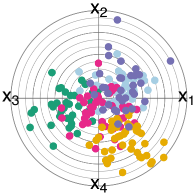
Higher-dimensional Datasets

- Display p anchor points infeasible, even for moderate p
  - placement of equally-spaced anchor points built on not inducing spurious positive correlations in display
    - with increasing p, harder to guarantee such outcome
- Project high-dimensional data to uncorrelated coordinates but preserve distinctiveness and variability in groups
  - Principal Components finds mutually orthogonal projections summarizing proportion of total variance, but does not account for groups.

Maximum-Ratio Projection (MRP)

- **Step 1:** Obtain PCs (orthogonal $\boldsymbol{V}_g$) for each group
  - Find orthogonal $\boldsymbol{W}$ closest to all $\boldsymbol{V}_g$
    - Project $\boldsymbol{X}$ with $\boldsymbol{W}$ and then obtain MRP
- **Step 2:** Obtain uncorrelated projections that maximize between-group sums of squares and cross products (SSCP) relative to the total SSCP.
  - Let $\boldsymbol{T}$, $\boldsymbol{W}$ be (p.d.) total & between-group corrected SSCP.
    - $\hat{\boldsymbol{v}}_j = \boldsymbol{T}^{-\frac{1}{2}}\hat{\boldsymbol{w}}_j / \|\boldsymbol{T}^{-\frac{1}{2}}\hat{\boldsymbol{w}}_j\|, j = 1, 2, \ldots, k$, $\hat{\boldsymbol{w}}_j, j = 1, 2, \ldots, k$ are, in decreasing order, the k largest eigenvalues of $\boldsymbol{T}^{-1/2}\boldsymbol{B}\boldsymbol{T}^{-1/2}$.
    - $k \leq G - 1$, chosen by scree plot/quality of display
    - $G \leq 4$ needs $4 - G + 1$ more projections w/ null contribution
    - needs p.d. $\boldsymbol{T}$, does not hold if $p > \min n_g$
- MRP maximizes separation between groups (in projected space) relative to total variability.

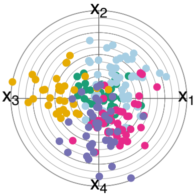
500D Examples



RadViz2D, $\ddot{\omega} = 10^{-3}$

Viz3D

RadViz3D

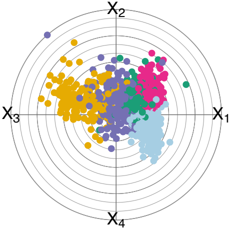


RadViz2D, $\ddot{\omega} = 10^{-2}$

Datasets with Skewed Attributes

- Consider a r.v. X with CDF $F_X(x)$.
  - $F_X(X) \sim U(0,1) \Rightarrow Y = \Phi^{-1}[F_X(X)] \sim N(0,1)$.
    - call the above (classical) Gaussianized Distributional Transform (CGDT)
    - marginal application of CGDT specifies distribution on $\boldsymbol{X}$ with desired marginal and correlation structure.
- CGDT standardizing transform, more stringent than usual affine 0-mean, unit-variance inducing transform
  - CGDT matches all marginal quantiles to N(0,1)
  - Apply to skewed datasets or with unclear marginals
- Before applying MRP and RadViz3D

Applications: Gamma Ray Bursts Dataset



RadViz2D

Viz3D

RadViz3D

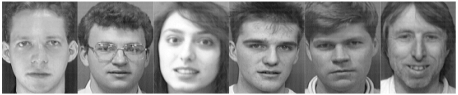
Groups ● 1 ● 2 ● 3 ● 4 ● 5

- Heavily skewed attributes, so CGDT appropriate
- Results indicate 5 overlapping clusters
  - some suggestion of 2, 3 super-types of GRBs

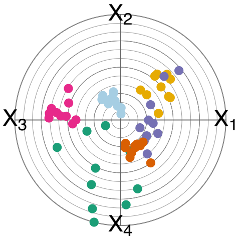
Applications: Face Recognition

- 112×92-images of 6/40 faces at 10 light angles/conditions.
- (20×14) DWT2 (LL band) of wavelet-transformed images with 280 features (Jadhav & Holambe, 2009)

Applications: Face Recognition



Persons ● A ● B ● C ● D ● E ● F



RadViz2D

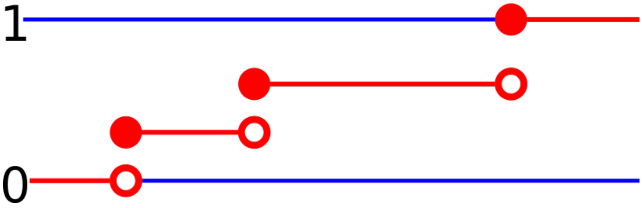
Viz3D

RadViz3D

- marginals unclear: use CGDT
- RadViz3D clarifies all 6 people the best

Datasets with Discrete Attributes

- For discrete-valued variable X, CDF $F_X(X) \nsim U(0,1)$ because of discreteness.
  - CGDT currently not applicable



- Note that the CDF is only right continuous
- Solution proposed by Rüschendorf (2013) via the generalized distributional transform

Generalized Distributional Transform (GDT)

Definition

Let X be a real-valued RV with CDF $F_X(\cdot)$ and let $V \sim U(0,1)$ be a RV independent of X. The generalized distributional transform of X is $U = F(X, V)$ where $F(x, \lambda) \doteq P(X < x) + \lambda P(X = x) = F_X(x-) + \lambda[F_X(x) - F_X(x-)]$ is the generalized CDF of X.

Theorem

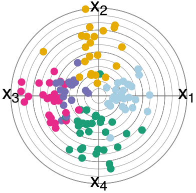
Let $U = F(X, V)$ be the generalized distributional transform of X. Then

$$U \sim \mathrm{Uniform}(0,1) \text{ and } X = F_X^{-1}(U) \text{ a.s.}$$

where $F^{-1}(t) = \inf\{x \in \mathbb{R} : F_X(x) \geq t\}$ is the generalized inverse, or the quantile transform, of $F_X(\cdot)$.

- Use $F(X, V)$ in place of $F_X(X)$, calculate GDT as before
    - use of GDT on non-discriminating coordinate can spuriously bestow it hyper-importance
        - suggest ANOVA test on each GDT-ed coordinate, control FDR

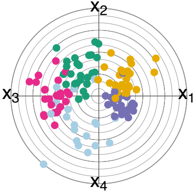
Illustration: Simulated Binary Datasets



RadViz2D, low clustering complexity

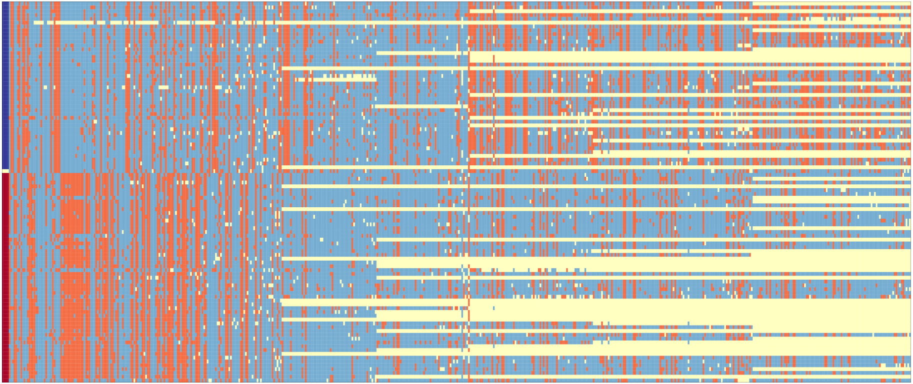
Viz3D

RadViz3D



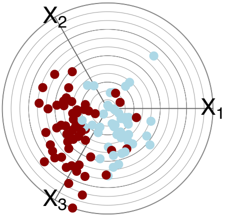
RadViz2D, high clustering complexity

Applications: Senate Voting Records



- 108th US Congress (2005-06) had 542 (Y/N/NR) Senate votes
  - 55 Republicans, 44 Democrats, 1 (D-caucus) Independent (VT) (Banerjee *et al*, 2008)
    - combine N/NR to get dataset of binary attributes

Applications: Senate Voting Records



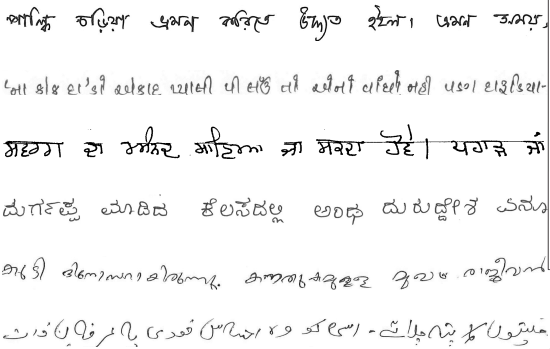
RadViz2D

Viz3D RadViz3D

● Democratic ● Republican

- $G = 2$ so only 1 MRP with postive eigenvalue
  - spring X_1 pulls members of one party towards itself more
  - X_2, X_3, X_4 pull senators from both parties with equally (non-discriminating) force

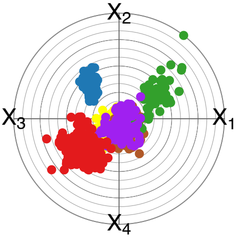
Applications: Handwritten Indic Scripts



(Map Acknowledgment: Surveyor-General of India)

- Handwritten scripts from Bangla (east), Gujarati (west), Gurmukhi (north), Kannada and Malayalam (southern states of Karnataka and Kerala), Urdu (Persian script), with 116 mixed features (Obaidullah et al, 2017).

Applications: Handwritten Indic Scripts



RadViz2D

Viz3D RadViz3D

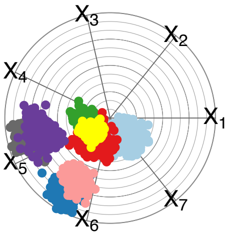
● Bangla ● Gujarati ● Gurmukhi ● Kannada ● Malayalam ● Urdu

- Viz3D (lesser extent RadViz2D) separates Urdu, Kannada and Gujarati, not the other 3 languages
- RadViz3D best in classifying all the 6 scripts
  - also points to difficulty of problem

Applications: *RNA* Sequences

- Gene expression levels, in FPKM, of RNA sequences from 13 human organs.
  - focus on 8 largest (in terms of the sample size) organs
    - esophagus (659), colon (339), thyroid (318), lung (313), breast (212), stomach (159), liver (115) and prostate (106)
  - p=20242 discrete features
    - some have many discrete values, essentially continuous
  - dataset of mixed attributes.
- Display for distinctiveness of samples from each organ

Applications: RNA Sequences



RadViz2D

Viz3D

RadViz3D

● Breast ● Colon ● Esophagus ● Liver ● Lung ● Prostate ● Stomach ● Thyrioid

- RadViz2D, Viz3D poorer at separating organs
- RadViz3D indicates clear separation between organs
  - colon and stomach have some marginal overlap.

Conclusions and Further Work

- Visualization tool for HD datasets
  - RadViz3D for more comprehensive display of grouped data
  - MRP, GDT for discrete, mixed, skewed variates
  - displays distinct groups more accurately
  - R package `https://github.com/fanne-stat/radviz3d`
  - manuscript `https://arxiv.org/abs/1904.06366`
- Number of issues merit further attention
  - MRP linear; non-linear projections better?
  - extend for categorical (non-binary) attributes
  - GDT/MRP with other tools for improved visualization