Exploratory Projection Pursuit

Katrin Newger

Department for Statistics
Ludwig-Maximilians-University of Munich

January 15, 2014
Table of Contents

1. Introduction

2. Exploratory Projection Pursuit
   - The Projection Index and the Algorithm
   - Examples of Indices
   - Robust Projection Pursuit

3. Application to Two Data Sets
   - Swiss Banknotes
   - Near-Infrared Spectra of Gasolines

4. Conclusion

5. References
Table of Contents

1 Introduction

2 Exploratory Projection Pursuit
   - The Projection Index and the Algorithm
   - Examples of Indices
   - Robust Projection Pursuit

3 Application to Two Data Sets
   - Swiss Banknotes
   - Near-Infrared Spectra of Gasolines

4 Conclusion

5 References
Recall: Principal Component Analysis

- Data set is reduced to its Principal Components
- Focus on total variance in the original data
- Coefficients easily calculated by eigenvalues and eigenvectors
- Downside: limitedness in application, e.g. with clustering or outliers

Alternative: Projection Pursuit (PP)
Start Off

How to tackle bigger and more complex data sets?

- First solutions using **Projection Pursuit** in 1969 by Kruskal and later by Friedman & Tukey
- PP reduces high-dimensional data sets to **1 to 3 dimensions**
- Focus on **non-linear structure** to find clusters etc.
- Methods: Projection Pursuit Density Estimation, Projection Pursuit Regression and **Exploratory Projection Pursuit** (EPP)
The Main Goal
Reduce high-dimensional data sets

- Kind of ‘generalisation’ of PCA
- Greatest advantage: Operator defines interesting structure themselves
- Different indices and algorithms
  → More complicated than PCA
1 Introduction

2 Exploratory Projection Pursuit
   • The Projection Index and the Algorithm
   • Examples of Indices
   • Robust Projection Pursuit

3 Application to Two Data Sets
   • Swiss Banknotes
   • Near-Infrared Spectra of Gasolines

4 Conclusion

5 References
The Projection Index (1/2)

- Forms the **heart** of EPP
- Can be understood as **analogon to total variance** in PCA
- Gives data analysis its direction (defines what is **interesting**)
- Is optimised by an **algorithm**
Interesting could be anything!

- Define what is **least interesting**: Multivariate Normal Density
- Index designed to deliver **low** value for **uninteresting** structure and
- To find departure from normality close to the **center** rather than in the tails
Exploratory Projection Pursuit—The Algorithm

Problem

Solution through eigenvalues not applicable

- Attribute of EPP: Rarely just one solution
  → Projection Algorithm has to discover most of these solutions
- Consequence of high-dimensional data: lots of uninformative space
  → Projection Algorithm has to discover informative optima first
- Modern technology enabled computing the algorithms

Very challenging from mathematical point of view
One-Dimensional Index

\[ I(a) = s(a)d(a), \]

where \( s(a) \) is the **spread** and \( d(a) \) the **local density** of the datapoints

- Index designed to reveal **clustering**
- Applicable to 1 or 2 dimensions
- Properties:
  - Very **stable**
  - **Easily applicable** using SAT (Solid Angle Transformation)
Jones & Sibson’s Index

- Jones & Sibson simplified Friedman & Tukey’s Index
- Considered uninteresting projections and maximised divergence from it
- Found that $d(a)$ was just an estimate of $\int f(x)^2 dx$ and that
- $s(a)$ was negligible when using Centering and Sphering
- Designed two important indices evolving from these insights
Usual Order-1 Entropy Basis of Index

\[- \int f(x) \log f(x) \, dx,\]

where \( f \) is the density of the projected data

- Use of **Centering & Sphering**
- Here it was discovered that Standard Normal Density was the **least interesting**
- **Maximising** Entropy Index **difficult**, motivating Jones & Sibson to design another
One-Dimensional Distributional Moment Index

\[ M = \left( k_3^2 + \frac{1}{4} k_4^2 \right) / 12, \]

\( k_3, k_4 \) being the 3\(^{rd}\) and 4\(^{th}\) order cumulants for projected distribution

- **Approximisation** of Entropy Index
- Based on **summary statistics** of data
- Index is rotationally invariant
- Index is **differentiable**
  - → Simplifies optimisation
Friedman’s Index (1/2)

Distributional Version

\[ F(a) = \int_{-\infty}^{\infty} \phi(x)^{-1} f(x)^2 \, dx, \]

with \( \phi \) being the Standard Normal Density

- Disadvantage: For heavy-tailed distributions \( F \) does not point out departure from normality
- Optimisation: Based on Lagrange multipliers (performs well)
Goal
Find as much **interesting** information as possible

Problem
Not self-evident that algorithm finds **most informative** projection first

Friedman’s **solution**:
- After an interesting view is found, **remove** interesting structure
  - Transform interesting parts of projection to Standard Normal Density
Most indices measure **departure from normality**
Index operates with **sphered data** (simplifies design)
First **derivatives** should exist (simplifies optimisation)
Index should be **rotationally invariant, affine invariant** and **computationally efficient**
The index is only as **good** as its **algorithm**!

**Recent Development**
- 3-Dimensional indices
- Measurement of departure from different distributions
- **Robust Projection Pursuit**
PCA combined with Projection Pursuit techniques

- PCA very fragile to outliers, as variance is no robust measure
  → Need of new kind of measure
- Solution through eigenvectors of covariance matrix no longer available
- Instead, solution through approximative algorithms
Two options for robust variance measure:

**Median Absolute Deviation (MAD)**

\[
\text{MAD}(z_1, \ldots, z_n) = 1.48 \text{med}_j \left| z_j - \text{med}_i z_i \right| \quad \text{and}
\]

**First Quartile of the Pairwise Differences between all Data Points (Q)**

\[
Q(z_1, \ldots, z_n) = 2.22 \left\{ |z_i - z_j|; 1 \leq i < j \leq n \right\} \left( \frac{n}{2} \right) / 4,
\]

in both formulae \( \{z_1, \ldots, z_n\} \) being the given data set.
Robust PP—The Croux–Ruiz Algorithm (CR)

- **Good results** especially for *large sample size* relative to number of variables
- **Problem** with *small sample size* relative to number of variables:
  - Algorithm may implode

### Implosion of the Algorithm

From certain order on, all subsequent ‘eigenvalues’ are zero (independent of the data)
Robust PP—The GRID Algorithm

→ Based on CR Algorithm

Advantages

- Problem of imploding is solved
- More precise than CR Algorithm
- All in all better and more stable results
- Easier to implement
- Index doesn’t have to be differentiable
Table of Contents

1 Introduction

2 Exploratory Projection Pursuit
   • The Projection Index and the Algorithm
   • Examples of Indices
   • Robust Projection Pursuit

3 Application to Two Data Sets
   • Swiss Banknotes
   • Near-Infrared Spectra of Gasolines

4 Conclusion

5 References
200 observations of *Swiss banknotes*

### One Discrete Variable $Y$

$$Y = \begin{cases} 
Y = 0 & \text{for genuine banknote} \\
Y = 1 & \text{for counterfeit banknote} 
\end{cases}$$

### Six Metric Variables, all measured in $mm$

- **Length**: Length of bill
- **Left**: Width of left edge
- **Right**: Width of right edge
- **Bottom**: Bottom margin width
- **Top**: Top margin width
- **Diagonal**: Length of image diagonal
Figure: Scatterplot Matrix for all Metric Variables
Two Scatterplots in Detail

Figure: Scatterplot of Right/Bottom

Figure: Scatterplot of Top/Diagonal
Andrews Curve

Figure: Andrews Curve

- Possibility to reveal structure in multivariate data
- Here: Two clusters
Best solution: virtually no departure from Standard Normal Density

Weakly suggesting two clusters

Figure: Density of Worst and Best Projection Compared to Standard Normal Density
Figure: Plot of all Indices
Best solution: Most departure from Standard Normal Density

Worst solution: close to Standard Normal Density

Suggests two clusters: counterfeit and genuine

**Figure**: Density of Worst and Best Projection Compared to Standard Normal Density
Application of Jones & Sibson’s Index (2/2)

Figure: Plot of all Indices

- Worst
- Best
Table of Contents

1 Introduction

2 Exploratory Projection Pursuit
   • The Projection Index and the Algorithm
   • Examples of Indices
   • Robust Projection Pursuit

3 Application to Two Data Sets
   • Swiss Banknotes
   • Near-Infrared Spectra of Gasolines

4 Conclusion

5 References
The Gasoline Data Set

→ 60 observations of near-infrared spectra of gasolines

401 Metric Variables
Wavelength from 900 to 1700 nm in 2-nm-intervals

High-dimensional data set with small sample size relative to number of variables
Near-Infrared Spectra of Gasolines

Figure: Spectra of all 60 Variables

- Impossible to filter single variables
- Extremely high correlation between all variables
- Many peaks
Comparison of GRID and CR Algorithm (MAD)

GRID always above CR
CR Algorithm seems to implode

**Figure**: Measurement with MAD for 20 Components
Comparison of GRID and CR Algorithm ($Q$)

Figure: Measurement with $Q$ for 20 Components

Again: GRID always above CR

Better performance of GRID Algorithm for both MAD- and $Q$-measure
Table of Contents

1 Introduction

2 Exploratory Projection Pursuit
   - The Projection Index and the Algorithm
   - Examples of Indices
   - Robust Projection Pursuit

3 Application to Two Data Sets
   - Swiss Banknotes
   - Near-Infrared Spectra of Gasolines

4 Conclusion

5 References
EPP’s Relevance in Data Analysis

- EPP very **good alternative to PCA**
- Helps discover **phenomena** in the data
- Yields very **comprehensive results**
- Becoming more and more **important** in analysis of **current research**
- Also **combinations** of EPP and PCA effective, see Robust Projection Pursuit
Advantages & Disadvantages:

- **Flexible** because of ability to use different indices
- Through this, **applicability to great variety** of data sets
- Sometimes **hard to interpret** the results
- **Implementation difficult** to conduct (no special R-package, etc.)
Table of Contents

1 Introduction

2 Exploratory Projection Pursuit
   • The Projection Index and the Algorithm
   • Examples of Indices
   • Robust Projection Pursuit

3 Application to Two Data Sets
   • Swiss Banknotes
   • Near-Infrared Spectra of Gasolines

4 Conclusion

5 References
Bartholomew, D. J., F. Steele, I. Moustaki, and J. I. Galbraith
*The Analysis and Interpretation of Multivariate Data for Social Scientists.*

Croux, C., P. Filzmoser, and M. R. Oliveira
Algorithms for projection-pursuit robust principal component analysis.
*Katholieke Universiteit Leuven, Faculty of Economics and Applied Economics, Department of Decision Sciences and Information Management,* 2006.

Croux, C. and A. Ruiz-Gazen
*Journal of Multivariate Analysis* 95, 2005.
Friedman, J. H.
Exploratory projection pursuit.

Friedman, J. H. and J. W. Tukey
A projection pursuit algorithm for exploratory data analysis.

Hall, P.
On polynomial-based projection indices for exploratory projection pursuit.

Huber, P. J.
Projection pursuit.
Jones, M. C. and R. Sibson
What ist projection pursuit?

Li, G. and Z. Chen

Nason, G. P.
Design and choice of projection indices.

Posse, C.
Tools for two-dimensional exploratory projection pursuit.