Peeling Algorithm in Financial Risk Analysis

Aijun Zhang

University of Michigan and Bank of America

October, 2008

A joint work with Agus Sudjianto, Bank of America
Introduction

Peeling Methodology
- Principal Direction of Anomaly
- MD-based Peeling Algorithm

Radar-chart Visualization
- CBSA GeoRisk
- Tracking Financial Storm

Discussion
Introduction

- **Univariate outlier detection:** e.g. Box-plot, QQ-plot


\[
\begin{align*}
&\gg \text{boxplot(zscore(X), 'PARAM', val,...);} \\
&\gg \text{qqplot(zscore(X(:,j)))};
\end{align*}
\]

- **Multivariate outlier detection:** Mahalanobis distance

\[
MD_i(\mu, \Sigma) = (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)
\]

1. Often used are the sample mean $\bar{x}$ and covariance $\hat{\Sigma}$
2. $MD \sim \chi^2_p$ asymptotically, for $x \sim N_p(\mu, \Sigma)$
3. However, problem with masking and swamping effects ...
Univariate case: Box-plot and QQ-plot

CSBA Economic Indicators, Aug 2008

Variable: BKs per US Pop

Standard Normal Quantiles
Multivariate case: Masking and Swamping
Our development

- We propose a sequential method, called the peeling algorithm
  1. Reasoning from projection pursuit
  2. MD-based peeling algorithm
  3. Visualization by polar coordinates

- Ideas mostly originated from real applications in financial risk
  a. Anti-Money Laundering project
  b. CBSA GeoRisk visualization
  c. Recent storm from Wall Street

- Examples will be provided throughout the talk ...
Introduction

Peeling Methodology
Principal Direction of Anomaly
MD-based Peeling Algorithm

Radar-chart Visualization
CBSA GeoRisk
Tracking Financial Storm

Discussion
Problem Setup

**Sphering:** For \( \{x_i\}_{i=1}^n \) in \( \mathbb{R}^p \), consider the “sphered” data,

\[
    z_i = \Sigma^{-1/2}(x_i - \bar{x}), \quad i = 1, \ldots, n
\]

given any \( \Sigma > 0 \) (positive definite).

**Projection:** For \( w \in \mathbb{R}^p \) with \( ||w|| = 1 \), the “projected” data

\[
    \{w^T z_1, \ldots, w^T z_n\}, \text{ in 1-D.}
\]

**Question:** What is the best direction \( w^* \) that would separate clearly the outlying observations in 1-D space?
Suppose $\exists 100\alpha\%$ outliers, define the separation $\text{Score}(w, \alpha)$ by

$$
\sum_{i=1}^{n} \left\{ \frac{1}{n\alpha} (w^T z_i - q_{w,\alpha}) + \frac{1}{n(1-\alpha)} (w^T z_i - q_{w,\alpha}) \right\}
$$

where $q_{w,\alpha} = (1 - \alpha)$-th quantile of projected data. Then,

\[ PDA: \quad w_{\alpha}^* = \arg \max_{||w||=1} \text{Score}(w, \alpha), \quad \alpha \in (0, 0.5] \]
Theorem

Given the projected data $\mathcal{D}_0 = \{z_1, \ldots, z_n\}$, let $\mathcal{D} \subset \mathcal{D}_0$, then

$$\text{Score}(w, \alpha) = \frac{1}{n\alpha(1 - \alpha)} \left\{ \max_{|\mathcal{D}| = \lceil n\alpha \rceil} \sum_{z \in \mathcal{D}} w^T z - \{n\alpha\} q_{w,\alpha} \right\}.$$

For $\alpha = j/n$ with $j = 1, \ldots, \lfloor n/2 \rfloor$, the score is bounded from above by

$$\text{Score}(w, j/n) \leq \frac{n}{n-j} \|\bar{z}^*_{(1:j)}\|,$$

where

$$\bar{z}^*_{(1:j)} = \frac{1}{j} \sum_{z \in \mathcal{D}^*} z, \quad \mathcal{D}^* = \arg \max_{|\mathcal{D}| = j} \left\| \sum_{z \in \mathcal{D}} z \right\|.$$

The maximum score is attained by the PDA $w^*_j/n \propto \bar{z}^*_{(1:j)}$. 
Corollary 1: Set \( \alpha = 1/n \) and let \( z^* \leftarrow \max_i \|z_i\| \) with maximal Euclidean distance. Then, the PDA is given by

\[
\mathbf{w}_{1/n}^* = \arg \max_{\|\mathbf{w}\|^2 = 1} \text{Score}(\mathbf{w}, 1/n) = z^*/\|z^*\|
\]

Corollary 2: Based on the raw data \( \mathcal{D}_0 = \{x_i\}_{i=1}^n \), the separation score

\[
\text{Score}(\mathbf{w}, j/n; \Sigma) = \frac{n}{j(n-j)} \max_{|\mathcal{D}|=j} \sum_{x \in \mathcal{D}} \mathbf{w}^T \Sigma^{-1/2} (x - \bar{x}), \quad \mathcal{D} \subset \mathcal{D}_0
\]

The PDA is given by \( \mathbf{w}_{j/n}^* \propto \Sigma^{-1/2}(\bar{x}^{(1:j)} - \bar{x}) \), where \( \bar{x}^{(1:j)} = \frac{1}{j} \sum_{x \in \mathcal{D}^*} x \) and \( \mathcal{D}^* = \arg \max_{|\mathcal{D}|=j} \left\| \Sigma^{-1/2} \sum_{x \in \mathcal{D}} (x - \bar{x}) \right\| \).

For \( \alpha = 1/n \) and \( \Sigma = \hat{\Sigma} \), let \( x^* \) attain the maximal Mahalanobis distance. Then

\[
PDA: \quad \mathbf{w}_{1/n}^* = \hat{\Sigma}^{-1/2} (x^* - \bar{x}) / \sqrt{\text{MD}(x^*)}
\]
Peeling Algorithm

- **One-by-one procedure**: detect one outlier every step, remove it before proceeding to next step.

- **Masking/swamping immunity**: the “intermediate” observations are likely affected by the extreme ones, but not vice versa.

- The recursive MD-based algorithm is simple to understand, and easy to implement; see Corollary 2.
Peeling Algorithm

MD-based Peeling Algorithm: input $\alpha_0$ (0.5 by default)

1. Initialize $\mathcal{D}_1 = \mathcal{D}_0 = \{x_i\}_{i=1}^{n}$ and $k = 0$

2. Compute Mahalanobis distance (MD) for sample $\mathcal{D}_1$; find one of the elements with max MD

   $$\text{MD} = \text{mahal}(D1,D1); \ [\text{maxMD}, \text{outId}] = \text{max}(\text{MD});$$

3. If $k < n\alpha_0$, flag $D_1_{\text{outId}}$ as outlier, update $k + 1 \rightarrow k$, $\mathcal{D}_1 \setminus \text{outId} \rightarrow \mathcal{D}_1$ and go back to Step 2.

{4} Determine the best $\alpha \in (0, \alpha_0]$ based on MD-histogram; output the indices of the corresponding $n\alpha$ outliers.
>> alpha0 = 0.2;
>> idx = peel(zscore(X), alpha0);
>> Ticker(idx)'
ans = 'WB' 'SOV' 'NCC' 'SFI' 'GNW'

Example: 2-D stock returns (financial sector) on Sep 29-30
Introduction

Peeling Methodology
Principal Direction of Anomaly
MD-based Peeling Algorithm

Radar-chart Visualization
CBSA GeoRisk
Tracking Financial Storm

Discussion
Radar-chart Visualization

- For each suspicious subject \( i \), the peeling algorithm gives us
  
  (a) Mahalanobis distance: \( D_i = \sqrt{MD_i} \) (scalar)
  
  (b) Outlying direction: \( w_i \) s.t. \( \|w\| = 1 \) (spherical)

- Radar-chart visualization is a natural choice, by converting

  \[
  D_i \rightarrow \text{Radius}, \quad w_i \rightarrow \text{Radian (angle)}
  \]

- Trivial case if \( w \in \mathbb{R}^2 \):
  
  \[
  \text{>> theta} = \text{acos}(W(:,1)).*\text{sign}(W(:,2));
  \]

- Nontrivial if \( w \) is high-dimensional. We need dimension reduction techniques, e.g. MDS (multidimensional scaling)
Robust PC1 and PC2 are used as reference coordinates

Better choices are under development ⇒ to report later
For $i = 1, \ldots, 288$ (Financial firms included in DJUSFN and KBW)

$$R_i(t) = \alpha_{ki} + \beta_{ki}R_0(t) + \varepsilon_{ki}(t), \quad t \in [\tau_{k-1}, \tau_k]$$

Portfolio anomaly detection via **idiosyncratic residuals** $\hat{\varepsilon}_{ki}$
Black September - Radar Tracking

Time permitting, show the animated radar chart in Matlab ...
Introduction

Peeling Methodology
Principal Direction of Anomaly
MD-based Peeling Algorithm

Radar-chart Visualization
CBSA GeoRisk
Tracking Financial Storm

Discussion
Discussion

- A whole family of interesting problems are being investigated:
  1. Robust estimate of location and scale, e.g. MCD (minimum covariance determinant) estimator
  2. Peeling-based projection pursuit, e.g. robust PCA
  3. Spherical clustering: Hierarchical linkage, K-means, the mixture vMF (von Mises-Fisher) model
  4. Multidimesional scaling onto polar coordinates

- We look forwards to more applications in financial risk analysis