# 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

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

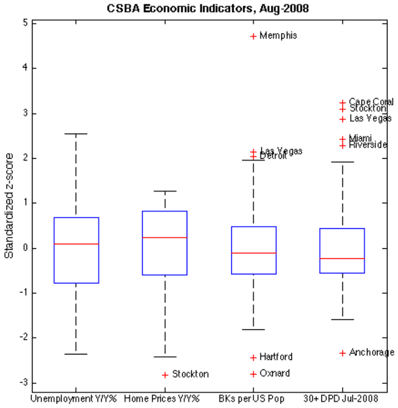
```
>> boxplot(zscore(X), 'PARAM', val,...);
>> qqplot(zscore(X(:,j)));
```

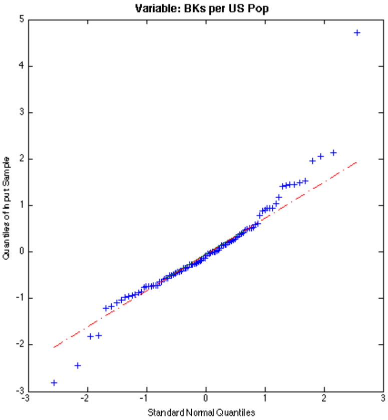
- **Multivariate outlier detection:** Mahalanobis distance

$$\mathrm{MD}_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})$$

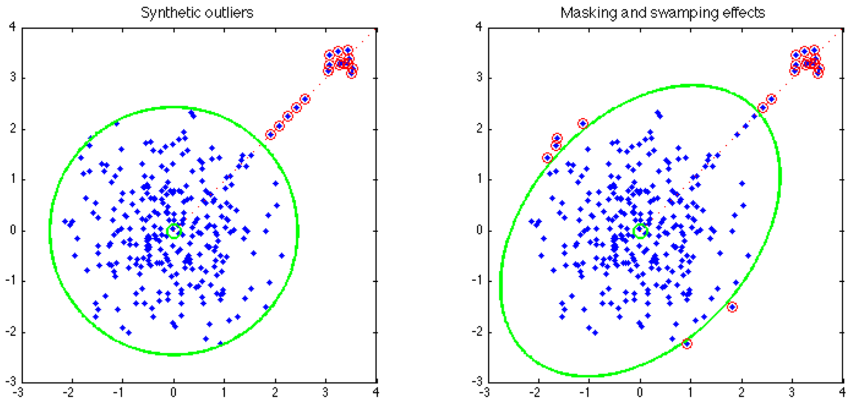
1. Often used are the sample mean $\bar{\mathbf{x}}$ and covariance $\widehat{\boldsymbol{\Sigma}}$
2. $\mathrm{MD} \sim \chi_p^2$ asymptotically, for $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
3. However, problem with masking and swamping effects ...

# Univariate case: Box-plot and QQ-plot



# 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 ...

# Problem Setup

**Sphering:** For $\{\mathbf{x}_i\}_{i=1}^n$ in $\mathbb{R}^p$, consider the "sphered" data,

$$\mathbf{z}_i = \mathbf{\Sigma}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}), \quad i = 1, \ldots, n$$

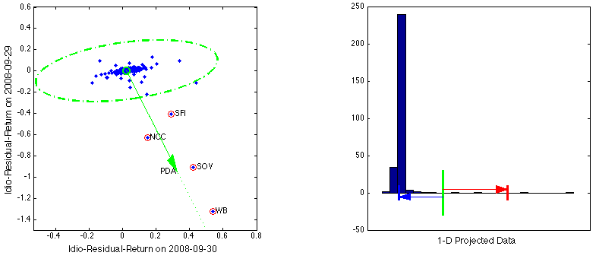
given any $\mathbf{\Sigma} > 0$ (positive definite).

**Projection:** For $\mathbf{w} \in \mathbb{R}^p$ with $||\mathbf{w}|| = 1$, the "projected" data

$$\{\mathbf{w}^T\mathbf{z}_1, \ldots, \mathbf{w}^T\mathbf{z}_n\}, \text{ in 1-D.}$$

**Question:** What is the best direction $\mathbf{w}^*$ that would separate clearly the outlying observations in 1-D space?

# Principal Direction of Anomaly




Suppose $\exists 100\alpha\%$ outliers, define the separation $\mathrm{Score}(\mathbf{w},\alpha)$ by

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

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

$$\textbf{PDA:}\quad \mathbf{w}^*_\alpha = \arg\max_{||\mathbf{w}||=1} \mathrm{Score}(\mathbf{w},\alpha), \quad \alpha \in (0, 0.5]$$

# Theorem

*Given the projected data* $\mathscr{D}_0 = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, *let* $\mathscr{D} \subset \mathscr{D}_0$, *then*

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

*For* $\alpha = j/n$ *with* $j = 1, \ldots, \lfloor n/2 \rfloor$, *the score is bounded from above by* $\mathrm{Score}(\mathbf{w}, j/n) \leq \frac{n}{n-j} \left\| \bar{\mathbf{z}}^*_{(1:j)} \right\|$, *where*

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

*The maximum score is attained by the PDA* $\mathbf{w}^*_{j/n} \propto \bar{\mathbf{z}}^*_{(1:j)}$.

**Corollary 1:** Set $\alpha = 1/n$ and let $\mathbf{z}^* \leftarrow \max_i \|\mathbf{z}_i\|$ with maximal Euclidean distance. Then, the PDA is given by

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

**Corollary 2:** Based on the raw data $\mathscr{D}_0 = \{\mathbf{x}_i\}_{i=1}^n$, the separation score

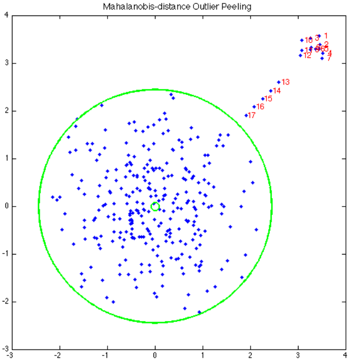
$$\mathrm{Score}(\mathbf{w}, j/n; \mathbf{\Sigma}) = \frac{n}{j(n-j)} \max_{|\mathscr{D}|=j} \sum_{\mathbf{x}\in\mathscr{D}} \mathbf{w}^T \mathbf{\Sigma}^{-1/2}(\mathbf{x} - \bar{\mathbf{x}}), \quad \mathscr{D} \subset \mathscr{D}_0$$

The PDA is given by $\mathbf{w}^*_{j/n} \propto \mathbf{\Sigma}^{-1/2}(\bar{\mathbf{x}}^*_{(1:j)} - \bar{\mathbf{x}})$, where $\bar{\mathbf{x}}^*_{(1:j)} = \frac{1}{j}\sum_{\mathbf{x}\in\mathscr{D}^*} \mathbf{x}$ and $\mathscr{D}^* = \arg\max_{|\mathscr{D}|=j} \left\|\mathbf{\Sigma}^{-1/2}\sum_{\mathbf{x}\in\mathscr{D}}(\mathbf{x} - \bar{\mathbf{x}})\right\|$.
For $\alpha = 1/n$ and $\mathbf{\Sigma} = \widehat{\mathbf{\Sigma}}$, let $\mathbf{x}^*$ attain the maximal Mahalanobis distance. Then

$$\textbf{PDA:} \quad \mathbf{w}^*_{1/n} = \widehat{\mathbf{\Sigma}}^{-1/2}(\mathbf{x}^* - \bar{\mathbf{x}})/\sqrt{\mathrm{MD}(\mathbf{x}^*)}$$

# 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

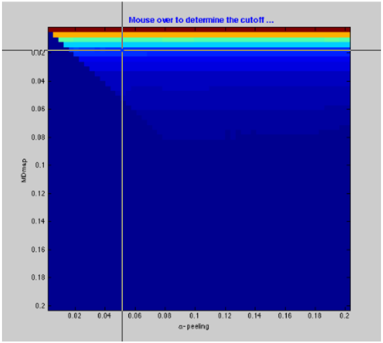
**MD-based Peeling Algorithm:** input $\alpha_0$ (0.5 by default)

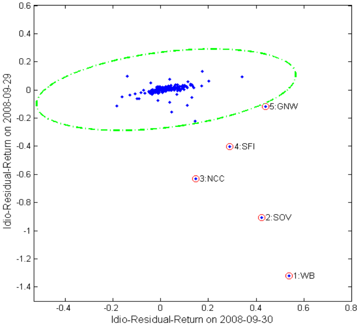
1. Initialize $\mathscr{D}_1 = \mathscr{D}_0 = \{\mathbf{x}_i\}_{i=1}^n$ and $k = 0$
2. Compute Mahalanobis distance $(\mathrm{MD})$ for sample $\mathscr{D}_1$; find one of the elements with $\max \mathrm{MD}$

```
MD = mahal(D1,D1); [maxMD, outId] = max(MD);
```

3. If $k < n\alpha_0$, flag $\mathtt{D1}_{\mathtt{outId}}$ as outlier, update $k+1 \to k$, $\mathscr{D}_1 \backslash \mathtt{outId} \to \mathscr{D}_1$ 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

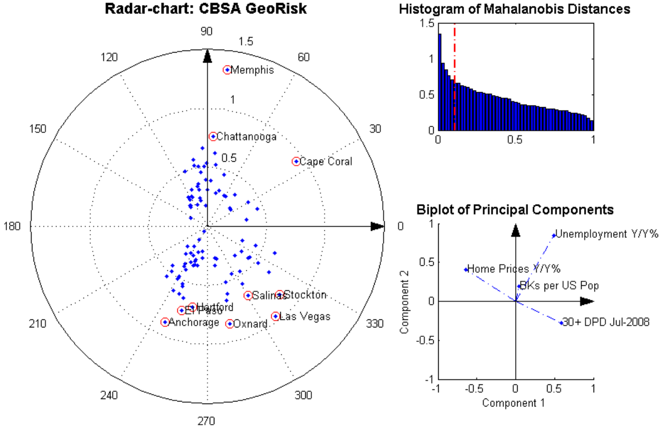
# Radar-chart Visualization

- For each suspicious subject $i$, the peeling algorithm gives us
  (a) Mahalanobis distance: $D_i = \sqrt{\mathrm{MD}_i}$ (scalar)
  (b) Outlying direction: $\mathbf{w}_i$ s.t. $\|\mathbf{w}\| = 1$ (spherical)
- Radar-chart visualization is a natural choice, by converting

$$D_i \to \text{Radius}, \quad \mathbf{w}_i \to \text{Radian (angle)}$$

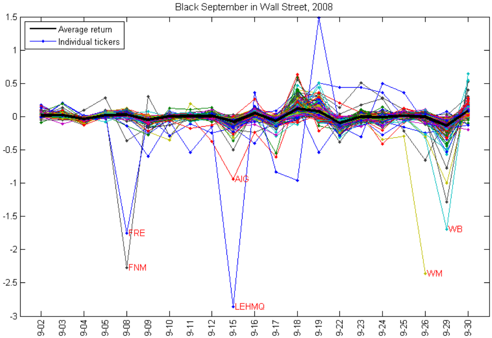
- Trivial case if $\mathbf{w} \in \mathbb{R}^2$:
```
>> theta = acos(W(:,1)).*sign(W(:,2));
```
- Nontrivial if $\mathbf{w}$ is high-dimensional. We need dimension reduction techniques, e.g. MDS (multidimensional scaling)

# CBSA GeoRisk




- Robust PC1 and PC2 are used as reference coordinates
- Better choices are under development ⇒ to report later

# Tracking Financial Storm

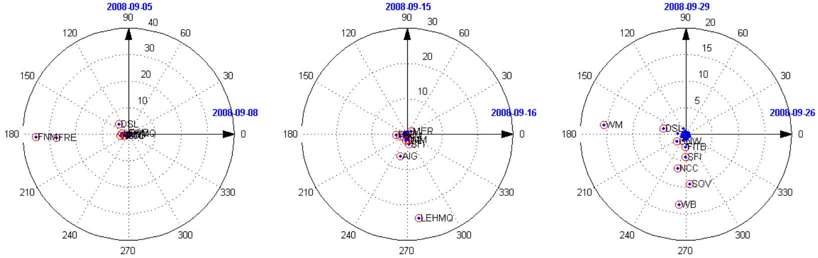



For $i = 1, \ldots, 288$ (Financial firms included in DJUSFN and KBW)

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

Portfolio anomaly detection via **idiosyncratic residuals** $\widehat{\varepsilon}_{ki}$ ...

# Black September - Radar Tracking




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

# 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