On a discrete Laplacian based method for outliers detection in Phase I of profile control charts

F. D. Moura Neto¹, M. S. De Magalhães²

¹ Polytechnic Institute, Rio de Janeiro State University, Nova Friburgo
² National School of Statistical Sciences, Brazilian Institute of Geography and Statistics, Rio de Janeiro
Outline

- Profiles
- Phases I and II for Profile Control Charts
- More on Phase I: Baseline; measures of performance
- Profiles, Projections and Graphs
- Laplacian, Fiedler vector and clustering of a graph
- Profiles, Fiedler vector and outliers
- Results and conclusions
Profiles: quality characteristics of products or production processes

Semiconductors manufacturing: linear profile (Kang and Albin 2000)

Calibration of MFC: measured pressure in the etcher chamber, $y$, is a linear function of the mass flux $x$ allowed by the MFC (ideal gas behavior)
Profiles: quality characteristics of products or production processes

Vertical Density Profile: nonlinear profile

24 profiles consists of the density of a wood board measured at fixed depths across the thickness of the board with 314 measurements taken 0.002 inches apart.

Walker and Wright, J. Quality Technology (2002), 34, 118-129
Williams, Woodall and Birch, (2003)
Zhang and Albin, IIE Transactions (2009), 41, 335-345

**Profiles**: quality characteristics of products or production processes

Potential versus current nonlinear profile in steel corrosion

**304 stainless steel**

**316 stainless steel**

11 profiles consisting of the resulting current due to imposed potential, in process of corrosion of stainless steel (a) 304 and (b) 316. Data from Prof. I. Bastos

The 316 family is a group of stainless steels with superior resistance to corrosion compared to 304 stainless steels, due to the presence of 4% molybdenum. Jumps in current are due to the formation of a pite or crevice.
Phase I x Phase II

• **Phase I**: determination of *baseline profile* (standard)

• **Phase II**: monitoring *current profile* to check if it matches the *baseline* (process is *in-control*) or fails to match (process is *out-of-control*)
Phase I: determination of standard profile

- From a set of initial (historical) profiles distinguish between
  - Outliers profiles (process out-of-control)
  - Non-outliers profiles (process in-control)
- Define baseline profile by averaging profiles classified as non-outliers
  - Misclassification of profiles as non-outliers is troublesome

Source: Zhang and Albin, IIE Transactions (2009), 41, 335-345

Fig. 2. 200 non-outlier profiles in gray and one outlier profile in bold.
Performance measure of a Phase I method

Type I Error = ‘Probability of classifying an non-outlier profile as an outlier profile’;

Type II Error = ‘Probability of classifying an outlier profile as a non-outlier profile’;

Best: in phase I, minimize type II error since one uses the profiles classified as non-outlier to set the baseline non-outlier (in-control) profile

Phase I: want small type II error

Phase II: want small type I error
Phase I for nonlinear profile

• Williams, Woodall and Birch (2003) uses nonlinear regression models and nonparametric regression.
• Zhang and Albin (2009) compares nonlinear regression method versus $\chi^2$ control chart method.

• We present a spectral method based on the Laplacian operator.

• Studies above work directly with a vector representation of a nonlinear profile.
A complex model of a nonlinear profile
(Zhang & Albin, 2009)

\[ y_{ij} = f_a(x) + \varepsilon_{ij}, \]

\[ f_a(x) = 10 - 20ae^{-ax_j} \sin(\sqrt{4 - a^2x_j})/\sqrt{4 - a^2} + 10e^{-ax_j} \cos(\sqrt{4 - a^2x_j}) \]

\[ \varepsilon_{ij} \sim N(0, \sigma^2) \]

Out-of-control = outlier
In-control = Non-outlier

Fig. 1. Plots of \( f_{0.5}(x) \) and \( f_{1.1}(x) \).

Source: Zhang and Albin, IIE Transactions (2009), 41, 335-345
Representation of a profile in finite dimensional vector space

\[ \mathbf{y} = P(f_\alpha) = (18.7, 6.6, 5.4, 11.3, 12.1, 9.4, 9.0, 10.2, 10.4, 9.9, 9.8) \in \mathbb{R}^{11} \]
Projection of profile in finite dimensional space

Baseline

non-outlier

outlier

+ BLACK: baseline  * BLUE: non-outliers  o RED: outliers
Associating a graph to a set of profiles

- Each node of the graph represents a profile.
- Each arc of the graph, connecting two profiles, has associated a weight, representing ‘closeness’ of profiles.
- Wants to split the graph in two groups, in such a way that:
  - one group has almost all non-outliers and very few outliers (type II error), and
  - the other has almost all outliers and few non-outliers (type I error).
- Wants closeness between non-outliers and between outliers to be big while closeness between a non-outlier and an outlier be small.
- The splitting in two groups is accomplished by Fiedler’s vector.
Laplacian of a graph and Fiedler vector

- Graph with 6 nodes and 8 edges
- Diagonal edge weight matrix $C$. Each edge $\{\nu_i, \nu_j\}$ has a weight $c_{ij}$ representing how much $\nu_i$ and $\nu_j$ are ‘close’ (the bigger $c_{ij}$ the closer they are);

\[
C = \text{diag}(c_{12}, c_{13}, c_{14}, c_{23}, c_{24}, c_{34}, c_{45})
\]

- Incidence matrix, 8x6

\[
A = \begin{pmatrix}
-1 & 1 & 0 & 0 & 0 & 0 \\
0 & -1 & 1 & 0 & 0 & 0 \\
0 & -1 & 0 & 1 & 0 & 0 \\
0 & -1 & 0 & 0 & 1 & 0 \\
0 & 0 & -1 & 0 & 1 & 0 \\
0 & 0 & 0 & -1 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & 1
\end{pmatrix}
\]

- Weight matrix, 8x8

\[
W = \begin{pmatrix}
c_{01} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
c_{12} & c_{12} & 0 & 0 & 0 & 0 & 0 & 0 \\
c_{13} & c_{13} & c_{13} & 0 & 0 & 0 & 0 & 0 \\
c_{14} & c_{14} & c_{14} & c_{14} & 0 & 0 & 0 & 0 \\
c_{23} & c_{23} & c_{23} & c_{23} & 0 & 0 & 0 & 0 \\
c_{24} & c_{24} & c_{24} & c_{24} & c_{24} & 0 & 0 & 0 \\
c_{34} & c_{34} & c_{34} & c_{34} & c_{34} & 0 & 0 & 0 \\
c_{45} & c_{45} & c_{45} & c_{45} & c_{45} & c_{45} & 0 & 0 \\
c_{45} & c_{45} & c_{45} & c_{45} & c_{45} & c_{45} & 0 & 0
\end{pmatrix}
\]

- Fiedler: eigenvector to first positive eigenvalue. Separates nodes in two groups, in such a way as to maximize ‘closeness’ in each group.
A graph and its Fiedler vector components

20 nodes and a bunch of edges with weights equal to 1. There are two subgroups with 10 nodes each.
Fiedler vector and partition of nodes in two disjoint sets

- Let us partition the nodes into two disjoint subsets $A$ and $B$. Let $z_i = \pm 1$ whether node $i \in A$ or $i \in B$.

- If nodes $i$ and $j$ are in same subset, their contribution to $\sum_{i,j=1}^{N} w_{ij}(z_i - z_j)^2$ is null. On the other hand, if they are allocated to different subsets, then their contribution is $4w_{ij}$.

- Let $\mathbb{1}$ denote the vector with $N$ entries equal to one. Then, $|z^T \cdot \mathbb{1}|$ equals the difference between the number of elements in $A$ and $B$. If $N$ is even, then $0 \leq |z^T \cdot \mathbb{1}| \leq N$ and if $N$ is odd, then $1 \leq |z^T \cdot \mathbb{1}| \leq N$.

- The problem of clustering the set of nodes in two subsets, in each set getting the nodes that are ‘closer’ to each other (have higher weights in the edges connecting them) can be written as:

$$\min_{\{z_i \in \{-1, 1\}, |z^T \cdot \mathbb{1}| \leq \alpha\}} \sum_{i,j=1}^{N} w_{ij}(z_i - z_j)^2$$

The parameter $\alpha$ in the constraint can be chosen from 1 to $N$. The smaller the parameter is, the closer will be the number of elements in each subset.
Fiedler vector and partition of nodes in two disjoint sets

• Instead of considering the discrete optimization problem, \( z \in \{-1, 1\}^N \), relax to a continuum problem by allowing \( z \in \mathbb{R}^N \), and since the objective function is quadratic, normalize the solution to have squared norm \( N \), (as it would be the case in the discrete problem)

\[
\min \{ z \in \mathbb{R}^N, |z^T \cdot \mathbf{1}| \leq \alpha, z^T \cdot z = N \} \sum_{i,j=1}^{N} w_{ij} (z_i - z_j)^2
\]  \hspace{1cm} (1)

• Have

\[
\sum_{i,j=1}^{N} w_{ij} (z_i - z_j)^2 = \sum_{i,j=1}^{N} z_i (L)_{ij} z_j = z^T L z
\]  \hspace{1cm} (2)

Due to the equality expressed by eq.(2), eq.(1) is a variation of Rayleigh-Ritz problem, and its solution can be related to the eigenfunctions of the Laplacian.
Fiedler vector and profiles

- Each profile gives rise to a node of the graph;
- Edge weights are scalar products of vectors associated with profiles.

Given a profiles $i$ and $j$, represented by $y_i = (y_{1i}, y_{2i}, \ldots, y_{Mi})^T$, and $y_j$, the weight associated with the edge connecting their nodes is

$$w_{ij} = y_i \cdot y_j = \sum_{k=1}^{M} y_{ik} y_{jk}$$

In matrix notation, given $N$ profiles, let $Y = (Y)_{ij} = y_{ij}$, be the $M \times N$ matrix whose columns represent the profiles. Also, denote by $W = (W)_{ij} = w_{ij}$ the $N \times N$ matrix of weights between profiles, that is

$$Y = (y_1, y_2, \ldots, y_N)$$

Then

$$W = Y^T Y$$

Let $D = (D)_{ij}$ be the diagonal matrix with entries in the diagonal given by

$$D = (D)_{ii} = \sum_{k=1}^{N} w_{ik}$$

**Sum of weights along a line**

The laplacian is then given by

$$L = D - W$$
**Type I Error:** Average percent (and standard deviation) of non-outlier profiles incorrectly identified as outliers by chi-square method and by laplacian method

<table>
<thead>
<tr>
<th>Number of outliers (P)</th>
<th>Parameter a for outliers (a = 0.5 for non-outliers)</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
<th>1.1</th>
<th>1.3</th>
<th>1.5</th>
<th>1.7</th>
<th>1.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td></td>
<td>7(1)</td>
<td>9(1)</td>
<td>7(1)</td>
<td>7(1)</td>
<td>8(1)</td>
<td>8(1)</td>
<td>8(1)</td>
<td>8(1)</td>
</tr>
<tr>
<td>40</td>
<td></td>
<td>7(1)</td>
<td>6(1)</td>
<td>3(1)</td>
<td>4(1)</td>
<td>4(0)</td>
<td>4(0)</td>
<td>4(0)</td>
<td>4(0)</td>
</tr>
<tr>
<td>60</td>
<td></td>
<td>5(1)</td>
<td>3(1)</td>
<td>1(0)</td>
<td>2(1)</td>
<td>2(0)</td>
<td>2(0)</td>
<td>3(2)</td>
<td>3(2)</td>
</tr>
<tr>
<td>80</td>
<td></td>
<td>7(1)</td>
<td>8(2)</td>
<td>3(1)</td>
<td>2(1)</td>
<td>3(1)</td>
<td>5(1)</td>
<td>5(1)</td>
<td>6(1)</td>
</tr>
</tbody>
</table>

Chi-square method  (Zhang & Albin (2009))

Laplacian method

200 profiles = (200-P) non-outliers + P outliers
300 simulations
8(3) = average Type I error 8% , standard deviation 3
**Type II Error:** Average percent (and standard deviation) of outlier profiles incorrectly identified as non-outliers by chi-square method and by laplacian method as parameter a shifts

<table>
<thead>
<tr>
<th>Number of outliers (P)</th>
<th>Parameter ( a ) for outliers ((a = 0.5 \text{ for non-outliers}))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td>20</td>
<td>N/A</td>
</tr>
<tr>
<td>40</td>
<td>N/A</td>
</tr>
<tr>
<td>60</td>
<td>N/A</td>
</tr>
<tr>
<td>80</td>
<td>N/A</td>
</tr>
</tbody>
</table>

\( \Rightarrow \) chi-square method (Zhang & Albin (2009))

Laplacian method

<table>
<thead>
<tr>
<th>( a \rightarrow )</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
<th>1.1</th>
<th>1.3</th>
<th>1.5</th>
<th>1.7</th>
<th>1.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-outliers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>N/A</td>
<td>39 (25)</td>
<td>2 (5)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>40</td>
<td>N/A</td>
<td>40 (24)</td>
<td>1 (3)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>60</td>
<td>N/A</td>
<td>48 (25)</td>
<td>1 (2)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>80</td>
<td>N/A</td>
<td>49 (31)</td>
<td>1 (3)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

200 profiles = (200-P) non-outliers + P outliers

300 simulations

8(3) = average Type II error 8%, standard deviation 3
Outlier Profiles
Williams et al. (2003): 4, 9, 15, 18, 24
Zhang & Albin (2009): 3, 6, 9, 10, 14
This talk: 2, 6, 9, 14, 19
Outlier Profiles
This talk: 2, 6, 9, 14, 19
6-blue
14-green
9-black
19- magenta
2-red

Outlier Profiles
Williams et al. (2003): 4, 9, 15, 18, 24

Outlier Profiles
Zhang & Albin (2009): 3, 6, 9, 10, 14

VDP data
Conclusions

• The method based on discrete Laplacian is able to better separate, in most artificial data cases studied, between outlier and non-outlier profiles for shifts in the parameters of nonlinear profile model presented; in fact it presents, in general, smaller type I errors as well as type II errors;

• Having smaller Type II errors leads to a better estimation of the nonlinear baseline profile;

• We present preliminary results and guess that the method discussed may be competitive for profile Phase I investigations;

• Further analysis is needed to fully access the capabilities of the discrete Laplacian-Fiedler vector method. In particular, appropriate notions of closeness between profiles must be addressed.