A Least Angle Regression Control Chart for Multidimensional Data

GIOVANNA CAPIZZI and GUIDO MASAROTTO
Department of Statistical Sciences
University of Padua
Italy

2ND INTERNATIONAL SYMPOSIUM ON STATISTICAL PROCESS CONTROL
Rio de Janeiro, Brazil
July 13-14, 2011
Outline

1. Multivariate Statistical Monitoring
2. Variable-selection Methods in SPC
3. Proposed Procedure
   - Reference Model
   - Least Angle Regression Algorithm
   - LAR-EWMA control chart
4. Simulation Results
5. Concluding Remarks and Future Research
Synopsis

Framework: Phase II monitoring of a normal multivariate vector of product variables.

Fault type: Persistent change in the process mean and/or increase in the total dispersion.

Problem: Simultaneous monitoring of several variables: how many and which variables are really changed?

Proposal: Combination of a variable-selection method (Least Angle Regression) with a multivariate control chart (multivariate EWMA).

Results: The LAR-based EWMA is a very competitive statistical tool for handling several change-point scenarios in the high-dimensional framework.
### Multivariate Statistical Monitoring

#### Possible Frameworks

**Unstructured:**
- multivariate vector of quality characteristics

**Structured:**
- analytical models describing process quality
  - linear and non-linear profiles
  - multistage processes
Multivariate Statistical Monitoring
Possible Approaches

- Monitoring the stability of the **whole** set of product variables (low sensitivity in a high-dimensional context)
- Monitoring a **reduced** set of out-of-control product variables.

But the shifted components are obviously unknown!
Promising approach: combine a suitable variable selection method with the multivariate statistical monitoring

- Forward search algorithm with a Shewhart-type control chart (Wang and Jiang, 2009) for handling mean changes in the unstructured scenario.
- LASSO algorithm with a multivariate EWMA, for detecting
  1. mean changes in the unstructured case (Zou and Qiu, 2009)
  2. changes in multivariate linear profiles (Zou et al. 2010)
LEAST ANGLE REGRESSION WITH A MULTIVARIATE EWMA

1. general model formulation unifying the unstructured and structured framework
2. use of a variable selection method yet unexplored in the SPC framework
3. competitive procedure for detecting changes in process mean and total dispersion for a wide variety of change point-scenarios
4. relatively simple tool for fault identification
**CHANGE-POINT MODEL**

\[ y_t \sim \begin{cases} 
N_n(\mu, \Sigma) & \text{if } t < \tau \\
N_n(\mu + \delta, \Omega) & \text{if } t \geq \tau 
\end{cases} \]

\((\text{in-control})\)

\((\text{out-of-control})\)

**MEAN SHIFT**

\[ \delta = F\beta \]

\(F_{n \times p} \): matrix of known constants; \(\beta_{p \times 1} \): vector of unknown parameters.

**DISPERSION INCREASE**

\[ E_{OC}[\xi_t^2] > n \iff \Omega - \Sigma \text{ positive definite matrix} \]

with \( \xi_t^2 = (y_t - \mu)\Sigma^{-1}(y_t - \mu)' \).
Reference Model

Frameworks and characterization of $\delta$

**Unstructured**

$$[\delta_i] = [\beta_i] \iff F = I_n$$

**Profile**

$$[\delta_i] = [g(x_i)] = \left[ \sum_{j=1}^{p} \beta_j f_j(x_i) \right]$$

e.g. $\delta_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2$ for $i = 1, \ldots, n$. 

**Multistage Process**

Standard **state space** representations of a process with $n$ stages can be written in the desired form.

$$\begin{cases} y_{t,i} = \mu_i + c_i x_{t,i} + v_{t,i} \\ x_{t,i} = d_i x_{t,i-1} + \beta_i l_{\{t \geq \tau\}} + w_{t,i} \end{cases} \quad (i = 1, \ldots, n)$$
Multivariate EWMA control statistic

\[ z_t = (1 - \lambda)z_{t-1} + \lambda(y_t - \mu) \]

with \( z_0 = 0_n, \ 0 < \lambda \leq 1 \).

Linear model

\[ z_t = F\beta + a_t, \]

with \( a_t \sim N_n(0_n, \lambda/(1 - \lambda)\Sigma) \)

**MONITORING STABILITY OF PROCESS MEAN**

\[ H_0 = \{ \text{the process is in control} \} \iff \{ \beta = 0 \} \]

\[ H_1 = \{ \text{the process is out of control} \} \iff \{ \text{How many } \beta_j \text{ are non zero? One, two,...all?} \} \]
Least Angle Regression algorithm

Main steps

1. Start with all the $p$ coefficients equal to zero.
2. Build up estimates of the unknown mean in successive steps: each step adding one variable.
3. Reach the full least square solution in $p$ steps (using all the variables).

**Step $k$**

\[
\{j_1, \ldots, j_k\} \quad \iff \quad \{\beta_{j_1}, \ldots, \beta_{j_k}\}
\]

variables selected by LAR \iff just $k$ parameters are assumed $\neq 0$
### Mean Change

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Control Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_{j_1} \neq 0, \ldots, \beta_{j_k} \neq 0, \beta_{j_{k+1}} = \cdots = \beta_{j_p} = 0 )</td>
<td>Likelihood ratio test: ( S_{t,k} \ (k = 1, \ldots, p) )</td>
</tr>
</tbody>
</table>

### Variation Increase

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Control Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta = 0_p ) and ( E[\xi_t^2] &gt; n )</td>
<td>( S_{t,p+1} = \max \left( 1, (1 - \lambda)S_{t-1,p+1} + \lambda \frac{\xi_t^2}{n} \right) ) with ( S_{0,p+1} = 1 )</td>
</tr>
</tbody>
</table>
LAR-EWMA control chart
Control statistic, stopping rule, fault identification

Control statistic

\[ W_t = \max_{k=1,\ldots,p+1} \frac{S_{t,k} - a_k}{b_k} \]

Alarm time

\[ t^* = \min \{ t : W_t > h \} \]

Fault Identification

\[ k^* = \min \left\{ k : \frac{S_{t^*,k} - a_k}{b_k} > h \right\} \]

\[ k^* \leq p : \{ \text{plausible mean shift} \} \]

\[ k^* = p + 1 : \{ \text{plausible dispersion increase} \} \]
**LAR-EWMA control chart**

**Control chart design**

**Smoothing Constant $\lambda$**
- $(0.1, 0.3)$ normal distribution
- $(0.03, 0.05)$ skewed and heavy-tailed distributions

**Control Limit $h$**
- $h$ is determined to obtain a desired value of the in-control ARL
- via simulation, using the Polyak-Ruppert stochastic approximation algorithm.
### Mean Change and No Variance Change

<table>
<thead>
<tr>
<th>Unstructured Scenario</th>
<th>Multistage Processes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEWMA (Lowry et al., 1992)</td>
<td>DEWMA (Zou and Tsung, 2008)</td>
</tr>
<tr>
<td>LEWMA (Zou and Qiu, 2009)</td>
<td></td>
</tr>
</tbody>
</table>

### Mean Change and Variance Increase

<table>
<thead>
<tr>
<th>Parametric Profile</th>
<th>Nonparametric Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>KMW (Kim, et al. 2003))</td>
<td>NEWMA (Zou et al., 2008)</td>
</tr>
<tr>
<td>PEWMA (Zou et al., 2007)</td>
<td></td>
</tr>
</tbody>
</table>
Simulation Results

Investigation of a wide variety of out-of-control scenarios (unstructured and structured)

1. large range of size shifts
2. different combinations of shifted components or locations: one, more than one.
3. mean shifts and/or variance increases

performance measure: Average Run Length

summary performance measure: the Relative Mean Index (Han and Tsung, 2006).
Very small RMI values → best or close to the best chart.
Simulation Results

The Relative Mean Index (Han and Tsung, 2006)

The RMI is a summary performance measure defined as

\[
RMI = \frac{1}{N} \sum_{l=1}^{N} RMI_l = \frac{1}{N} \sum_{l=1}^{N} \frac{ARL_{\delta_l} - MARL_{\delta_l}}{MARL_{\delta_l}}
\]

1. \(N\) total number of shifts
2. \(ARL_{\delta_l}\) out-of-control ARL for detecting \(\delta_l\)
3. \(MARL_{\delta_l}\) smallest out-of-control ARL, among the compared charts, for detecting \(\delta_l\)
4. \(RMI_l\) relative efficiency of a chart in detecting \(\delta_l\) compared to the best chart.
IC: $y_{t,i} \sim N(0, 1)$. OC: $y_{t,i} \sim N(\beta_1 + \beta_2 x_i, \omega^2)$, $i = 1, \ldots, 4$, $x_i = -1, -1/3, 1/3, 1$

<table>
<thead>
<tr>
<th>Shifts</th>
<th>PEWMA</th>
<th>KMW</th>
<th>LAR-EWMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1 = 0.1$</td>
<td>293.08</td>
<td>296.40</td>
<td>272.39</td>
</tr>
<tr>
<td>$\beta_1 = 0.3$</td>
<td>46.16</td>
<td>43.04</td>
<td>39.55</td>
</tr>
<tr>
<td>$\beta_1 = 1$</td>
<td>4.65</td>
<td>4.29</td>
<td>4.22</td>
</tr>
<tr>
<td>$\beta_2 = 0.2$</td>
<td>180.73</td>
<td>182.79</td>
<td>163.19</td>
</tr>
<tr>
<td>$\beta_2 = 0.4$</td>
<td>46.84</td>
<td>43.66</td>
<td>40.14</td>
</tr>
<tr>
<td>$\beta_2 = 1.2$</td>
<td>5.41</td>
<td>4.99</td>
<td>4.90</td>
</tr>
<tr>
<td>$\beta_1 = 0.1, \beta_2 = 0.1$</td>
<td>230.79</td>
<td>243.60</td>
<td>215.11</td>
</tr>
<tr>
<td>$\beta_1 = 0.4, \beta_2 = 0.2$</td>
<td>20.69</td>
<td>21.14</td>
<td>18.80</td>
</tr>
<tr>
<td>$\beta_1 = 0.4, \beta_2 = 0.6$</td>
<td>10.39</td>
<td>12.34</td>
<td>10.05</td>
</tr>
<tr>
<td>$\beta_1 = 0.6, \beta_2 = 0.6$</td>
<td>7.11</td>
<td>8.16</td>
<td>6.88</td>
</tr>
<tr>
<td>$\beta_1 = 0.1, \omega = 1.2$</td>
<td>45.88</td>
<td>50.95</td>
<td>38.78</td>
</tr>
<tr>
<td>$\beta_1 = 0.3, \omega = 1.2$</td>
<td>21.55</td>
<td>23.77</td>
<td>20.14</td>
</tr>
<tr>
<td>$\beta_1 = 1, \omega = 1.2$</td>
<td>4.44</td>
<td>4.29</td>
<td>4.08</td>
</tr>
<tr>
<td>$\beta_2 = 0.2, \omega = 1.2$</td>
<td>38.68</td>
<td>43.46</td>
<td>33.71</td>
</tr>
<tr>
<td>$\beta_2 = 0.4, \omega = 1.2$</td>
<td>21.80</td>
<td>23.93</td>
<td>20.13</td>
</tr>
<tr>
<td>$\beta_2 = 1.2, \omega = 1.2$</td>
<td>5.12</td>
<td>4.97</td>
<td>4.72</td>
</tr>
<tr>
<td>$\beta_1 = 0.1, \beta_2 = 0.1, \omega = 1.2$</td>
<td>42.47</td>
<td>47.36</td>
<td>36.26</td>
</tr>
<tr>
<td>$\beta_1 = 0.4, \beta_2 = 0.2, \omega = 1.2$</td>
<td>13.74</td>
<td>15.14</td>
<td>13.02</td>
</tr>
<tr>
<td>$\beta_1 = 0.4, \beta_2 = 0.6, \omega = 1.2$</td>
<td>8.60</td>
<td>9.96</td>
<td>8.36</td>
</tr>
<tr>
<td>$\beta_1 = 0.6, \beta_2 = 0.6, \omega = 1.2$</td>
<td>6.41</td>
<td>7.29</td>
<td>6.19</td>
</tr>
<tr>
<td>$\omega = 1.2$</td>
<td>53.39</td>
<td>58.17</td>
<td>43.24</td>
</tr>
<tr>
<td>$\omega = 1.5$</td>
<td>11.06</td>
<td>12.35</td>
<td>7.90</td>
</tr>
<tr>
<td>$\omega = 2$</td>
<td>4.37</td>
<td>4.71</td>
<td>3.03</td>
</tr>
<tr>
<td>RMI</td>
<td>0.13</td>
<td>0.19</td>
<td>0.00</td>
</tr>
</tbody>
</table>
### Simulation Results: RMI

\[ ARL_0 = 500, \lambda = 0.2, \tau = 1 \]

<table>
<thead>
<tr>
<th>“Unstructured”</th>
<th>LAR-EWMA</th>
<th>MEWMA</th>
<th>REWMA</th>
<th>LEWMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n = 15 )</td>
<td>0.02</td>
<td>0.24</td>
<td>0.32</td>
<td>0.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Linear Profile</th>
<th>LAR-EWMA</th>
<th>PEWMA</th>
<th>KMW</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n = 4 )</td>
<td>0.00</td>
<td>0.13</td>
<td>0.19</td>
</tr>
<tr>
<td>( n = 10 )</td>
<td>0.01</td>
<td>0.07</td>
<td>0.11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cubic Profile</th>
<th>LAR-EWMA</th>
<th>PEWMA</th>
<th>KMW</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n = 8 )</td>
<td>0.00</td>
<td>0.24</td>
<td>0.36</td>
</tr>
<tr>
<td>( n = 15 )</td>
<td>0.00</td>
<td>0.23</td>
<td>0.31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Non-parametric profile</th>
<th>LAR-EWMA</th>
<th>NEWMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n = 20 )</td>
<td>0.05</td>
<td>0.28</td>
</tr>
<tr>
<td>( n = 40 )</td>
<td>0.06</td>
<td>0.30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Multistage process</th>
<th>LAR-EWMA</th>
<th>MEWMA</th>
<th>DEWMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n = 20; c_i = d_i = 1 )</td>
<td>0.02</td>
<td>0.24</td>
<td>0.10</td>
</tr>
<tr>
<td>( n = 20; c_i = 1.2; d_i = 0.8)</td>
<td>0.06</td>
<td>0.22</td>
<td>0.18</td>
</tr>
</tbody>
</table>
The LAR-based multivariate EWMA

- offers an unifying approach for handling the multivariate statistical monitoring in both the unstructured and structured framework
  - Flexible choices of $F$ in a quite general model formulation
- makes use of a relatively easier and faster variable selection method
- shows some appealing enhancements:
  - a control statistic for detecting increases in total dispersion
  - a simple tool for fault detection.
Concluding remarks and future research

Future Research

- A more general formulation for dispersion changes.
- Generalization of the LAR algorithm for non-normal data (but how is possible to handle dispersion changes in a distribution-free way?)
- Design of variable sampling interval version of the LAR-EWMA.
Since 1222

“Universa Universis Patavina Libertas”
(Paduan Freedom is Complete and for Everyone)

Anatomy Theatre (1594)

Galileo Galilei’s desk (≈ 1605)