Robust Multivariate Outlier Diagnostics in Chemometrics with Application to Spectrally Overlapping Drugs

Aylin ALIN ¹  Claudio AGOSTINELLI ²  Georgi GERGOV ³  Plamen KATSAROV ⁴  Yahya AL-DEGS ⁵

¹Department of Statistics, Dokuz Eylul University, Izmir, Turkey

²Dipartimento di Matematica, Università di Trento, Trento, Italy

³Faculty of Pharmacy, Medical University, Sofia, Bulgaria

⁴Department of Pharmaceutical Sciences, Faculty of Pharmacy, Medical University-Plovdiv, Bulgaria

⁵Department of Chemistry, Hashemite University, Jordan
<table>
<thead>
<tr>
<th>Outline</th>
<th>Introduction</th>
<th>PLSR</th>
<th>RWSIMPLS</th>
<th>Outlier Diagnostics</th>
<th>Numerical Studies</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Introduction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Robust SIMPLS Algorithm-RWSIMPLS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Alin and Agostinelli, 2017]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Robust Diagnostics for Detecting Outliers and Influential Observations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Numerical Example</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
\[ Y = E(Y|x_1, \ldots, x_k) + \epsilon = \beta X + \epsilon \] (1)

For drug concentration problem:

- \( Y \Rightarrow nxm \) matrix of \( m \) concentration values.
- \( X \Rightarrow nx(k + 1) \) matrix of \( k \) absorbance values calculated as 
  \[ A = -\log_{10} R \] where \( R \) represents the reflectance.
- \( \beta = (\beta_1, \beta_2, \ldots, \beta_m) \Rightarrow (k + 1)xm \) matrix of unknown parameters with \( \beta_s = (\beta_{j0}, \ldots, \beta_{sk}) \) for \( s = 1, \ldots, m \).
- \( \epsilon = (\epsilon_1, \ldots, \epsilon_m) \) is a \( nxm \) matrix of errors.
SIMPLS is one of the mostly used PLSR algorithm.

It is extremely sensitive to outliers.

For multivariate response variable ⇒ RSIMPLS by [Hubert and Branden, 2003].

For univariate response variable ⇒ Partial version of M-regression estimator (PRM) by [Serrneels et al., 2005]

For multivariate response variable ⇒ Robustly Weighted SIMPLS algorithm (RWSIMPLS) by [Alin and Agostinelli, 2017]
We propose new robust diagnostic tools based on the estimates from the RWSIMPLS to identify leverage points, regression outliers, influential observations and points causing non-normality in both concentration and absorbance values of overlapping binary drug systems.

We compare the performance of the proposed diagnostics with the ones obtained from the classical SIMPLS on two data sets:

- Biscuit dough data
- Paracetamol (PAR) and Caffeine (CAF) data set
\[ Y = E(Y|x_1, \ldots, x_k) + \epsilon = \beta X + \epsilon \]  

\[ X = TP^T + E \quad Y = UQ^T + F \]  

- For convenience, we assume \( X \) and \( Y \) are mean-centered.  
- \( T = XW = (t_1, \ldots, t_A) \) \( \Rightarrow n \times A \) component score matrix for \( X \)  
- \( U = YC = (u_1, \ldots, u_A) \) \( \Rightarrow n \times A \) component score matrix for \( Y \)  
- \( A \) \( \Rightarrow \) the number of components \( \leq \min(k, m, n) \).  
- \( P = (p_1, \ldots, p_A) \) \( \Rightarrow k \times A \) loading matrix for \( X \)  
- \( Q = (q_1, \ldots, q_A) \) \( \Rightarrow m \times A \) loading matrix for \( Y \)  
- \( W \) \( \Rightarrow k \times A \) weight matrix for \( X \)  
- \( C \) \( \Rightarrow m \times A \) matrix are the weight matrix for \( Y \).
The first score vectors $t_1$ and $u_1$ are the solutions to the following maximization problem.

$$\text{cov}(t, u) = \text{cov}(Xw, Yc) = t^T u$$

$$= (Xw)^T Yc = w^T X^T Yc \Rightarrow \max$$

under the constraints of $\|t\| = \|u\| = 1$.

The maximum of Equation (4) is attained for the vectors $w_1$ and $c_1$ corresponding to the largest singular values of $X^T Y$.

Orthogonality of the previous score vectors; i.e. $t_at_j = 0$ and $u_au_j = 0$ for $1 \leq a < j \leq A$.

In SIMPLS orthogonality is provided by finding the subsequent score vectors on deflated $X^T Y$. 
Below are the steps of the SIMPLS algorithm.

**Step 1:** Centralize $X$ and $Y$ as in (5)

$$x_{ij} = x_{ij} - \text{mean}(x_j) \quad j = 1, \ldots, k$$

$$y_{is} = y_{is} - \text{mean}(y_s) \quad s = 1, \ldots, m$$

and calculate $S_1 = X^T Y$

**Step 2:** $w_1$ is obtained as the left singular vector of $S_1$

**Step 3:** $w_1 = \frac{w_1}{\|w_1\|}$

**Step 4:** $t_1 = Xw_1$

**Step 5:** $t_1 = \frac{t_1}{\|t_1\|}$

**Step 6:** $p_1 = X^T t_1$

**Step 7:** $q_1 = Y^T t_1$
Step 8: \( S_{a+1} = S_a - v_a(v_a^T S_a) \) where \( v_1 = p_1 \) and \( v_a = p_a - V_{a-1}(V_{a-1}^T p_a) \) for \( a > 1 \) with \( V_{a-1} = (v_1, v_2, \ldots, v_{a-1}) \).

Repeat the steps 2-8 until all \( A \) components are extracted.

Calculate the parameter estimates \( \hat{\beta} \) for the model in (2) as in Equation (6).

\[
\hat{\beta} = W(P^T W)^{-1} Q^T
\]
The weight takes a value in $[0, 1]$.

The function $A(.)$ ([Lindsay, 1994]) is a residual adjustment function.

$A(\delta) = 2(\delta + 1)^{1/2} - 1$ is for the Hellinger distance residual function.
The RWSIMPLS algorithm is based on deflating the weighted covariance matrix $S^\omega = X^\omega \top Y^\omega$. The weighted matrices $X^\omega$ and $Y^\omega$ are obtained by multiplying each row by the squared root of

$$
\omega_{ri} = \text{median}_s \{ \omega_{ris} \} \quad \text{for} \quad s = 1, \ldots, m; \quad i = 1, \ldots, n. \quad (8)
$$

The quantity $\omega_{ris}$ is the weight of the $i$th residual corresponding to the $s$th response.
Outliers:

(1) Regression (Vertical) outliers

- The Classical Mahalonobis Distance for Residuals (9)

\[ MD(r_i) = \sqrt{\frac{1}{\hat{\Sigma}_\epsilon} r_i} \]  

- The cut-off value for \( RMD(r_i^\omega) \Rightarrow \sqrt{\chi^2_{m,(1-\alpha)}} \).

(2) Leverage points

- Diagonal elements of the hat matrix \( H = X(X^\top X)^{-1} X^\top \)
- Mahalonobis Distance for Explanatory Variables

\[ MD_i = \sqrt{(x_i - \mu_X)^\top \hat{\Sigma}_X^{-1} (x_i - \mu_X)} \]  

- \( h_{ii} = \frac{MD_i^2}{n-1} + \frac{1}{n} \).
- The cut-off value for \( RMD(r_i^\omega) \Rightarrow \sqrt{\chi^2_{k,(1-\alpha)}} \).
- Good Leverage and Bad Leverage

Influential Points

- Leave-one-out diagnostics (such as Cook’s Distance [Cook, 1977]) are frequently used to flag the influential points.
Robust Mahalonobis Distance for Scores

\[ RMD(t_i^w) = \sqrt{(t_i^w - M(T^w))\top (\hat{\Sigma}_{T^w})^{-1}(t_i^w - M(T^w))} \]  

(11)

\[ t_i^w = (t_{i1}^w, t_{i2}^w, \ldots, t_{iA}^w)\top \Rightarrow \text{The } i\text{th score from score matrix } T^w \text{ calculated as} \]

\[ T^w = X W^\omega \]  

(12)

\[ W^\omega \Rightarrow \text{The robust weight matrix of } X \text{ from the RWSIMPLS method} \]

\[ X \Rightarrow \text{The explanatory variable matrix unweighted but robustly centered as defined in [Alin and Agostinelli, 2017]} \]

\[ M(T^w) \Rightarrow \text{The } A \times 1 \text{ } L_1\text{-median vector} \]

\[ \hat{\Sigma}_{T^w} = \frac{\sum_{i=1}^{n} \omega_r t_i^w t_i^w\top}{\sum_{i=1}^{n} \omega_r - 1}. \]  

(13)

The cut-off value for \( RMD(t_i^w) \) is \( \sqrt{\chi^2_{A, (1-\alpha)}} \).
Robust Mahalonobis Distance for Residuals (14)

\[ RMD(r_i^\omega) = \sqrt{r_i^\omega^\top (\hat{\Sigma}_\epsilon)^{-1} r_i^\omega} \]  

- \( r_i^\omega = (r_{i1}^\omega, r_{i2}^\omega, \ldots, r_{im}^\omega)^\top = r_i(\hat{\beta}^\omega) - M(r(\hat{\beta}^\omega)). \)
- \( r_i(\hat{\beta}^\omega) = y_i - x_i^\top \hat{\beta}^\omega. \)
- \( M(r(\hat{\beta}^\omega)) \Rightarrow \) the \( m \times 1 \) \( L_1 \)-median vector computed from the collection of \( (r_1(\hat{\beta}^\omega), r_2(\hat{\beta}^\omega), \ldots, r_m(\hat{\beta}^\omega)). \)
- \( x_i^\top = (x_{i1}, x_{i2}, \ldots, x_{ik})^\top \Rightarrow i\text{'th row of unweighted but robustly centered } X \text{ matrix.} \)
- \( \hat{\beta} \Rightarrow m \times k \) RWSIMPLS coefficient estimate matrix.
- The \( m \times m \) robust covariance matrix for residuals is

\[ \hat{\Sigma}_\epsilon^\omega = \frac{\sum_{i=1}^{n} \omega_i r_i^\omega r_i^\omega^\top}{\sum_{i=1}^{n} \omega_i - A}. \]  

- The cut-off value for \( RMD(r_i^\omega) \Rightarrow \sqrt{\chi^2_{m,(1-\alpha)}}. \)
Robust Cook’s Distance

- In classical ordinary least squares based regression, the Cook’s Distance gives the squared distance from $\hat{\beta}$ to $\hat{\beta}(i)$ relative to the fixed geometry of $X^\top X$ [Cook and Weisburg, 1982].

- For SIMPLS, the Cook’s Distance can be calculated as to measure the squared distance from $\hat{\beta}$ to $\hat{\beta}(i)$ relative to the fixed geometry of $T^\top T$.

- The Robust Cook’s Distance (16)

$$RCD^w_{is} = \left( \frac{\tilde{r}_{is}(T^w)}{\hat{\sigma}(T^w)s\sqrt{1 - h^w_{ii}}} \right)^2 \frac{h^w_{ii}}{A(1 - h^w_{ii})}$$ (16)

$$\hat{\sigma}(T^w)s = \sqrt{\frac{\sum_{i=1}^{n} \omega_r \tilde{r}(t)_{is}^2}{\sum_{i=1}^{n} \omega_r - A}}$$ (17)

$$\tilde{r}_{is}(T^w) = r_{is}(T^w) - M(r(T^w))s$$
- $r_{is}(T^w) = y_{is} - \hat{y}^w_{is}$.
- $M(r(T^w))_s \Rightarrow$ the $s$th element of the $m \times 1$ $L_1$ median vector of $(r_1(T^w), r_2(T^w), \ldots, r_m(T^w))$.
- $\hat{y}^w_{is} \Rightarrow$ the fitted value at the $i$th row and $s$th column of the matrix in (18), and $h^w_{ii}$ is the $i$th diagonal element of $H^w$.

$$\hat{Y}^w = H^w Y = T^w (T^w^\top T^w)^{-1} T^w^\top Y \quad (18)$$

- Cut-off value=1 [Cook and Weisburg, 1982].
Robust diagnostic to identify outliers with respect to

\[ X = TP^\top + E \]  \hspace{1cm} (19)

\[ ROD(t_i^w) = \sqrt{x_i - P^\omega t_i^w} \]  \hspace{1cm} (20)

- \( x_i \Rightarrow k \times 1 \) robustly centered explanatory variable vector for \( i \)th data.
- \( P^\omega \Rightarrow \) Robust loading matrix for \( X \) from RWSIMPLS.
- Bootstrap method is used to obtain the cut-off value for \( ROD(t_i^w) \).
- Plot of Mahalonobis distance for scores \( RMD(t_i^w) \) against the orthogonal distance for scores \( ROD(t_i^w) \).
The aim is to predict the concentrations of flour, sucrose and water of biscuit doughs based on 40 NIR spectra with measurements every 2nm from 1200 up to 2400 nm.

After applying same preprocessing as in [Hubert and Branden, 2003], we end up with 600 measurements for 40 biscuit dough samples.

Observation 23 is known to be an outlier.

3 components were used for diagnostics.
SIMPLS and RSIMPLS Diagnostics for Biscuit Dough Data-[Hubert and Branden, 2003]
RWSIMPLS Diagnostics for Biscuit Dough Data

(a) Residual Distance vs. Score Distance

(b) Orthogonal Distance vs. Score Distance

(c) Cook's Distance vs. Score Distance
Paracetamol which is one of the most popular and most commonly used analgesic and antipyretic drugs around the world, available without a prescription, both in mono and multi-component preparations.

Coffee and caffeine-containing products affect the cardiovascular system, with their positive inotropic and chronotropic effects, and the central nervous system, with their locomotor activity stimulation and anxiogenic-like effects [Cappelletti et al., 2015].

100 mg/L standard solutions of both drugs were prepared from the corresponding pure materials.

Both drugs were kindly donated from local pharmaceutical company (Hikma Pharmaceuticals, Amman, Jordan).
For both drugs, initial pH was adjusted to 7.0 using diluted 0.01 M NaOH solution. Initial screening studies indicated insignificant effect of pH on spectral behavior of both drugs.

For calibration sample, 25 mixtures of PAR and CAF were prepared.

15-mixture validation set with the new concentration levels outside the ranges of the calibration set were prepared in order to test the prediction performance of the models and to decide for the number of components.

Absorbency of drugs were obtained using a double beam spectrophotometer (Thermo Electron Corporation Nicollet evolution 100).

The spectra of drugs for both calibration and validation sets were recorded over the spectral range of 216 – 300 nm at 2 nm intervals.
### Table: The Concentration Levels for Calibration and Validation Sets for PAR/CAF Drug System

<table>
<thead>
<tr>
<th>Number</th>
<th>Paracetamol</th>
<th>Caffeine</th>
<th>Paracetamol</th>
<th>Caffeine</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>1</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>20</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>5</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>20</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>10</td>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>5</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>15</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>15</td>
<td>20</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>15</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>13</td>
<td>15</td>
<td>10</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>20</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>15</td>
<td>20</td>
<td>20</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>15</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Binary Drug System: Paracetemol/Caffeine (PAR/CAF)

\[
RMSEP_j = \sqrt{\frac{\sum_{i=1}^{\nu} (\hat{y}_{i(valid)}^j - y_{i(valid)})^2}{\nu - A}} \tag{21}
\]

- \(\nu \Rightarrow\) the number of samples in validation set.
- \(\hat{y}_{i(valid)}^j = x_{i(valid)}^\top \hat{\beta}_j\) for \(j = \text{RWSIMPLS, SIMPLS, PRM}\) with \(\hat{\beta}\) estimated from the calibration sample.
Figure: Root Mean Square Error for Prediction (RMSEP) vs The Number of Components for PAR/CAF data
Figure: Coefficient Estimates for PAR/CAF Data with $A=2$

Table: The Scale Estimates For PAR/CAF Drug System with $A=2$

<table>
<thead>
<tr>
<th></th>
<th>RWSIMPLS</th>
<th>SIMPLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAR</td>
<td>0.2170177</td>
<td>0.2300721</td>
</tr>
<tr>
<td>CAF</td>
<td>0.2818257</td>
<td>0.2923287</td>
</tr>
</tbody>
</table>
Figure: RWSIMPLS Diagnostics for PAR/CAF Drug System

(a) Residual Distance vs. Score Distance

(b) Orthogonal Distance vs. Score Distance

(c) Cook's Distance vs. Score Distance
Figure: RWSIMPLS Weights for PAR/CAF Drug System
Binary Drug System: Paracetemol/Caffeine (PAR/CAF)

(a) SIMPLS

(b) SIMPLS

(c) SIMPLS

Figure: SIMPLS Diagnostics for PAR/CAF Drug System

A Alin
SDSS 2018 16-19 May 2018
Binary Drug System: Paracetemol/Caffeine (PAR/CAF)

Thanks for your attention! Any questions? ...


