Class-modelling vs Discriminant classification

A practitioner reflections

Dr. Marina Cocchi
Researcher Analytical Chemistry (Chemometrics)
outline

- The one class modeling context
- Retrospective on SIMCA
- PLS-DA and its implementation
The one class modeling context

Classification task

Class Modeling

› single class information is used
› useful in authentication

Discriminant methods

› other/others class/es information always used
The one class modeling context

Why class-modeling?

- e.g. Authentication issue

Need to identify non-conformity, adulterations...

- Sometimes \textit{we know what we want!!!}
  - there is a category to contrast with

- Most often \textit{we would like to know!!!}
  - to find if an adulterant is present even not knowing what it could be

Chemical Signature

- we expect something to change in the signal profiles if an "unknown" constituent is added
The one class modeling context

the “unknown” has similar profile but some peculiar features

modeled class/es

- in model space the “unknown” will be accepted /close to class but it will have high residuals

A discriminant method will assign it to one of the classes by definition

- Does the paradigm one class against the rest of words make sense? (e.g. ill vs healthy)
The one class modeling context

Distinguish different valued Italian products

Modeled categories

<table>
<thead>
<tr>
<th>CLASS</th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liguria</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>Apulia</td>
<td>12</td>
<td>10</td>
</tr>
</tbody>
</table>

ECVA model 3LV in inner PLS*

- how the models consider “not Italian samples”?

* model dimensionality according to minimum CV classification rate or **CV Efficiency
The one class modeling context

ECVA weights

+ Apulia
- Liguria
different concentration common

SIMCA PC1
loadings Liguria
main variance for liguria

std for each category

Apulia
Liguria
not Italian

Weekend, March 16, 2014
The one class modeling context

discriminant PLS model 3LV

The model is built on 2 classes Apulia and Liguria.

- how the models consider “not italian samples”? 
The one class modeling context

discriminant PLS model 3LV

again the model is built on 2 classes Apulia and Liguria

- how the models consider “not italian samples”?
SIMCA methods

1. Build a distinct PCA model for each class (separate centering/scaling & dimensionality choice)

2. Define a distance measure to the class model

3. Define a limit/boundary for acceptability (classification rule)

4. For a “new” sample estimate distances from each class model

5. objects may be assigned to one, more than one, or none of the classes
**SIMCA critical/differing steps**

2. Define a distance measure to the class model

- **Common frame**
  - \( SD = \text{Scores} \text{ Distance} \)
  - distance within model space
  - Use PCA scores

- **OD = Orthogonal Distance**
  - distance from model space
  - Use Residuals

- **Variations**
  - how they are defined
  - \( SD \)
    - Euclidean / Mahalonobis / Leverage
    - from origin / boundaries
  - \( OD \)
    - DoF correction
    - Use CV residuals and scores or not
    - Use both / only one of the two
    - how to combine

3. Define limit(s)/boundary(ies) for acceptability (classification rule)

- **alternative-SIMCA**
  - Use \( SD \) & \( OD \) distributions/statistics
  - same / different
  - Use \( SD \) & \( OD \) limits Robust / distributions free

- **original-SIMCA**
  - combine \( SD \) & \( OD \) for “external” object in a single “distance to class” measure: \( D \)
  - Use F-test to compare \( D \) with class residuals variance (\( RSD \))

**Retrospective on SIMCA**

Sunday, March 16, 2014
Implementation I

\[ d_p^{(q)} = \sqrt{s_p^{(q)} + \sum_a \Phi_a^2 (t_a - \theta_a^{(q),\text{lim}})^2} \]

\[ s_p^{(q)} = \sqrt{\sum_k e_{pk}^2 / (M-A)} \]

Distance from a class, i.e. \( q \), of a test object, i.e. \( p \)

Total RSD of a class, i.e. \( q \):

\[ s_0^{(q)} = \sqrt{\sum_{ik} e_{ik}^2 / (N-A-1)(M-A)} \]

Classification rule

\[ F = d_p^{(q)2} / s_0^{(q)2} < F_{\text{crit}} (M-A),(N-A-1)(M-A) \]

If true for both \( q \) and \( r \) Unique assignment only if

\[ F = d_p^{(q)2} / d_p^{(r)2} > F_{\text{crit}} (M-A_r),(M-A_q) \]

\[ \Phi_a = s_p^{(q)2} / \left( \sum_k t_{ak}^{(q)2} / N_q \right) \]

\[ SD = 0 \text{ for calibration samples} \]

\[ SD \text{ is weighted by } \Phi_a^2 \]

\[ \sum \Phi_a^2 \]

\[ \theta_a^{(q),\text{lim}} \text{ can be: } t_{\text{max}}, t_{\text{max}} + \text{std}(t), \text{CI}(t) \]

\[ F \text{-test because it is a comparison of variances} \]

DoF, corrections:

\[ F = \frac{d_p^{(q)2}/s_0^{(q)2}}{s_0^{(q)2}} \]

- correct \( F \ast \frac{N-1}{N-A-1} \)

\[ F_{crit} = \left( M-A \right), \frac{(N-A-1)(M-A)}{(M-A) (N-A-1)(M-A)} \]

Variants:

- \( \frac{1}{2}[(N-A-1)(M-A)] \) (to enlarge acceptance)
- Use: \( (r-A)(N-A-1)/N, \frac{(N-A-1)(r-A)}{N} \)
  \( r = M \) if \( N>M \)
  \( r = N-1 \) if \( N<=M \)

- correct \( F \ast \frac{N}{N-A-1} \) and use \( F_{crit} 1, \frac{(N-A-1)}{(N-A-1)} \)

See references 15-18 in [1]

**Implementation in Software**

**SIMCA**

- Umetrics
  - **Classification rule**
    - \( \left( \frac{s_i^{(q)}}{s_0^{(q)}} \right) \) is approximately F-distributed with DoF of observation and the model

- SIMCA
  - **pooled RSD of a class, i.e.** \( q: \)
    - \( s_0^{(q)} = \sqrt{\Sigma e_{ik}^2/(N-A-1)(M-A)} \)
  - **absolute DModX**
    - \( s_i^{(q)} = \sqrt{\Sigma e_{ik}^2/(M-A)} \times v \)
  - **(calibration samples)**
    - where \( v \) is a correction factor slightly higher than 1
  - accounting for the fact that distance to model is expected to be slightly smaller for a sample that is part of it

- **absolute DModXPS**
  - \( s_p^{(q)} = \sqrt{\Sigma e_{pk}^2/(M-A)} \)
  - ("external" samples), i.e. OD
  - \( d_p^{(q)} = \text{augmented DModXPS}^{+}(q) = \sqrt{s_p^{(q)}^2 + SD^2} \)

**SIMCA**

- PARVUS
  - Classification Rule is as in Umetrics but correct \( F*(N-1)/(N-A-1) \)
  - Moreover variants are introduced in:
  - \( \vartheta_{a}^{(q)}\),lim can be restricted
  - \( D \) is considered as the hypotenuse from class box
  - Residuals are corrected by leverage

---

Retrospective on SIMCA

Sunday, March 16, 2014
Implementation in Software

SIMCA Unscrambler

Classification Membership Limits

- Leverage Limit
  \[ \leq 3(A+1)/N \]

- Sample to model distance Limit
  \[
  \left( \frac{s_p^{(q)}}{s_0^{(q)}} \right)
  \]
  is approximately F-distributed

Leverage correction: \( F_{crit} 1, (N-A-1) \)

Cross Validation/Test: \( F_{crit} 1, N_{val} \)

\[ s_0^{(q)} = \sqrt{\text{ResXValTot}} \]
\[ = (?) \sqrt{\sum_{ik} e_{ik}^2 / [N_{val} * (M-A)/M]} \]

ModelDist\(^{(q)}\) =

\[ s_p^{(q)} = \sqrt{\sum_k e_{pk}^2 / (M-A)} \]

("external" samples)

i.e. OD

not clear to me how the two limits are used
alternative SIMCA

Common frame

**SD** = **Scores** **Distance**
use Mahalonobis Distance / Hotelling's T-square ($T^2$)

**OD** = **Orthogonal** **Distance**
$Q$ is the sum of squared residuals: $\Sigma_k e_{pk}^2$ (same role as $s_p^{(q)2}$)

$Q_{lim}$ is calculated by assuming a $\chi^2$ distribution

Classification rule

Use a summary of reduced distance

$SD/SD_{lim}$ & $OD/OD_{lim}$

Variations

- Which Reference Distribution for SD
- Use of Robust estimation
- Use CV scores

Variations

- Unique assignment /not
Implementation II

alternative SIMCA

© PLS-Toolbox SIMCA

\[ Q = \sum_k e_{pk}^2 \]

\[ Q_{\text{lim}} \chi^2 \text{ distribution JM approximation} \]

\[ T_i^2 = t_i \lambda^{-1} t_i \]

Hotelling’s T-square statistics

\[ T_{\text{lim}}^2 = [A*(N-1)/N*(N-A)] F_{\text{crit}} A, (N-A) \]

Classification rule

Assign an object to a class if its reduced combined distance satisfies:

\[ \sqrt{\left( \frac{Q}{Q_{\text{lim} \text{ ite}}} \right)^2 + \left( \frac{T^2}{T_{\text{lim}}^2} \right)^2} < \sqrt{2} \]
variations on Theme

alternative SIMCA

robust SIMCA [1,2]
- use robust PCA for class models
- use a weighted sum of “reduced” orthogonal (OD, i.e. $s_p^{(q)2}$ in previous slide) and Mahalonobis (MD, in scores space) distances
- thus classification rule will be to assign a new sample to the class for which $R$ is minimal:

$$R = \gamma \frac{OD}{OD_{\text{lim}}} + (1 - \gamma) \frac{MD}{MD_{\text{lim}}}$$

where $MD_{\text{lim}}$ is estimated by using as reference the chi-squared distribution with $dof$ equal to the number of retained components (A) and $OD_{\text{lim}}$ is estimated by assuming that a scaled version of chi-squared distribution is appropriate (Wilson-Hilferty approximation)

The same method as for SD is used:

\[ h_i = t_i^T (T_A^T T_A)^{-1} t_i = \sum_{a=1}^{A} \frac{t_{ia}^2}{\lambda_a}, \quad i = 1, \ldots, I \]

mean Leverage

\[ h_0 = \frac{1}{I} \sum_{i=1}^{I} h_i = \frac{A}{I}, \]

Scores Distance = Leverage

The author estimated DoF by the method of moment

\[ \hat{N}_h = \frac{2 h_0^2}{S_h} = \frac{2A^2}{I S_h} \]

The author prefers a Robust estimate:

\[ \frac{1}{N_h} \left[ \chi^{-2}(N_h, 0.75) - \chi^{-2}(N_h, 0.25) \right] = \frac{1}{h_0} \text{IQR}(h_1, \ldots, h_I) \]

\[ \hat{N}_h = \exp \left( 4.36 \ln \frac{1.24}{\text{IQR}} \right)^{0.72} \]


alternative SIMCA
**variations on Theme**

**alternative SIMCA**

- **Pomerantsev acceptance area [1]**

Scores Distance = Leverage

\[ h_i = t_i^T(T_A^T T_A)^{-1} t_i = \sum_{a=1}^{A} \frac{t_i^2}{\lambda_a}, i = 1, \ldots, l \]

mean Leverage

Orthogonal Distance

\[ v_i = \sum_{j=1}^{J} e_{ij}^2, \quad v_0 = \frac{1}{l} \sum_{i=1}^{l} v_i \]

The same method as for SD is used:

\[ N_v v / v_0 \sim \chi^2(N_v) \quad \hat{N}_v = \frac{2v_0^2}{S_v} \]

**SD distribution**

DoF = \( N_h \)

DoF = \( A \) if scores are normally distributed

The author estimated DoF by the method of moment

\[ V(h) = \frac{2h_0^2}{N_h} = \frac{2A^2}{N_hI^2} \]

\[ \hat{N}_h = \frac{2h_0^2}{S_h} = \frac{2A^2}{I^2S_h} \]

\[ S_h = \sum (h_i - h_0)^2 / (l-1) \]

The author prefers a Robust estimate:

\[ \frac{1}{N_h} [\chi^{-2}(N_h, 0.75) - \chi^{-2}(N_h, 0.25)] = \frac{1}{h_0} \text{IQR}(h_1, \ldots, h_l) \]

\[ \hat{N}_h = \exp \left( 4.36 \ln \frac{1.24 \text{IQR}}{\text{IQR}} \right)^{0.72} \]


marina.cocchi@unimore.it

Arctic Analysis 10-15 March 2014

Dr. Marina Cocchi

Dpt Chemical and Geological Sciences - UniMORE

Sunday, March 16, 2014
I, II and V were considered acceptable

I

$$H_y = \left[ 0, \frac{h_0}{N_h} \chi^{-2}(N_h, \alpha) \right] \otimes \left[ 0, \frac{V_0}{N_V} \chi^{-2}(N_V, \alpha) \right].$$

II

$$H_y = \left\{ (h, v) : N_h \frac{h}{h_0} + N_V \frac{v}{V_0} \leq \chi^{-2}(N_h + N_V, \gamma) \right\}.$$
Orthogonal Distance: $Q$

Scores Distance:

Leverage:

$H_{fit} = \text{diag} \left[ T_{fit} (T_{fit}^TT_{fit})^{-1}T_{fit}^T \right]$

$H_{CV} = \text{diag} \left[ T_{CV} (T_{fit}^TT_{fit})^{-1}T_{CV}^T \right]$

$S$: scores Covariance matrix; $T$: Scores matrix;

$D$-statistic

$D_{fit} = \text{diag} \left[ T_{fit} (S_{it})^{-1}T_{fit}^T \right]$

Classification rule (reduced distances as in PLS toolbox)

$\cdot$ $Q$, $H$ limits:

$Q_{lim \, fit} \sim \chi^2$ distribution JM approximation

$H_{lim \, fit}^{[2]} \sim h_i N/(N-1) \sim \beta(M/2, (N-M-1)/2)$

$D_{lim \, fit}^{[1]} \sim [A(N^2 - 1)/N(N-A)] F(A, N-A)$

H CV 95%: the 95 percentile of $H_{CV}$

Q CV 95%: the 95 percentile of $Q_{CV}$

---


Implementation I & II comparison

**original in SIMCA (I):**
- score and orthogonal distances are combined as a “rsd” measure
- square distance/rsd of a “new” sample is compared to square rsd of class by F-test
- Score distance is computed from the “box” boundaries (but can be taken from average problematic
- degree of freedom for F-test
- scaling factor needed in order to combine score distance with orthogonal one

**alternative in PLS-Toolbox SIMCA (II):**
- reduced score distance and orthogonal distance are combined, thus being directly comparable
- Score distance is computed from the average problematic
- two different reference distribution are assumed
- degree of freedom (?)
- giving the same weight to both type of distances

**Variants**
- using $T^2$, MD avoids to define “box” boundaries
- using only Orthogonal distances does not take into account deviations inside class space
- using $\chi^2$ distribution for both OD and SD
- do not combine OD and SD in a single rule
$D_{\text{lim}} \text{ fit}^{[1]} \sim \left[ \frac{A(N^2 - 1)}{N(N-A)} \right] F(A, N-A)$
Distances:

✓ Use a modified Mahalonobis distance defined by using both **SD** and **OD**

\[
\text{Bayes rule: } \quad D_{\text{DASCO}} = D_{P_k}^2 + D_{C_k}^2 + \ln |C_k| - 2 \ln \pi_k
\]

\[
\begin{aligned}
\text{DASCO: } \quad \min D_{\text{DASCO}} &= D_{P_k}^2 + D_{C_k}^2 + \ln |C_k| - 2 \ln \pi_k \\
&= \min_k \left[ (x - \bar{x}_k)^T C_k^{-1} (x - \bar{x}_k) \\
&\quad + \ln |C_k| - 2 \ln \pi_k \right]
\end{aligned}
\]

Regression on dummy variables (DPLS, PLS-DA)

Y coding: as many dummy y-variables as classes 1/0
Fit a PLS2 model

DPLS

Predict the dummy values ($y_{i,\text{pred}}$) and assign the object to the group with highest predicted value, $k = \arg\max(y_{i,\text{pred}})$

with more than 2 categories can be sub-optimal (masking effect Hastie\(^1\))
a better alternative would be to use LDA, QDA, etc. on PLS scores

PLS-DA

It has been demonstrated PLS on dummy variables has connection with LDA\(^2\) or with Fisher canonical discriminant analysis, in fact it correspond to eigendecomposition of a slightly altered version (M) of the Between-group variance-covariance matrix (B)

$$M_{jl} = \sum_{k=1}^{q} n_k^2 \bar{x}_{j(k)} \bar{x}_{l(k)},$$

$$B_{jl} = \sum_{k=1}^{q} \frac{n_k}{n} \bar{x}_{j(k)} \bar{x}_{l(k)}.$$

Nocairi\(^3\) et al. implemented PLS-DA bringing to eigendecomposition of B

\(^2\) Barker and Rayens., J. Chemometrics 2003 (17): 166-173
\(^3\) Nocairi et al. Comp. Stat & Data Anal. 2005 (48) 139-147
Regression on dummy variables (DPLS, PLS-DA)

Y coding: as many dummy y-variables as classes 1/0
Fit a PLS2 model

PLS-DA extension
A general framework has been proposed\(^1\) by including prior probability in the derivation

Given a classification problem with specified prior probabilities \(\Pi\), a feasible strategy is to extract PLS loading weights according to the dominant eigenvector of \(B_\Pi\), found by SVD of:

\[
\Delta Y^t X
\]

\[
\Delta = \sqrt{\Pi (Y^t Y)^{-1}}
\]

Moreover using LDA on PLS2 scores or Y fitted values (not redundant, i.e. with k column) give the same results

---


* X has to be centered by the weighted global means taking into account the prior
- Y coding: as many dummy y-variables as classes 1/0
- Fit a PLS2 model

Classification rule

- define a threshold
  - **UNSCRAMBLER**: fixed, e.g. 0.5; (Also suggest to use LDA... on PLS scores)
  - **SIMCA-Umetrics**: <0.35 reject; 0.35-0.65 border-line; 0.65-1.35 accept; >1.35 do not belong, possible extreme/outlier
  - **PLS-toolbox**: Threshold selected on best SENS/SPEC compromise from ROC curve (CV
  - Bayesian threshold

The Bayesian threshold calculation assumes that the predicted y values follow a distribution similar to what will be observed for future samples. Using these estimated distributions, a threshold is selected at the point where the two estimated distributions cross; this is the y-value at which the number of false positives and false negatives should be minimized for future predictions (either in fit or CV prediction).
Practitioners reflections:

- there is a clear imbalance between literature and implementations
- there is a need for tutorials (back to basic.. but a little beyond...)
- need to define a problem oriented strategy, e.g. modeling vs. discrimination, etc..

for discussion:

- using the threshold rule(s) in the DPLS spanned space how to relate to the theoretical FCDA framework or other?
- when there are more than 2 classes an “hybrid” use of threshold can be found, i.e. use it as class boundaries and treat results as class modeling... make sense (??)
- to interprete PLS weights, regression coefficients in the usual way forgetting Y is a dummy matrix is correct ?
- what about target projection and SR in the discriminant PLS framework ?

- could we think of unified/generalized SIMCA making a synthesis of the different derivations ?
Thanks for attention