Paper B800104A

Improving the Linearity of Spectroscopic Data Subjected to Fluctuations in External Variables by the Extended Loading Space Standardization

1) Spectral Standardization using LSS

Assume the rows of $\mathbf{X}_{I\times J}(t_1)$, $\mathbf{X}_{I\times J}(t_2)$, …, and $\mathbf{X}_{I\times J}(t_N)$ are the corresponding spectra for the same I training mixture samples measured at training temperatures t_1, t_2, , , …, t_N (I and J are the number of samples and discrete wavelength positions, respectively).

Suppose $\overline{\mathbf{X}} = \sum_{n=1}^{N} \mathbf{X}_{I\times J}(t_n)/N$, the singular value decomposition of $\overline{\mathbf{X}}$ can be expressed as:

$$\overline{\mathbf{X}} = [\mathbf{U}_c, \mathbf{U}_r]\begin{bmatrix} \mathbf{\Lambda}_c^{1/2} & \\ & \mathbf{\Lambda}_r^{1/2} \end{bmatrix}[\mathbf{V}_c, \mathbf{V}_r]' = \mathbf{T}_c\mathbf{P}_c + \mathbf{T}_r\mathbf{P}_r \quad (1)$$

$$\mathbf{T}_c = \mathbf{U}_c\mathbf{\Lambda}_c^{1/2},\ \ \mathbf{T}_r = \mathbf{U}_r\mathbf{\Lambda}_r^{1/2},\ \ \mathbf{P}_c = \mathbf{V}_c',\ \ \mathbf{P}_r = \mathbf{V}_r' \quad (2)$$

Where the subscripts '*c*' and '*r*' signify the first C principal factors representing the spectral information and the rest factors representing noise, respectively. Superscript ''' denotes the transpose. Assuming Beer's law is valid, then $\overline{\mathbf{X}} = \mathbf{Y}\overline{\mathbf{S}} + \mathbf{E}$ (where $\mathbf{Y}$ denotes the concentration matrix; $\overline{\mathbf{S}} = \sum_{n=1}^{N} \mathbf{S}(t_n)/N$ are the mean of the pure spectra matrices that are not available for grey chemical systems; and $\mathbf{E}$ is the matrix of residuals). It can be proven that there exists a full rank matrix $\mathbf{R}$ that satisfies equation 3:

$$\mathbf{Y} = \mathbf{T}_c\mathbf{R} \quad (3)$$

Therefore, the following equation holds:

$$\mathbf{X}(t_n) = \mathbf{Y}\mathbf{S}(t_n) + \mathbf{E}(t_n) = \mathbf{T}_c\mathbf{R}\mathbf{S}(t_n) + \mathbf{E}(t_n) \quad (4)$$

Letting $\mathbf{P}(t_n) = \mathbf{R}\mathbf{S}(t_n)$ (matrix $\mathbf{P}(t_n)$ has a size of $C\times J$. where C is the number of principal components representing spectral information):

$$\mathbf{P}(t_n) = \mathbf{R}\mathbf{S}(t_n) \approx (\mathbf{T}_c)^{+}\mathbf{X}(t_n), \quad n = 1,2,\cdots,N \quad (5)$$

Here, superscript '+' denotes the Moore-Penrose Matrix Inverse.

In order to model the temperature effects on loading matrixes, a polynomial can be fitted to each element of $\mathbf{P}(t_n)$ (n = 1, 2, …, N) against the temperature t_n:

$$p_{i,j}(t_n) = \alpha_{i,j} + \beta_{i,j}t_n + \gamma_{i,j}t_n^2,\ \ i = 1,2,\cdots,C,\ \ j = 1,2,\cdots,J,\ \ n = 1,2,\cdots,N \quad (6)$$

Parameters $\alpha_{i,j}$, $\beta_{i,j}$ and $\gamma_{i,j}$ can be estimated out as follows:

$$\mathbf{p} = \mathbf{M}\mathbf{A}$$
$$\mathbf{p} = \left[p_{i,j}(t_1);\cdots;p_{i,j}(t_n);\cdots;p_{i,j}(t_N)\right],\ \ \mathbf{t} = \left[t_1;\cdots;t_n;\cdots;t_N\right], \mathbf{M} = \left[\mathbf{1},\mathbf{t},\mathbf{t}^2\right],\ \ \mathbf{A} = \left[\alpha_{i,j};\beta_{i,j};\gamma_{i,j}\right] \quad (7)$$

Here, the regressor matrix $\mathbf{M}$ is known and expected to have full column rank, the parameter matrices $\mathbf{A}$ can then be simply calculated out as $\mathbf{A} = (\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{p}$.

The above procedure can be applied to all the elements of loading matrices $\mathbf{P}(t_n)$ (n = 1, 2, …, N). After all the parameters $\alpha_{i,j}$, $\beta_{i,j}$ and $\gamma_{i,j}$ ($i = 1,2,\cdots,C$, $j = 1,2,\cdots,J$) have been estimated out, the loading matrix $\mathbf{P}(t)$ at any temperature t ($t_1 \le t \le t_N$) (which should be measured along with the spectrum of the sample) can be calculated out as: $p_{i,j}(t) = \alpha_{i,j} + \beta_{i,j}t + \gamma_{i,j}t^2$

For a spectrum $\mathbf{x}(t)$ measured at temperature t, it can then be standardized into an arbitrarily selected reference temperature t_{ref} ($t_1 \le t_{ref} \le t_K$) as if it were measured under the reference temperature:

$$p_{i,j}(t_{ref}) = \alpha_{i,j} + \beta_{i,j}t_{ref} + \gamma_{i,j}t_{ref}^2, \quad \mathbf{z}(t_{ref}) = \mathbf{x}(t)\mathbf{P}(t)^{+}(\mathbf{P}(t_{ref}) - \mathbf{P}(t)) + \mathbf{x}(t) \tag{8}$$

Where $\mathbf{z}(t_{ref})$ denotes the standardized spectrum.

2) MATALAB code for the estimation of multiplicative parameters by OPLEC

```
%  [b] = OPLEC(Z, y, CompNumb);
%  This is an m-file to find the multiplicative scattering parameter vector b for calibration samples;
%  Z contains z_i in its rows; z_i (i = 1,2,...,m) are the standardized spectra of calibration samples.
%  y is the concentration vector of the target chemical component in calibration samples;
%  CompNumb is the number of chemical components in mixture samples;
%  b is a vector containing the multiplicative scattering parameters for calibration samples;

function [b]=OPLEC(Z,y,CompNumb);

[m,n]=size(Z);
Y=[ones(m,1),y];
b=zeros(m,m);

for i=1:m

    OriginalPositionIndex=(1:m)';
    Z1=Z(i,:);
     PositionIndex=i;

    for j=1:(CompNumb-1)
        NormVect=sqrt(sum(Z.^2,2));
        ResMatrix=diag(1./NormVect)*Z*(eye(n,n)-pinv(Z1)*Z1);
        ResVect=diag(ResMatrix*ResMatrix');
```

```
        [MaxRes,MaxPosition]=max(ResVect);
        Z1=[Z1;Z(MaxPosition,:)];
        PositionIndex=[PositionIndex,MaxPosition];

    end

    OriginalPositionIndex(PositionIndex)=[];

    ZZ=Z;
    ZZ(PositionIndex,:)=[];
    Z2=ZZ;

    Y11=Y(PositionIndex,1);
    Y21=Y(PositionIndex,2);
    YY=Y;
    YY(PositionIndex,:)=[];
    Y12=YY(:,1);
    Y22=YY(:,2);

    Regv=Z2*pinv(Z1);
    P1=diag(Y22)*Regv*diag(Y11);
    P2=Regv*diag(Y21);
    P11=P1(:,1);
    P12=P1(:,2:CompNumb);
    P21=P2(:,1);
    P22=P2(:,2:CompNumb);

    b0=lsqnonneg(P12-P22,P21-P11);
    b1=[1;b0];

    Q1=Regv*diag(Y11);
    Q2=Regv*diag(Y21);
    b2=diag(1./(Y12+Y22))*(Q1+Q2)*b1;

    b(PositionIndex,i)=b1;
    b(OriginalPositionIndex,i)=b2;

end

b=mean(b,2);

% After obtaining the model parameter vector b for calibration samples, two calibration models are
  built by standard PLS toolbox. One is between the concentration vector ( y ) of the target
```

chemical component and the spectral data **Z**, the other is between $diag(\mathbf{y})\mathbf{b}$ and **Z**. The multiplicative effect on the test sample can then be corrected through dividing the prediction of the second calibration model by the prediction of the first calibration model.