SUPPLEMENTRY INFORMATION

Chemometric assisted Correlation optimized warping of the chromatograms: Optimizing the computational time for correcting the drifts in the chromatographic peak positions

Keshav Kumar

Institute for Wine analysis and Beverage Research, Hochschule Geisenheim University, Geisenheim 65366, Germany

Email ID: keshavkumar29@gmail.com

Conceptual difference between CRNNFA, MCR-ALS and NNMF

NNMF [1, 2] and MCR-ALS [3-7] analysis are the two commonly used chemometric techniques to decompose a two-way array Y of dimension I×J (sample × variable). The decomposition of Y (I×J) with NNMF can be summarized as $Y=WH^T$. The matrices W (I×F) and H (J×F) explain the contribution and response of each of the F factors. The NNMF model ($Y=WH^T$) is fitted using an iterative algorithm and it involves initialization of both W and H matrices with random values. The NNMF model uses alternate least square criteria for convergence. There are two problems with NNMF algorithm (i) both W and H are initialized in random fashion therefore at several occasion NNMF algorithms wanders with W and H matrices that has no physical and chemical relevance and (ii) NNMF algorithm often gets lingered in the local minima and never provides the solutions that corresponds to true minima. It essentially means that algorithm does not converge to true solution and the obtained solutions do not reflect the real phenomena-taking place.

The decomposition of Y (I×J) with MCR-ALS can be summarized as $Y=CS^{T.}$ The matrices W (I×F) and H (J×F) explain the contribution and response of each of the F factors. MCR-

ALS also uses least square convergence criteria for the convergence. The fitting of MCR-ALS model requires initialization of either C or S matrix. The matrix C can be initialized using evolve factor analysis (EFA) [8, 9]. This approach has certain limitations on practical grounds and it is mainly so because EFA requires data must belong to the sequential process. The initialization of the matrix S for fitting the MCR-ALS model can be achieved using the pure variable approach e.g. SIMPLISMA algorithm [5, 6, 10]. This approach too has the practical limitations because it requires that there must be pure variable for each factors of the data set, often with real life samples due to the compositional complexity it is difficult to find pure variable for each factor. It can be easily realized that in both cases (i.e. EFA and pure variable approach), it is difficult to meet the essential pre-requisites and it is even more difficult in the case of real life samples belonging to the agricultural or biological samples. In addition to this, convergence of MCR-ALS algorithm is controlled by alternate least square criteria. Therefore, there is always a possibility that MCR-ALS algorithm lingers in the local minima and never attains the true convergence and thus obtained solution might not explain the actual chemical or biochemical phenomena-taking place.

As discussed above, CRNNFA also decomposes a two-way array as a product of two smaller matrices (i.e. $X=\Phi\Psi+E$). However, CRNNFA approach is different from these two approaches in the following sense

(i) It involves the initialization of the variables in a constraint manner. The variables are initialized with random number in a space spanned by the minima (i_{min}) and maxima (i_{max}) of each i^{th} variable. The constraint initialization of the variables are summarized in equation S1

$i^{th} \in [i_{max}, i_{min}] =(i_{max} - i_{min})*random\ number +i_{min}$ .... (S1)

(ii) It does not involve any convergence criteria. The analysis terminates only when the iteration limit is reached.

The constraint initialization of the variables ensures a good initial estimate for the subsequent iterations and it ensures the algorithm does not wander with random numbers having no chemical or biochemical significance. Unlike MCR-ALS, the initialization of the variables in CRNFFA approach does not require that data must come from sequential process or it should contain pure variables for each factor. The CRNNFA does not involve any convergence criteria. The analysis terminates only when the iteration limit is reached, thus, circumventing the issue of premature convergence. As a result, the CRNNFA obtained solutions will correlate well the real chemical biochemical phenomena-taking place in the analyzed samples.

MATLAB code and Commands

Inbuilt SVD MATLAB command can be used for finding the optimum number of factors.

```
[u,s,v]=svd(x,’econ’)
Plot(diag(s),’o-‘)
```

MATLAB code for carrying out the CRNNFA analysis

```
 x=input('data =');
maxit= input ('maximum iteration =');
iter=0
[m,n]=size(x);
f=input('f');
s=zeros(m,f);
for kk=1:f
for ii=1:m
min1= min (x(ii,:));
max1=max(x(ii,:));
s(ii,kk)=(max1-min1).*rand(1)+ min1;
end
end
while (iter < maxit)
c=(pinv(s))*x ;
for ll=1:f
for k=1:n
if c(ll,k)<0
c(ll,k)=10^-6;
end
```

```
end
end
s=x*pinv(c);
for i=1:f
s(:,i)=s(:,i)./(sum(s(:,i)));
end
for ll=1:f
for k=1:m
if s(k,ll)<0
s(k,ll)=10^-6;
end
end
end
for i=1:f
s(:,i)=s(:,i)./(sum(s(:,i)));
end
iter=iter+1;
end
```

The COW analysis can be performed using the code available from http://www.models.life.ku.dk/algorithms.

References

1. D. D. Lee and S. H. Seung, *Nature*, 1999, **401**, 788–791.

2 D. D. Lee and S. H. Seung, *Adv. Neural Inform. Process. Syst.*, 2001, **13**, 556–562.

3 R. Tauler, B. Kowalski and S. Fleming, *Anal. Chem.*, 1993, 65, 2040–2047.

4. R. Tauler, *Chemom. Intell. Lab. Syst.*, 1995, 30, 133–146.

5. T. Azzouz and R. Tauler, *Talanta*, 2008, 74, 1201–1210.

6. M. Garrido, F. X. Rius and M. S. Larrechi, *Anal. Bioanal. Chem.*, 2008, **390**, 2059– 2066.

7. G. Ahmadi, R. Tauler and H. Abdollahi, *Chemom. Intell. Lab. Syst.*, 2015, 142, 143–115.

8. M. Maeder and A. Zilian, *Chemom. Intell.Lab. Syst.*, 1998, 3, 205–213.

9. B. M. Wise, N. B. Gallaghar, R. Bro and J. M. Shaver, PLS_Toolbox 4.0, Eigenvector Research, 2006.

10. W. Winding and J. Guilment, *Anal. Chem.*, 1991, **63**, 1425–1432.