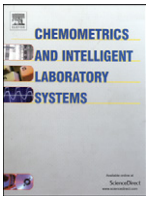
Chemometrics and Intelligent Laboratory Systems 93 (2008) 108–119

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab



Identification of unknown pure component spectra by indirect hard modeling

E. Kriesten [a], D. Mayer [a], F. Alsmeyer [b], C.B. Minnich [c], L. Greiner [c], W. Marquardt [a,*]

[a] *AVT - Process Systems Engineering, RWTH Aachen University, Turmstraβe 46, 52064 Aachen, Germany*
[b] *AixCAPE e.V., Turmstraβe 46, 52064 Aachen, Germany*
[c] *Institute for Technical and Macromolecular Chemistry, RWTH Aachen University, Worringerweg 1, 52074 Aachen, Germany*

ARTICLE INFO

Article history:
Received 4 September 2007
Received in revised form 21 April 2008
Accepted 5 May 2008
Available online 13 May 2008

Keywords:
Spectroscopic analysis
Indirect hard modeling (IHM)
Self modeling curve resolution (SMCR)
Nonlinear optimization
Peak fitting
Hard modeling
Factor analysis
Identification
Collinearity

ABSTRACT

Indirect hard modeling (IHM) is a physically motivated spectral analysis principle. It utilizes nonlinear spectral hard models generated by peak fitting of the pure spectra. This approach allows the consideration of various nonlinear effects such as peak variations or spectral shifts. Compared to established methods, less calibration samples are required and basic calibration transfer is performed inherently. To extend the applicability of IHM, which currently requires knowledge of the pure component spectra, two methods for the identification of pure spectra are presented in this work. These methods work automatically on a mathematically objective basis and do thus not depend on the expertise of the user. As IHM relies on an underlying physical picture of the spectra, the relevant information in the input data is exploited very efficiently especially for selective spectra, and nonideal spectral behavior is captured throughout the identification process. Compared to established SMCR methods the number of required spectra is reduced. The first method, complemental hard modeling (CHM), is introduced for the case that a single pure spectrum is unknown. The method is based on a deconvolution approach and only requires a single mixture spectrum as input data. The second method, hard modeling factor analysis (HMFA), is conceptually related to SMCR methods. It allows the identification of all pure spectra in a completely unknown mixture from a limited set of mixture spectra. As shown in this work, even highly collinear data can be employed.

1. Introduction

Today, spectroscopy is an established noninvasive measurement technique utilized in diverse industrial applications. For the analysis of the spectral raw data, mainly linear multivariate methods such as principal component analysis (PCA) or partial least squares (PLS) are applied [1]. However, most chemical systems of practical interest are not ideal and behave nonlinearly. Examples for nonlinearities are peak shape variations due to molecular interactions or shifts of the full spectrum due to variations of the spectrometer adjustments. These effects can only be modeled to some extent, e.g. by application of laborious preprocessing or calibration transfer methods [2,3].

To overcome these limitations the authors have introduced a novel analysis method, called indirect hard modeling (IHM), which combines the advantages of multivariate linear methods with those of hard modeling techniques [4,5]. In this context, hard modeling techniques are understood as methods that exploit the physical structure of spectra by modeling the data with peak shaped functions, i.e. by peak fitting, which was first applied to spectroscopy in 1953 [6]. A good overview of recent methods is given in [7–9]. IHM represents a mixture spectrum as a weighted sum of parameterized pure component models built by hard modeling. The pure component weights are estimated in a nonlinear model fit that takes nonlinear effects into account by allowing for a constrained variation of the hard model parameters. This allows the use of a simple linear calibration model with high extrapolation capabilities to predict the concentrations. IHM has been successfully applied in several fields ranging from ATR measurements in reactive mixtures to 1D-Raman measurements in liquid mixtures [10–12].

A requirement of IHM is the availability of all pure component spectra. If some pure components are not known or cannot be measured at acceptable cost, they have thus to be identified with suitable methods such as self modeling curve resolution (SMCR). SMCR provides a useful tool for exploring multicomponent phenomena in complex systems in case that only limited a priori information on the data is available [13]. Two-way multivariate data changing due to some physical reason over time is deconvolved into factors for single species. Spectral data is thus factorized in concentration profiles and pure component spectra. For this reason SMCR can be directly utilized for the purpose of identifying pure component spectra. In a comprehensive review article [13], Jiang, Liang and Ozaki divide SMCR methods into two groups: *unique resolution methods* such as evolving factor analysis (EFA) [14] or window factor analysis (WFA) [15] and *rational solution methods* such as multivariate curve resolution (MCR-ALS) [16–18] or SIMPLISMA [19]. To increase the uniqueness of the solution, both groups apply certain constraints to the data. Unique resolution methods require the user to define unique feature regions. These are

---

* Corresponding author.
E-mail address: wolfgang.marquardt@avt.rwth-aachen.de (W. Marquardt).

0169-7439/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2008.05.002

either selective wave number regions where only a single component contributes to the spectrum or zero concentration regions where certain components do not exhibit spectral intensities. To some degree the quality of the results is thus dependent on the expertise of the user. For rational solution methods time series data with preferably weakly correlated concentration profiles are most suitable, as the accuracy of the solution depends on the degree of collinearity among the pure component profiles [13]. Some methods also require initial guesses for either the concentration profiles or the pure component spectra. To further enhance the mathematical uniqueness of the solution, non-negativity or unimodality constraints can be imposed [20]. This requires generic knowledge about the variables and a certain data structure. For multivariate measurements on mixtures with varying compositions this is naturally satisfied [13]. The user-friendly implementation of rational solution methods has made it a common practice in real-world chemical applications [13].

In this work, two novel methods for the identification of unknown pure component spectra are presented that work automatically on a mathematically objective basis. Both methods are based on the IHM approach, i.e. they exploit physical information in the data by using peak-based nonlinear spectral hard models. For this reason they are especially suited for selective spectra and can even be applied to highly collinear data. The first method, complemental hard modeling (CHM), is suited for the identification of a single unknown pure spectrum in an otherwise completely known mixture. Here, a single mixture spectrum is sufficient as input. The second method, hard modeling factor analysis (HMFA), permits the identification of all pure component spectra even from few unordered data with highly correlated concentration profiles. Initial guesses for concentration profiles or pure spectra are not required, and no constraints have to be selected. This method can be applied to all systems that exhibit at least one "distinctive" peak per component, i.e. a peak that does not occur in any other pure component spectrum. However, the distinctive peak is not required to stand free but may be highly overlapped even by peaks of other pure components. This distinctive peak is automatically identified by HMFA.

After a short introduction to IHM in Section 2 the CHM and HMFA methods are presented in detail in Section 3. In Section 4, the methods are validated with three increasingly challenging data sets: an uncorrelated time series, a highly correlated time series, and a very small data set containing only 8 independent measurements. Both, the quality of the identified spectra and the concentration prediction quality of IHM using the identified spectra are assessed. Finally, conclusions are given in Section 5.

2. Indirect hard modeling

The quantitative spectral analysis method IHM [4,5] proceeds in two steps, as depicted in Fig. 1. In the first step, a phenomenologically motivated nonlinear spectral model is fitted to the data to correct the nonlinear effects and to determine the pure component weights. A linear calibration model calculates the concentrations from the weights in the second step. Both, the phenomenological nonlinear spectral model and the calibration model are built during calibration. This process requires that all pure spectra are available, the number of pure components is known and at least one calibration mixture is measured. The whole procedure will be described in more detail in the following.

The nonlinear mixture model

$$\mathcal{X}(v,\boldsymbol{w},\boldsymbol{\theta}) = \mathcal{B}(v,\boldsymbol{\theta}_B) + \sum_{k=1}^{K} w_k \mathcal{S}_k\left(v,\boldsymbol{\theta}_{P,k},\theta_{S,k}\right) \tag{1}$$

is composed of a weighted sum of K parameterized pure component submodels $\mathcal{S}_k(\cdot)$ and a baseline function $\mathcal{B}(\cdot)$, which usually is a polynomial and thus linear in the parameters. Besides the pure component weights there are 3 sets of model parameters. The baseline parameters $\boldsymbol{\theta}_B$, the peak parameters of the pure component models $\boldsymbol{\theta}_{P,k}$, and the shift parameters $\theta_{S,k}$, which serve to shift the individual pure component spectra as a whole. The vector $\boldsymbol{\theta}$ concatenates all model parameters, i.e. $\boldsymbol{\theta}=(\boldsymbol{\theta}_B^T,\boldsymbol{\theta}_P^T,\boldsymbol{\theta}_S^T)^T$ with $\boldsymbol{\theta}_P=(\boldsymbol{\theta}_{P,1}^T,\ldots,\boldsymbol{\theta}_{P,K}^T)^T$ and $\boldsymbol{\theta}_S=(\theta_{S,1}^T,\ldots,\theta_{S,K}^T)^T$. The pure component models are generated by hard modeling during calibration and can be expressed by

$$\mathcal{S}_k\left(v,\boldsymbol{\theta}_{P,k},\theta_{S,k}\right) = \sum_{l=1}^{L_k} \mathcal{V}_l\left(v-\theta_{S,k},\boldsymbol{\psi}_{k,l}\right) \tag{2}$$

as sums of L_k Voigt functions $\mathcal{V}_l(\cdot)$. Each peak function is parameterized by four parameters $\boldsymbol{\psi}_{k,l}=(\alpha_{k,l},\beta_{k,l},\gamma_{k,l},\omega_{k,l})^T$, the width γ, the height α, the position ω, and the Gauss–Lorentz weight β. Except for the height α, all peak parameters are nonlinear parameters of the spectral model $\mathcal{X}(\cdot)$. The vector $\boldsymbol{\theta}_{P,k}=(\boldsymbol{\psi}_{k,1}^T,\ldots,\boldsymbol{\psi}_{k,L_k}^T)^T$ concatenates all individual peak parameters.

In summary, IHM builds the mixture spectrum hard model (Eq.(1)) in an indirect way by fitting ready-made hard models of each pure component spectrum. This explains the name of the method.

The clear attribution of all peaks to the individual components encourages a multivariate analysis of the pure component weights $\boldsymbol{w}$, which are influenced by the full wave number range, rather than a univariate analysis of certain characteristic peaks. The weights are estimated by fitting the spectral mixture model $\mathcal{X}(\nu,\boldsymbol{w},\boldsymbol{\theta})$ to the mixture spectrum $\mathbf{x}\epsilon\mathbb{R}^{1\times n_\nu}$ using a nonlinear least-squares approach:

$$\min_{\boldsymbol{w},\boldsymbol{\theta}} \quad \sum_{i=1}^{n_\nu}[x_i-\mathcal{X}(\nu_i,\boldsymbol{w},\boldsymbol{\theta})]^2. \tag{3}$$

During optimization, nonlinear effects such as peak variations or spectral shifts are taken into account by allowing for a constrained variation of the model parameters $\boldsymbol{\theta}_{P,k}$ and $\theta_{S,k}$, $k=1,\ldots,K$. To minimize the number of flexible parameters in this correction step, the relevant parameters are identified automatically by an algorithmic procedure as part of the solution of Eq. (3) [21]. The modeling of nonlinear effects is an important correction step for the pure component weights in many practical situations. A typical example is spectrometer replacement. Due to slight differences in the optical units, often spectral shifts are induced. If unshifted pure component spectra are fitted to the data, the pure component weights are incorrect as the spectra do not fit properly. Consequently a wrong mixture composition is predicted. If however the spectra are shifted in the correct manner by adapting the nonlinear shift parameters $\theta_{S,k}$ with $k=1,\ldots, K$ the pure component weight is corrected.

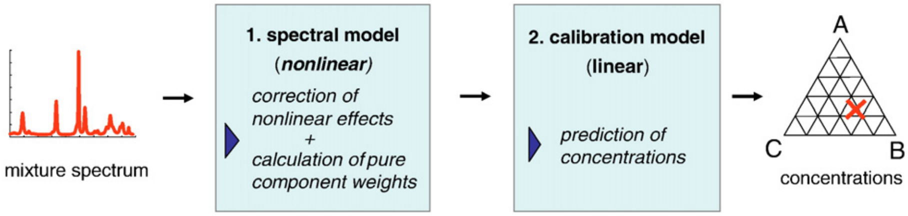


Fig. 1. Two-step procedure of indirect hard modeling (IHM).