A Semi-parametric Approach for Decomposition of Absorption Spectra in the Presence of Unknown Components

Payman Sadegh[1,2], Henrik Aalborg Nielsen[1],
and Henrik Madsen[1]

Abstract

Decomposition of absorption spectra using linear regression has been proposed for calculating concentrations of mixture compounds. The method is based on projecting the observed mixture spectrum onto the linear space generated by the reference spectra that correspond to the individual components comprising the mixture. The computed coefficients are then used as estimates for concentration of the components that comprise the mixture. Existence of unknown components in the mixture, however, introduces bias on the obtained concentration estimates. We extend the usual linear regression model to an additive semi-parametric model to take the unknown component into account, estimate the absorption profile of the unknown component, and obtain concentration estimates of the known compounds. A standard back-fitting method as well as a mean weighted least squares criterion are applied. The techniques are illustrated on simulated absorption spectra.

Keywords: Parameter estimation, non-parametric methods, unbiased estimates, chemometry, absorption spectra, additive models, semi-parametric models, mean weighted least squares, back-fitting.

[1] Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark

[2] E-mail: ps@imm.dtu.dk

1 Introduction

Chemometric spectroscopy is a simple way for examination of gases and liquids. UV examination of wastewater for instance has been proposed for quality control purposes (Thomas, Theraulaz & Suryani 1996). The technique is based on the analysis of the absorption spectrum obtained from the sample of interest. Depending on the concentrations of comprising compounds, the spectral absorption of the mixtures vary at different wavelengths. This information may in principle be used to *encode* the concentration of an existing compound, given information about the absorption pattern of the compound of interest. The functional dependency of absorption spectra upon concentrations and absorption spectra of comprising compounds is in general unknown. Several simple models have been proposed, most notably, a linear regression model (Gallot & Thomas 1993). In this model, it is assumed that the absorption spectrum of a mixture is a linear combination of absorption spectra of the comprising compounds where each coefficients determines the concentration of the corresponding compound. Hence, if spectral absorption measurements are performed at minimum p wavelengths, where p is the number of existing compounds, the concentrations may be estimated by the least squares technique (Gallot & Thomas 1993). The technique fails when unknown compounds are present. Even though, it is not of interest to estimate the concentration of the unknown components, the presence of such will introduce bias on the concentration estimates for the known ones unless the spectrum of the unknown component is orthogonal to the spectra of the known ones. Such a situation is unlikely to occur for most decomposition problems of interest in chemometry. We propose a semi-parametric model to account for the presence of unknown compounds. We apply both a standard back-fitting method for estimation in additive models and a novel technique based on a mean weighted least squares criterion (MWLS). MWLS provides a promising easy way to embed prior information into kernel estimation schemes. While kernel estimation of a function at N data points may be regarded as N independent weighted least squares problems, the MWLS combines the N optimization problems into one. The advantage is that any global information about the behavior of the process may be imposed as hard or soft constraints in the resulting single optimization criterion.

The rest of the paper is organized as follows. In Section 2, we present the semi-parametric formulation of spectral absorption decomposition problem, we review techniques from the theory of general additive models, and introduce the MWLS technique. In Section 3, we present a simple numerical study based on simulated data, and finally, Section 4 presents concluding remarks.

2 Problem formulation

Consider the problem of decomposing the observed function $f(t)$, $t \in T$, into known functions $f_i(t)$, $i = 1, ..., p$, by estimating the parameter $\theta = [\theta_1, \cdots, \theta_p]^\top$ of the linear regression

$$f(t) = \sum_{i=1}^{p} \theta_i f_i(t) + R(t) + e(t), \tag{1}$$

where $R(t)$ accounts for the superposition of all unknown components comprising $f(t)$, and $\{e(t)\}$ is a sequence of zero mean independently distributed random variables representing the measurement noise. Disregarding $R(t)$, the usual least squares estimate of θ is given by

$$\arg\min_{\theta} \sum_{t \in T} [f(t) - \sum_{i=1}^{p} \theta_i f_i(t)]^2. \tag{2}$$

where $T = \{t_1, \cdots, t_N\}$ is the set of N sampled observations. Inserting $f(t)$ from (1) in the solution to (2) shows that the bias on the least squares estimate is given by

$$(X^\top X)^{-1} X^\top \begin{bmatrix} R(t_1) \\ \vdots \\ R(t_N) \end{bmatrix}$$

where

$$X = \begin{bmatrix} f_1(t_1) & \dots & f_p(t_1) \\ f_1(t_2) & \dots & f_p(t_2) \\ \vdots & \vdots & \vdots \\ f_1(t_N) & \dots & f_p(t_N) \end{bmatrix}.$$

Hence depending on $R(t)$, the bias might be arbitrarily large.

For chemometric data, the function $f(t)$ is the measured absorption spectrum at various wavelengths t and $f_i(t)$ is the known absorption spectrum for component i. The presence of unknown components introduces the remainder term $R(t)$ which should be simultaneously with the coefficients θ_i estimated from data. The back-fitting algorithm (Hastie & Tibshirani 1990) can be applied under such circumstances. Back-fitting is an iterative method for decomposition of a number of unknown functions in an additive model. Starting from an initial guess, the algorithm iteratively estimates each one of the functions, fixing all others to their corresponding latest updated values. The algorithm has an explicit solution for problems of the type (1), see (Hastie & Tibshirani 1990), page 118. The solution involves a smoother function for estimation of $R(\cdot)$ and a weighted least squares criterion to estimate θ. Only in the case the smoother is a spline smoother, the back-fitting can be explicitly related to an optimization criterion (Hastie & Tibshirani 1990).

Another approach, which is novel to the best of our knowledge, is based on a mean weighted least squares criterion. The approach is particularly appealing since its solution is explicitly related to an optimization criterion which is a missing element for back-fitting using other smoothers than splines. The MWLS approach is as follows. Consider the following optimization

$$\min_{\theta,\{\phi(\tau)\}} \sum_{\tau\in T}\sum_{t\in T} w_h(|t-\tau|)[f(t)-\sum_{i=1}^{p}\theta_i f_i(t) - q\left(t-\tau;\phi(\tau)\right)]^2, \qquad (3)$$

where $q\left(t-\tau;\phi(\tau)\right)$ and $\phi(\tau)$ respectively denote a local approximation to $R(\cdot)$ around τ and its corresponding (τ-dependent) parameter and $\{w_h(|d|)\}$ is a weight sequence that falls monotonically with $|d|$. One typical choice for $q\left(t-\tau;\phi(\tau)\right)$ is a low order polynomial in $t-\tau$. The weight sequence is obtained from the kernel $K_h(|d|)$ according to

$$w_h(|d|) = \frac{K_h(|d|)}{\sum\limits_d K_h(|d|)}.$$

Some typical selections for the kernel $K_h(|d|)$ are Gaussian kernel

$$K_h(|d|) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{d^2}{2h^2}\right),$$

and Epanechnikov kernel

$$K_h(|d|) = \frac{3}{4h}\left(1-\frac{d^2}{h^2}\right) I(|d|\le h),$$

where $I(|d|\le h) = 1$ if $|d|\le h$ and zero otherwise.

The criterion (3) may be explained as follows. The optimization problem obtained by considering the inner sum in (3) as the cost function, i.e.

$$\min_{\theta,\phi(\tau)} \sum_{t\in T} w_h(|t-\tau|)[f(t) - \sum_{i=1}^{p} \theta_i f_i(t) - q\,(t-\tau;\phi(\tau))]^2, \tag{4}$$

provides the usual weighted least squares problem for non-parametric estimation of $R(\tau)$, $\tau \in T$, based on the local approximation $R(t) \approx q(t-\tau;\phi(\tau))$ around τ. Hence a non-parametric estimate for $R(\tau)$ is obtained by inserting the optimal value of $\phi(\tau)$ in $q(0;\phi(\tau))$. In connection with estimating θ, on the other hand, (4) is of no immediate use since the obtained estimates of θ vary with τ. This dependency is eliminated in (3) by the outer summation over τ. This may be thought of as restricting the solutions to the independent optimization problems (4) to yield a common estimate for θ.

Now assume that the local approximation $q(t-\tau;\phi(\tau))$ is linear in $\phi(\tau)$, i.e.

$$q\,(t-\tau;\phi(\tau)) = \sum_{i=1}^{m} \phi_i(\tau) g_i(t-\tau) \tag{5}$$

where $\phi(\tau) = [\phi_1(\tau), \cdots, \phi_m(\tau)]^\top$. Let W_τ denote a diagonal $N \times N$ matrix with the (i,i) element being equal to $w_h(|t_i - \tau|)$. Further denote

$$X_\tau = \begin{bmatrix} g_1(t_1-\tau) & \dots & g_m(t_1-\tau) \\ g_1(t_2-\tau) & \dots & g_m(t_2-\tau) \\ \vdots & \vdots & \vdots \\ g_1(t_N-\tau) & \dots & g_m(t_N-\tau) \end{bmatrix},$$

and finally $Y = [f(t_1), \cdots, f(t_N)]^\top$.

Proposition 1 Assume that the local approximation $q(t-\tau;\phi(\tau))$ is linear in $\phi(\tau)$ (see (5)). The optimal value of θ according to (3) is equivalent to the solution to the weighted least squares problem

$$\min_{\theta} (Y - X\theta)^\top W (Y - X\theta)$$

where

$$W = \sum_{\tau\in T} \left(W_\tau - W_\tau X_\tau (X_\tau^\top W_\tau X_\tau)^{-1} X_\tau^\top W_\tau \right)$$

Proof: Since $\phi(\tau)$ varies with τ in (3), $\phi(\tau)$ may be simply computed by finding the optimal value of $\phi(\tau)$ in (4) as a function of θ. Inserting the optimal values for $\phi(\tau)$ in (3) and collecting terms yields the desired result.

3 Numerical example

In this section, we apply the techniques discussed earlier to a spectral decomposition problem. Consider two absorption spectra $f_1(t)$ and $f_2(t)$ as given in Figure 1. These spectra contain COD, TOC, TSS, and BOD with concentrations 63, 0, 15, and 15 for f_1 and 36, 12.5, 0, 11.5 for f_2. The "unknown component" is assumed to consist of concentrations of nitrates with spectrum $R(t)$ as illustrated in Figure 2. The spectra of Figure 1 and Figure 2 are taken experimentally from real samples.

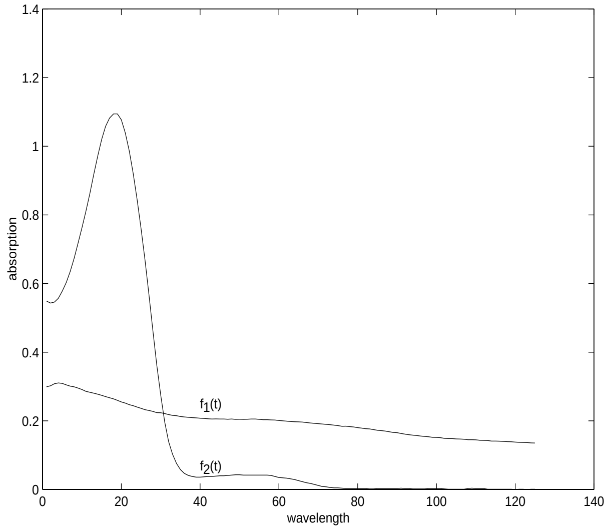

Figure 1: Reference absorption spectra.

We simulate the spectrum illustrated in Figure 3 by the linear combination:

$$f(t) = f_1(t) + f_2(t) + R(t).$$

The least squares solution yields coefficient estimates 1.35 and 1.81 for f_1

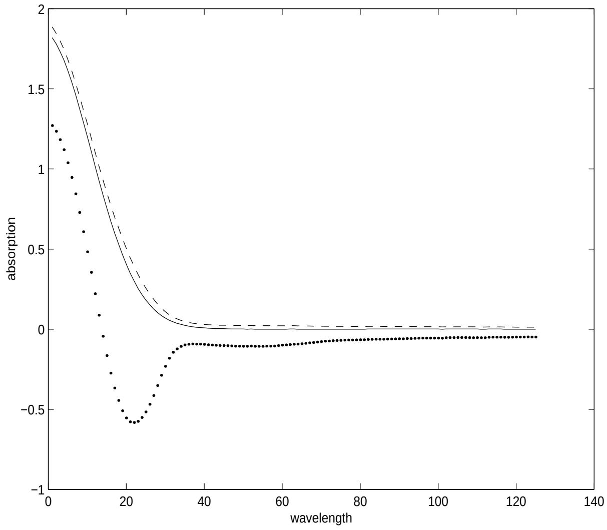

Figure 2: Spectrum for the unknown component. The solid curve, dashed, and dotted curves respectively represent the true spectrum, the estimated spectrum using a mean weighted least squares criterion, and the estimated spectrum using regular least squares.

and f_2, and an estimate for $R(t)$ as illustrated in Figure 2. These results clearly indicate the inappropriateness of the least squares solution.

We apply the result presented in Proposition 1 where the local approximators are polynomials of second order and the weights are computed according to a unit bandwidth Gaussian kernel. The coefficients of f_1 and f_2 are respectively estimated to 0.91 and 0.93.

We further apply the back-fitting solution to the above estimation problem. The best result is obtained by applying spline smoother with a large degree of freedom, yielding estimates of 1.25 and 0.73 for the coefficients of f_1 and f_2 respectively. These estimates are noticeably more biased than the MWLS solution.

Finally, we investigate the effect of measurement noise. We simulate 25 independent samples according to

$$f(t) = f_1(t) + f_2(t) + R(t) + e(t)$$

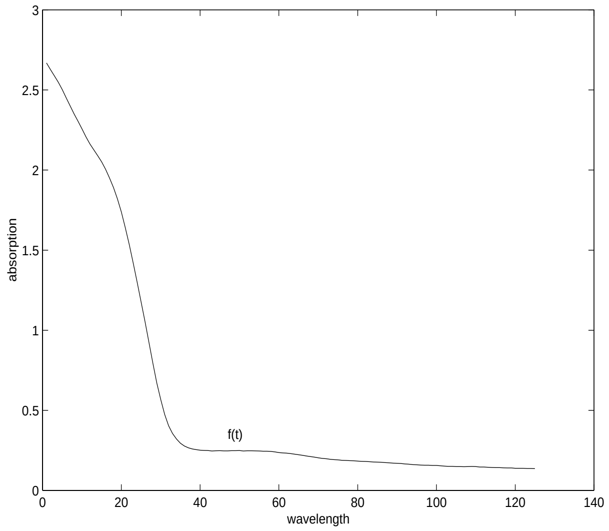

Figure 3: Simulated sample.

where $\{e(t)\}$ is a sequence of i.i.d. $N(0, 0.001^2)$ random variables, and empirically compute mean and covariance of the concentration estimates $\hat{\theta}$. These quantities are computed to

$$E\{\hat{\theta}\} = \begin{bmatrix} 0.9039 \\ 0.9306 \end{bmatrix}^\top, \; COV\{\hat{\theta}\} = \begin{bmatrix} 0.0576 & -0.0007 \\ -0.0007 & 0.0030 \end{bmatrix}$$

where $E\{\cdot\}$ and $COV\{\cdot\}$ as usual denote mean value and covariance. Numerical experimentation indicates that the estimation procedure fails for noise variances above 0.01^2.

4 Discussion and conclusion

2 We have proposed a solution to decomposition of absorption spectra in presence of correlated error (e.g. due to existence of unknown components). The underlying assumption throughout the paper is a linear additive model. We have applied both the back-fitting solution and a mean weighted least squares criterion. Numerical experience with back-fitting iterations for typ-

ical chemometric spectra fails due to high correlated among data. The back-fitting method yields reasonable estimates only if the explicit end solution of the iterations, which exists for a model linear in parameters, is applied. Contrary to back-fitting, the mean weighted least squares solution is not tied to an iterative algorithm but to a well defined optimization problem. The mean weighted least squares solution performs reasonably well for decomposition of absorption spectra in the presence of unknown components. The solution is rather sensitive to measurement noise which is again due to high correlation among reference spectra on the one hand and high correlation between the unmodelled error spectrum and the reference spectra on the other.

To further elaborate on the discussion above, Figure 4 on page 10 shows a scatter-plot matrix of the wavelength and five typical absorption spectra f_i, $i = 1, \cdots, 5$. The actual spectra are shown in the left column and, with the axes swapped, in the bottom row. From these it is seem that $f_2(t)$ and $f_3(t)$ are very similar, correspondingly the plot of $f_2(t)$ against $f_3(t)$ shows an almost linear relation. Consequently, concentrations of substances corresponding to these spectra will be difficult to distinguish, i.e. the estimated concentrations will be highly correlated.

For data simulated according to some arbitrary linear combination of the illustrated spectra and some typical spectrum for the unknown component $R(t)$ (simulated as a Guassian bell curve around some bandwidth), the R-squared value is above 0.9999 when omitting $R(t)$ from the model and replacing it with a intercept term. This indicate that the simulated spectrum $f(t)$ lies almost entirely in the space spanned by the reference spectra, making estimation of $R(t)$ difficult if $f(t)$ is measured with noise. Consequently, to reduce bias on the estimates of concentrations $R(t)$ must, to some extend which is determined by the level of noise, lie in an other space than the one spanned by the reference spectra.

The above considerations indicate that, if possible, reference spectra should be chosen so that (1) all explain different aspects of the unknown spectra (as opposed to $f_2(t)$ and $f_3(t)$ above), and (2) the unknown $R(t)$ lies, to some extend, in an other space than the one spanned by the reference spectra. Furthermore, measurement noise should be reduced as much as possible, e.g. by performing several measurements on the sample of interest and averaging.

Finally, the mean weighted least squares criterion introduced in this paper has application potentials far beyond the scope of the present paper. The approach provides a simple yet powerful tool to embed “global” information about the process of interest in local estimation techniques, hence combining the noise reduction and interpretability of global models with the generality, minimal model reliance, and convenience of non-parametric methods. Contrast this to usual ways of embedding prior information in non-parametric methods which concern local or smoothness properties such as selection of a suitable kernel, bandwidth, and degree of local approximators. This poses an interesting direction for future research and forthcoming publications.

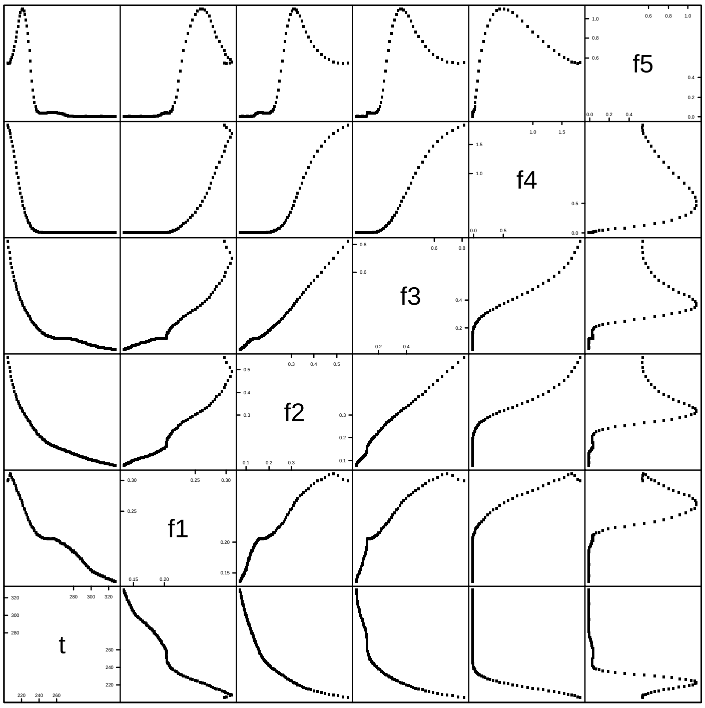

Figure 4: Scatter-plot matrix of wavelength (t) and reference spectra $(f_1(t), \ldots, f_5(t)$.

References

Gallot, S. & Thomas, O. (1993), 'Fast and easy interpretation of a set of absorption spectra: theory and qualitative applications for UV examination of waters and wastewatres', *Fresenius Journal of Analytical Chemistry* **346**, 976–983.

Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman & Hall.

Thomas, O., Theraulaz, F. & Suryani, S. (1996), 'Advanced UV examination of wastewaters', *Environmental Technology* **17**, 251–261.