Ultrasonic concentration measurement of aqueous solutions using PLS regression

Robert Schäfer a,*, Johan E. Carlson b, Peter Hauptmann a

a Institute of Micro- and Sensor Systems (IMOS), University of Magdeburg, P.O. Box 4120, 39016 Magdeburg, Germany
b EISLAB, Luleå University of Technology, SE-971 87 Luleå, Sweden

Available online 12 June 2006

Abstract

This work demonstrates the use of a multivariate statistical technique called partial least squares (PLS) to extract material related data by analyzing spectra of ultrasonic pulses. We show how PLS can be used to estimate the concentration of sodium chloride in an aqueous solution. The paper describes the use of PLS and discusses pre-processing of ultrasonic data, the PLS algorithm as well as model validation. The measured concentrations are compared to reference values. The influence of disturbances and parameter changes is highlighted. The proposed method is easily adaptable to similar applications and permits a cost-saving implementation using existing and approved hardware.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Partial least squares; Concentration measurement

1. Introduction

Modern industrial process control requires a high availability of accurate in situ measurement data. Additionally, rough conditions like extreme temperatures or aggressive substances induce a demand on robust measurement techniques. Ultrasonic solutions have become standard in this field of applications [1]. The focus is generally set to time-of-flight measurements as it is known from flowmeter or levelmeter devices, while amplitude information is often ignored. The amplitude information is related to mechanical properties of the propagation medium, which permits material characterization based on analyzing amplitude variations.

Material properties have become more and more important control variables for industrial processes. In addition to cost aspects one reason that limits the use of ultrasonic NDT/NDE application in process control might be the complexity of the systems, resulting in unmanageable or inaccurate physical models. One remedy to this problem is statistical modeling. This means finding a connection between some responses Y that are not directly measurable by studying some directly measurable descriptor variables, X, using Multiple Linear Regression (MLR), so that

\[ Y = X \cdot A. \] (1)

In this paper, we use partial least squares (PLS) to model the variations in spectra of ultrasonic pulses transmitted through aqueous solutions of sodium chloride (NaCl), for varying concentrations of NaCl. The PLS Regression (PLSR) delivers the coefficients A.

2. Theory

2.1. Physical background

Changing the concentration of a dissolved specimen in an aqueous solution causes changes in the mechanical properties of the fluid, in terms of density and bulk modulus. This affects the shape of an ultrasonic pulse transmitted through the fluid in several ways. Variations cause changes in acoustic impedance of the fluid \( Z_L \), while the acoustic impedance
of the transducer $Z_1$ remains unchanged. Reflections at the transducer/fluid interface are determined by a reflection coefficient $r = r(Z_1, Z_2)$ [2]. That is, $Z_1$ affects the amount of reflected as well as transmitted energy, which consequently has an influence to the wave shape and therefore can be measured at the electrical port of a piezoelectric transducer [3,4]. Usually, the reflection coefficient is frequency dependent. The same is true for the sound velocity [5].

Deriving a physical model to characterize the relationship of pulse distortion and fluid properties very soon becomes difficult or even impossible due to the complexity of the system. In this case, the use of statistical tools makes sense and enables rapid modeling of the phenomena of interest.

2.2. Specification of input and output, pre-processing

MLR requires pre-processing raw data [6]. This section describes how measured pulses, temperature values, and known concentrations are processed and organized in matrices. The response variable is the concentration of NaCl (by weight) \( \{c_{n}\}_1 \). The values are stored in the column vector $Y = \{c_n\} \in \mathbb{R}^{n \times 1}$.

The descriptor variables of the experiments, i.e., directly measured quantities, are the spectra of the ultrasound pulses sampled at the receiving transducer as well as the temperature of the medium. The sampled pulses $p(t)$ were transformed using the Discrete Fourier Transform (DFT), giving the spectral representation $P(f)$ in terms of magnitude and phase.

Once magnitude $P = \{|P_n|\}$ and (unwrapped) phase $\Phi = \{\text{arg}(P_n)\}$ were calculated, these were stored as rows of the descriptor matrix along with the measured temperature $\vartheta = \{\vartheta_n\}$.

$$
\mathbf{X} = \begin{bmatrix} P & \Phi & \vartheta \end{bmatrix}.
$$

Finally, the columns of $\mathbf{X}$ and $\mathbf{Y}$ were scaled to unit variance, and centered by subtracting their mean values. In this way, no column was (a priori) given any greater significance than the others. This is a standard procedure for regression problems.

2.3. Motivation to use PLSR

For an MLR model the relationship between descriptors and responses is given by Eq. (1). If $\mathbf{X}$ has full column rank, a least-squares estimate $\hat{\mathbf{A}}$ of $\mathbf{A}$ is obtained by

$$
\hat{\mathbf{A}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y},
$$

otherwise the inverse $(\mathbf{X}^T \mathbf{X})^{-1}$ will not exist. One remedy for this problem is called principal component regression (PCR). First a principal component analysis (PCA) is made on $\mathbf{X}$ [7]. The significant principal components are then used to form the Moore–Penrose pseudo-inverse of $\mathbf{X}$ [8]. This is a common approach, but there is an underlying assumption that is rarely fulfilled. Nothing says that large variation in $\mathbf{X}$ is necessarily of greater importance when it comes to describing $\mathbf{Y}$. That is, the discarded principal components might obtain valuable information about cross-covariance between $\mathbf{X}$ and $\mathbf{Y}$.

An alternative approach is PLSR, a method developed by Herman Wold [9], that has been applied to many areas of experimental sciences. We will give a brief description of the PLS method. An extensive overview can be found in [10]. The central idea of PLS, as opposed to PCR and ordinary least-squares (OLS) estimation, is that a set of basis functions (PLS components) for $\mathbf{X}$ and $\mathbf{Y}$ is determined in such a way that it best describes the cross-covariance of descriptors and responses.

2.4. Calculating the PLS components

The PLS components are calculated in an iterative process. For each iteration step $h = 1, \ldots, H$ the sample cross-covariance matrix

$$
\mathbf{C}^h_{XY} = \mathbf{Y}^T \mathbf{X}^h
$$

is calculated. Then the first principal component of $\mathbf{C}^h_{XY}$ must be found, i.e., the eigenvector of $(\mathbf{C}^h_{XY})^T \mathbf{C}^h_{XY}$ corresponding to the largest singular value of $\mathbf{C}^h_{XY}$ [7]. This vector is called weight vector $\mathbf{w}_h$. The corresponding score vector is

$$
t_h = \mathbf{X}_h \mathbf{w}_h.
$$

The so-called loading vectors of $\mathbf{X}$ and $\mathbf{Y}$ are then given by

$$
\mathbf{p}_h = (\mathbf{X}^h_1 t_h : (t_h^T t_h)^{-1})
$$

and

$$
\mathbf{q}_h = (\mathbf{Y}^h_1 t_h : (t_h^T t_h)^{-1}),
$$

respectively. The residuals $\mathbf{E}_h$ and $\mathbf{F}_h$ of $\mathbf{X}$ and $\mathbf{Y}$ are then

$$
\mathbf{E}_h = \mathbf{X}_h - t_h \mathbf{p}_h^T
$$

and

$$
\mathbf{F}_h = \mathbf{Y}_h - t_h \mathbf{q}_h^T.
$$

For iteration step $h + 1$, the new starting matrices are $\mathbf{X}_{h+1} = \mathbf{E}_h$ and $\mathbf{Y}_{h+1} = \mathbf{F}_h$. This procedure is repeated for $H$ PLS components to determine $\hat{\mathbf{A}}$. The scores, loadings and weights are stored in matrices as follows:

$$
\mathbf{T} = \begin{bmatrix} t_1 & t_2 & \cdots & t_H \end{bmatrix}
$$

$$
\mathbf{P} = \begin{bmatrix} p_1 & p_2 & \cdots & p_H \end{bmatrix}
$$

$$
\mathbf{Q} = \begin{bmatrix} q_1 & q_2 & \cdots & q_H \end{bmatrix}
$$

$$
\mathbf{W} = \begin{bmatrix} w_1 & w_2 & \cdots & w_H \end{bmatrix}.
$$

2.5. Predicting the responses

After the iterations are completed, the original matrices $\mathbf{X}$ and $\mathbf{Y}$ can be expressed in terms of PLS components according to

$$
\mathbf{X} = \mathbf{T} \mathbf{P}^T + \mathbf{E}_H
$$

$$
\mathbf{Y} = \mathbf{T} \mathbf{Q}^T + \mathbf{F}_H,
$$

where $\mathbf{E}_H$ and $\mathbf{F}_H$ are the residuals.
where $E_H$ and $F_H$ are the model residuals, i.e., the variations not described by the PLS components. Given an existing set of PLS components a prediction, $\hat{Y}$, of $Y$ is obtained by [10]:

$$\hat{Y} = X \hat{A}, \quad (16)$$

where $\hat{A} = XW(P^T W)^{-1} Q^T$, compare to Eqs. (1) and (3).

### 2.6. Determining the PLS model order

As in all linear modeling, finding an appropriate model order is an important and difficult problem. There are numerous criteria available for model order selection, e.g., the Akaike information criterion [11] and cross-validation [12]. Some of the available methods tend to overestimate the model order, while others are overly conservative [13].

For PLS modeling problems, cross-validation is the most commonly used method. In this paper we split $X$ and $Y$ to obtain a training data set $\{X_s, Y_s\}$ and a validation data set $\{X_k, Y_k\}$. To calculate $\hat{A}$, the training data set is used as input to the PLSR. The validation set is used for cross-validation to estimate an optimal model order, $H$, to ensure a minimum prediction error

$$H_{cv} = \arg \min_H \sum \frac{1}{n}(y_j - x_j \hat{A}_H)^2, \quad (17)$$

where $A = N - T$ is the number of subsets of $\{X_j, Y_j\}$.

### 3. Experimental setup and results

#### 3.1. Experiment description

This section describes a representative experiment, conducted to estimate the concentration of NaCl in an aqueous solution. $N = 50$ measured concentrations $c_n$ are covering a range of 0–5% (by weight).

A Tektronix AWG520 arbitrary waveform generator was used to form short sine pulses with a fundamental frequency of 1 MHz. These were converted into pressure pulses by an emitting transducer. The transmitted pulses were then recorded at a receiving transducer, opposite to the first one. A Data Acquisition Unit (Acquiris DC440, Compact PCI) sampled the incoming signals (14 bits, 100 MHz) as voltage signals $V(t)$ at the electrical port of the receiving transducer, and transferred them to a PC for data processing (see Fig. 1). For each recorded pulse, the temperature of the solution was also recorded. The length of each sampled sequence is $M = 2^{13}$. Fig. 2 shows a typical pulse.

Every $c_n$ was obtained by dissolving a known mass (using a Kern 870-13 precision balance) of NaCl in a known mass of distilled water. The reference density was measured using an Anton Paar DMA5000 densitometer.

Performing an FFT on $V(t)$, to obtain the spectral representation, $P(f) = \text{FFT}\{V(t)\}$, constitutes the first step of the pre-processing. Fig. 3 shows the magnitude spectrum of a received pulse from DC to 10 MHz. One of the steps in the processing is to calculate the unwrapped phase of the pulse spectra. This process is highly sensitive to noise, and therefore we were limited to a bandwidth where the signal-to-noise ratio (SNR) was sufficiently high. This bandwidth is indicated in the figure.

For calibration and validation the data sets are split. For building the statistical model as described in Section 2, every $2^{\text{nd}}$ row of $X$ was used. That means for $H_{max} = N/2 = 25$ PLS components. To achieve a reliable model and a minimum error respectively, cross-validation was used, i.e., the parameter $H$ was modified until Eq. (17) was fulfilled. A minimum error of $N^{-1}\sum |c_n| = 1.04\%$ was found for $H = 9$. Fig. 4 displays all $N$ concentrations obtained via PLSR against reference values. The error bars show the $2\sigma$ interval, corresponding to approximately 95% confidence level, where $\sigma$ is the sample standard deviation from the 25 estimated values of each concentration.

![Fig. 1. Experimental setup.](image1)

![Fig. 2. Typical shape of a recorded ultrasound pulse.](image2)

![Fig. 3. Magnitude spectrum of a sampled pulse. The relevant information is located within a frequency range of [0.5;1.5] MHz.](image3)
4. Discussion

For aqueous solutions of NaCl concentrations in the observed measurement range, variations of the wave shape are visible as attenuation changes as well as a varying sound velocity. The results presented in the previous section were obtained using only one pulse transmitted through the medium, and thus no knowledge about the energy transmitted by the transducer was available. This prevents estimation of the absolute attenuation of the medium, and only relative changes, due to changes in sodium chloride concentration were observed. Also, accurate measurement of the absolute sound velocity was difficult to obtain.

However, variations in the frequency dependent phase of the echoes give information about changes in speed of sound. Since both attenuation and speed of sound are highly temperature dependent, it is important to include the measured temperatures $T$ as a column of the descriptor matrix $X$, to avoid misinterpretations of changes in pulse shape originating from temperature dependent attenuation and sound velocity. If the temperature is available only in the calibration experiments, but not in the final application, the temperature vector can be regarded as a response variable, and thus included as a column of $Y$. We tried this, and the predictions of both concentrations and temperature still work, although at the cost of higher variance of the estimates.

The proposed technique can easily be extended to multicomponent mixtures and provides a tool for process control where the complexity of the system makes reliable physical modeling difficult. Even long-time affecting disturbances like sedimentation can be compensated by a simple re-calibration of the model. High measurement rates can be achieved, since a single vector multiplication is required to estimate the concentration from the descriptors.

5. Conclusion

In this paper we demonstrate how PLS regression can be used to accurately estimate the concentration of sodium chloride in an aqueous solution using spectra of ultrasound pulses. A reliable model could be found, using a cross-validation technique. The applied method affords a separation of several mechanisms affecting the wave shape. That is, a direct interrelation of concentration and characteristic wave shape could be found. An important requirement is, however, considering temperature as additional descriptor data.

References