Sensors and Actuators B 188 (2013) 440–453

Contents lists available at ScienceDirect

Sensors and Actuators B: Chemical

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

A bootstrapping-based statistical procedure for multivariate calibration of sensor arrays

Zongyu Geng [a], Feng Yang [a,*], Minqi Li [a], Nianqiang Wu [b]

[a] Industrial and Management Systems Engineering Department, West Virginia University, Morgantown, WV 26506, USA
[b] Department of Mechanical and Aerospace Engineering, West Virginia University, Morgantown, WV 26506, USA

ARTICLE INFO

Article history:
Received 22 April 2013
Received in revised form 9 June 2013
Accepted 11 June 2013
Available online 22 June 2013

Keywords:
Multivariate calibration
Sensor array
Optimal design of experiments
Bootstrapping
Statistical inference

ABSTRACT

One of the major challenges of calibrating a sensor array lies in the typically large samples required to estimate a high-quality multivariate calibration (MC) model, which functionally relates the array responses to the target analyte concentrations. For the efficient calibration of sensor arrays, this work develops a multi-stage procedure to guide the sampling in a sequential manner: Preliminary experiments are performed in the initial stage to collect some data; in each subsequent stage, information is derived from all the data collected from the previous stages and is employed to obtain the optimal design of the current-stage experiments. The design optimization at each stage seeks to optimize the quality of the MC model with a given sample size, and is performed based on the new statistical inference method, which quantifies the dependence of the MC model quality (the uncertainty/variability of the model estimates) upon the design of experiments. The proposed statistical inference takes advantages of both forward and inverse calibration modeling in the literature, is able to accommodate nonlinear sensor arrays, and utilizes the bootstrapping resampling method to handle the statistical inference issues that cannot be adequately addressed by existing methods. Substantial simulation studies have been performed to demonstrate the efficiency of the multi-stage procedure over the traditional once-and-for-all sampling.

1. Introduction

Sensor arrays are widely used to quantify analytes of interest in applications such as manufacturing control, environment monitoring, homeland security, biomedicine, and food industry. A sensor array consists of a number of sensors whose responses can provide a "fingerprint" for the target analytes in an environment. Individual sensors are subject to cross-sensitivity (i.e., responds to multiple components present in a background). Hence, a sensor array has to be coupled with its multivariate calibration (MC) model to quantify target analytes. The MC model is a mathematical function relating the analyte concentrations, denoted as vector **c**, to the array responses, denoted as vector **r**; and it is estimated from experimental data obtained by exposing the sensor array to a range of samples (e.g., a gas mixture) with pre-specified analyte concentrations. The ultimate goal of such a sensing system – both the device and the MC model – is to allow for the accurate quantification of target analytes in an unknown environment based on array responses. Apparently, the MC model plays a critical role in analyte quantification, and the quality of the model has a direct effect on the accuracy of the estimated analyte concentrations given by the sensing system.

However, efficiently calibrating a sensor array (or obtaining its MC model) remains one of the major difficulties in developing a reliable sensor array system [1]: How to achieve an MC model of desired quality using the least experimental effort? This is by nature a design of experiments (DOE) question, which has been largely ignored in the literature of array calibration. The vast majority of the existing work has been focused on the estimation of an MC model assuming that the experimental data is given (e.g., [2–4]). The few papers that have briefly mentioned DOE for array calibration include [5–8]. In those reports, the DOE methods that have been adopted or mentioned were restricted to classic designs such as fractional factorial designs [9, chapter 4, p. 135] and criteria-based (e.g., D-optimality, A-optimality, etc.) designs [10, chapter 11, p. 151]. However, these traditional designs have two major limitations. First, they are based on standard statistical inference for classic regression modeling that does not involve the forward-inverse complication present in array calibration (detailed later). Second, they are built for an MC model of fixed functional form, for which everything is pre-specified except the values of model parameters; at the stage of designing experiments for multivariate calibration, it is usually difficult (if not impossible) to have sufficient information to specify the MC model to that degree.

* Corresponding author. Tel.: +1 304 293 9477; fax: +1 304 293 4970.
E-mail addresses: zgeng@mix.wvu.edu (Z. Geng), feng.yang@mail.wvu.edu, fengyang08@gmail.com (F. Yang).

0925-4005/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.snb.2013.06.037

Nomenclature

$\mathbf{c}$	the variable vector representing the analyte concentrations of an environment
$\mathcal{C}$	the feasible region (or region of interest) of $\mathbf{c}$
c_p	the concentration of the pth component analyte
$\mathbf{c}^{(f)}$	the true concentration vector for the analytes in an environment
$\mathbf{c}^{(Grid)}$	the vector including all the grid points specified within $\mathcal{C}$
G	the number of grid points in $\mathcal{C}$
P	the number of target analytes
Q	the number of sensors in the array
$\mathbf{r}$	the random vector representing the sensor array responses
$\mathbf{r}^{(f)}$	an observed response vector of the sensor array after being exposed to an environment with analyte concentrations $\mathbf{c}^{(f)}$
δ_p	the desired standard error on the estimated concentration of the pth component analyte
$\boldsymbol{\epsilon}$	the random error vector for array responses
$\boldsymbol{\Gamma}$	vector of parameters in the forward MC model
$\boldsymbol{\Theta}$	vector of parameters in the inverse MC model

To overcome the shortcomings of the existing DOE methods, we developed a multi-stage procedure for sequential optimal design/sampling. The optimal design is performed based on the new bootstrapping statistical inference particularly developed to quantify the uncertainties of the estimated MC model for a sensor array.

2. Problem statement and method overview

For a sensor array, the MC model is estimated from a set of experimental data $\{(\mathbf{c}_i, \mathbf{r}_i)\,;\, i = 1, 2, \ldots, I\}$, where $\mathbf{c}_i$ represents the analyte concentrations in the ith sample and $\mathbf{r}_i$ the observed array responses. The quality of the MC model directly depends on the data set used for estimation. The task of DOE is to provide a set of design points $\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \ldots\}$ so that sampling at those points with limited experiment budget leads to an estimated MC model of the highest quality.

There are two major difficulties that traditional DOE methods are not able to handle in designing experiments for array calibration: the statistical inference issues in the MC model estimation, and the lack of information regarding the underlying MC model at the stage of designing experiments. We next discuss these two difficulties and describe how they are addressed in our method in Sections 2.1 and 2.2 respectively.

2.1. Statistical inference issues in array calibration

The objective of DOE herein is to optimize the quality of the MC model with respect to the design points. To solve such an optimization problem, the prerequisite is to quantify how the design points may affect the quality of the MC model, which is usually measured in terms of the variability (or uncertainty) of the estimated quantities of interest. These quantities may include the fitted model parameters and the analyte concentrations inferred by the MC model from observed array responses. How to quantify the relationship between the variability of model estimation and the design points? This is a statistical inference issue and is particularly interesting for multivariate calibration of sensor arrays, which involves both forward and inverse directions.

We consider $\mathbf{c} \to \mathbf{r}$ as the forward direction, obtaining $\mathbf{r}$ for a given $\mathbf{c}$; and the inverse direction refers to $\mathbf{r} \to \mathbf{c}$, estimating $\mathbf{c}$ for a given $\mathbf{r}$. In contrast to regular statistical modeling where only the forward direction is involved, array calibration involves both directions.

- Forward design: at the stage of calibrating a sensor array, the array is exposed to samples with known analyte concentrations (specified by the DOE strategy), and the corresponding array responses are observed. In the sample data, the pre-specified concentrations $\mathbf{c}$ can be considered as free of errors, since the mixture samples are prepared with extremely high accuracy; whereas the observed sensor responses $\mathbf{r}$ are subject to random errors due to instrument noise, environment interference, variation of sensing materials, etc.
- Inverse estimation: Once a sensor array has been calibrated, it is integrated with its MC model to infer the analyte concentrations in an unknown environment based on the sensor responses, which is represented as $\mathbf{r} \to \mathbf{c}$.

Given the existence of these two directions in array calibration, two types of MC models have been developed in the literature: forward and inverse models [2]. Both types of MC models intend to quantify the relationship between $\mathbf{r} = (r_1, r_2, \ldots, r_Q)$ and $\mathbf{c} = (c_1, c_2, \ldots, c_P)$. The vector $\mathbf{r}$ includes Q elements corresponding to the Q sensors in the array. The vector $\mathbf{c}$ involves P analytes, which include the target analytes as well as interferants that can cause a substantial change in array responses, that is, a systematic change that cannot be considered as random noise. Generally, it is required that $Q \geq P$ since otherwise $\mathbf{c}$ cannot be identified even from an error free $\mathbf{r}$ [2, 11, chapter 4].

The forward model is written as

$$\mathbf{r} = \mathbf{F}(\mathbf{c}, \boldsymbol{\Gamma}) + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\Gamma}$ is the vector of parameters and $\boldsymbol{\epsilon}$ the random error vector on the sensor responses. The advantage of using Model (1) as the MC model is that the standard statistical inference methods [12] can be applied on the model estimation, with $\mathbf{c}$ being deterministic and $\mathbf{r}$ random (as explained earlier). However, Model (1) has an obvious drawback: It does not allow for a direct calculation of $\mathbf{r} \to \mathbf{c}$. Hence, in operational use of the sensor array, additional inverse computation needs to be carried out based on $\mathbf{F}$. If $\mathbf{F}$ is a pure linear function of $\mathbf{c}$, the $\mathbf{F}$-based inverse computation can be relatively easily performed, and an estimate of $\mathbf{c}$ can be obtained for a given $\mathbf{r}$ [3]. However, when $\mathbf{F}$ involves some nonlinear terms of $\mathbf{c}$ (e.g., quadratic terms), the F-based inverse computation becomes much more complicated, and it is subject to issues such as the uniqueness of the estimated $\mathbf{c}$ for a given $\mathbf{r}$; in addition, such inverse computation will likely be time-consuming, which hinders the real-time monitoring ability of the sensing system. We believe that this is at least one of the reasons why the existing forward MC models all assume a linear dependence of $\mathbf{r}$ upon $\mathbf{c}$ [3], which unfortunately may well not hold in reality.

On the other hand, the inverse model

$$\mathbf{c} = \mathbf{G}(\mathbf{r}, \boldsymbol{\Theta}) \tag{2}$$

seeks to approximate $\mathbf{c}$ as a function of $\mathbf{r}$ with $\boldsymbol{\Theta}$ being the model parameters. Apparently, Model (2) can be used directly for real-time quantification of the analytes, and it can take practically any functional form adequate to describe the relationship between $\mathbf{c}$ and $\mathbf{r}$. A range of functional forms including kernel basis functions and neural network have been adopted for the inverse MC model [13–18]. But with the predictor variable $\mathbf{r}$ being random and the response $\mathbf{c}$ deterministic, standard statistical inference methods cannot be applied on the estimation of $\mathbf{G}$, and the inference (e.g.,