Chemometrics and Intelligent Laboratory Systems 176 (2018) 149–156

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Bayesian estimation of the analyte concentrations using the sensor responses and the design optimization of a sensor system

David Han [a,*], Kevin Johnson [b]

[a] Department of Management Science and Statistics, University of Texas, San Antonio, TX 78249, United States
[b] Chemistry Division, Code 6181, U.S. Naval Research Laboratory, Washington, DC 20375, United States

ARTICLE INFO

Keywords:
Bayesian analysis
Chemical sensors
Noninformative prior
Optimal design of sensory arrays
Sensitivity
Statistical inference

ABSTRACT

Using an array of sensors with well calibrated but different tuning curves, it is possible to appreciate a wide range of stimuli. In this work, we first revisit the statistical estimation of the stimuli concentrations given the responses of a sensor array, discussed in Sanchez-Montanes & Pearce [18]. Since it is not a typical regression problem, the Bayesian concept is adopted to develop an estimation method by elucidating the dynamic and uncertain nature of the environment-dependent stimuli with a proper choice of the probability distribution. Other studies confirm that the proposed method can demonstrate a superior performance in terms of accuracy and precision when compared to the popular frequentist methods in addition to the theoretical soundness it enjoys as a statistical estimation problem. Under the proposed framework, the design optimization of an artificial sensory system is also formulated using the expected Bayes risk as an objective function to minimize. The same approach may be equally applied to any sensory system in order to optimize its performance within a population of sensors. Finally, illustrative examples are provided to describe how the proposed method can be applied for the optimal configuration of a sensory system for a given sensing task.

1. Introduction

A chemical olfactory system is a sensor array consisting of hundreds of olfactory receptor neurons. Generally, these receptor neurons do not exhibit specificity to any single chemical compound but rather provide varying levels of response to multiple compounds. The precise mechanism for olfactory perception remains still unclear but the aggregate responses from a sensor array can provide the fundamental chemical information to detect tens of thousands of unique chemical vapors. By combining non-specific, general-purpose sensors with well calibrated but different tuning curves, it is possible to distinguish a wide range of stimuli or achieve a given sensing task without striving to develop fully selective or specific sensors to each chemical analyte [4]. Hence, sensor arrays are often proposed as potentially powerful and relatively inexpensive methods to characterize complex chemical mixtures.

Moreover, using information theoretic approaches, neural receptor systems have been explored in order to understand how these systems are structured, how the structure informs chemical recognition capability, and what the resulting implications are when designing a sensor-based chemical detection system [1,6]. Despite these efforts, the literature and application of non-specific sensor arrays for general-purpose chemical detection present general disappointment upon implementation. Even though such arrays have been reported frequently and have been the subject of numerous reviews over the past decades, relatively few instances of successfully commercialized devices exist to date. One reason is because the research community has mainly focused on the development of individual sensors or on the application for which the sensor array is to be used without considering the design and evaluation issues of sensor arrays. Consequently, design, optimization, and implementation of sensor arrays still remain time-intensive and costly, leading to sensory systems which tend to underperform when fielded, compared to their laboratory performance.

In this work, we revisit the statistical estimation problem of chemical stimuli given the responses of a sensor array, discussed in Ref. [18]. The concept of Bayesian analysis is adopted to develop an estimation method by explaining the dynamic and uncertain nature of the environment-dependent stimuli through a choice of the prior distribution. Under the proposed framework, the optimal configuration of an artificial sensory system is then discussed using the expected Bayes risk as a suitable objective function to characterize the performance of the sensory system. The proposed approach is generalizable and could be applied to other sensory systems for the stimuli estimation and/or the

* Corresponding author.
E-mail address: David.Han@utsa.edu (D. Han).

https://doi.org/10.1016/j.chemolab.2018.03.015
Received 20 January 2018; Received in revised form 28 March 2018; Accepted 29 March 2018
Available online 4 April 2018

optimal sensory system designs. The rest of the paper is organized as follows. Section 2 provides the model preliminaries and assumptions, and the formal model based on the Bayesian framework is developed in Section 3. The estimation of analyte concentrations is discussed in Section 4 while Section 5 addresses the optimal design of a sensory system using the expected Bayes risk. Two illustrative examples are provided in Section 6 to describe how the proposed method can be applied for the optimal configuration of a sensory system. Finally, Section 7 concludes the paper.

2. Preliminaries

Here we consider a system or an array of chemical vapor sensors in which each sensor is associated with certain parameters that control the response characteristics of the receptor to various chemical vapors. These parameters, denoted by the vector $\boldsymbol{\theta}$, define the tuning curve to the stimuli and need to be estimated to maximize the predictive power of the sensory system to different analytes. More specifically, let $\boldsymbol{Y}$ denote a vector of the (random) responses from a sensory system under consideration with a given set of the analyte concentrations, denoted by the vector $\boldsymbol{x}$, it is exposed to. Then, the following model describes the functional relationship between $\boldsymbol{Y}$ and $\boldsymbol{x}$.

$$\boldsymbol{Y} = h(\boldsymbol{\theta}, \boldsymbol{x}, \varepsilon), \tag{1}$$

where $h(\cdot)$ expresses an appropriate model or system to predict $\boldsymbol{Y}$ from $\boldsymbol{x}$ with the tuning provided by the parameters $\boldsymbol{\theta}$. The associated model errors are captured by the random vector ε whose dimension is equal to that of $\boldsymbol{Y}$. Let the dimensions of $\boldsymbol{Y}$ and $\boldsymbol{x}$ be m and n, respectively. In case one considers an additive error model, Eq. (1) becomes

$$\boldsymbol{Y} = h^*(\boldsymbol{\theta}^*, \boldsymbol{x}) + \varepsilon, \tag{2}$$

where $\boldsymbol{\theta}^*$ is a subset of $\boldsymbol{\theta}$ so that the parameters concerning the spread or dispersion of $\boldsymbol{Y}$ are associated with ε only. When ε is a white noise, $h^*(\boldsymbol{\theta}^*, \boldsymbol{x})$ in Eq. (2) describes the expected or mean responses of the sensory system given $\boldsymbol{x}$ (*viz.*, $E[\boldsymbol{Y}|\boldsymbol{x}] = h^*(\boldsymbol{\theta}^*, \boldsymbol{x})$). Since the sensory system may be composed of non-specific or general purpose sensors, the elements of $\boldsymbol{Y}$ may be correlated and their linear dependencies are dictated by the variance-covariance structure of ε, expressed as the $m \times m$ symmetric, positive definite matrix $\boldsymbol{\Sigma}$. Let us further assume that the probability distribution of ε follows a multivariate normal (or Gaussian) distribution with zero means, which is a reasonable and popular choice to describe scientific experimental errors. Then, $\langle \boldsymbol{Y}|\boldsymbol{x} \rangle$ naturally follows a multivariate normal distribution with the mean vector specified by $h^*(\boldsymbol{\theta}^*, \boldsymbol{x})$ and the variance-covariance matrix specified by $\boldsymbol{\Sigma}$.

As noted in Ref. [18], by using an array of the sensors, each with different tuning curves, one can implement a sensory system that can appreciate a wide range of stimuli with relatively few sensors (*i.e.*, $m \leq n$). In order to achieve this goal, the model parameters $\boldsymbol{\theta}^*$ and $\boldsymbol{\Sigma}$ need to be estimated with high accuracy (*viz.*, low bias) and high precision (*viz.*, low variance) based on an observed *i.i.d.* random sample of $(\boldsymbol{Y}, \boldsymbol{x})$. There are a number of well established statistical inferential tools available to chemometricians to estimate these parameters. These include but are not limited to the ordinary/weighted/generalized least squares (LS) methods, the method of moments (MOM), the (penalized) maximum likelihood methods (MLE), the EM algorithm, the Jackknife methods, the Monte Carlo simulation based methods, the bootstrap methods, the LASSO and ridge regression methods, the latent variable regression (LVR) methods, the sliced inverse regression (SIR) methods, the principal component regression (PCR) methods, the partial least squares (PLS) regression methods, the partial robust M-regression, the classification trees, and even the artificial neural networks (ANN), the support vector machines (SVM), and the ensemble approaches.

There are even several techniques to mitigate the problems associated with small sample sizes when they produce inadmissible estimates such as out-of-bounds solutions or cause non-convergence of some methods. These include the restricted maximum likelihood (REML) estimation with finite sample corrections [9], the Kenward-Roger standard error and degree of freedom corrections [12,13], and the Skene-Kenward corrections [19,20]. By virtue of these sophisticated statistical techniques, here we assume that these critical model parameters were well calibrated, passed rigorous goodness-of-fit tests, and cross-validated through a series of lab and field tests of the sensory system under consideration, rendering the estimation errors practically negligible. In other words, the sensor response functions to the library of target chemicals are known a priori. It is because the ultimate goal of this modeling process is not about predicting the sensor responses given the analyte types and concentrations but the exact opposite, which is about estimating the analyte types and concentrations given the responses of the sensor array (*i.e.*, $\langle \boldsymbol{X}|\boldsymbol{y} \rangle$). Depending upon the tuning curves of the individual sensor elements, the accuracy of the overall sensory system in estimating the stimulus will vary greatly in addition to the range of stimuli that may be appreciated.

3. Bayesian-based model

Needless to say, this is not a typical regression problem which usually aims to estimate $E[\boldsymbol{Y}|\boldsymbol{x}]$. In addition, since the stimulus population and their respective concentrations can vary greatly from environment to environment the sensor array is exposed to, the covariate vector $\boldsymbol{x}$ cannot be treated as static parameters to be estimated like in Ref. [18]. Rather, its dynamic and uncertain nature has to be elucidated in the model via a proper choice of its probabilistic distribution (*i.e.*, prior). For the efficient estimation of the unknown stimuli given an observed set of the sensory responses, here we adopt the concept of the random effects or the popular Bayesian framework by treating the unknown stimuli vector as a random vector $\boldsymbol{X}$. We also express the prior historical information, the experts' opinions or beliefs about the stimuli through a distribution function of choice, denoted by $f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\psi})$ with predetermined hyperparameters $\boldsymbol{\psi}$. Then, by the Bayes' theorem, the (posterior) probability distribution function of $\langle \boldsymbol{X}|\boldsymbol{y} \rangle$ is expressed as

$$f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\psi}) = \frac{f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}) f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\psi})}{\int_{\boldsymbol{x}} f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}) f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\psi}) d\boldsymbol{x}}, \tag{3}$$

where $f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta})$ is the joint distribution function of $\langle \boldsymbol{Y}|\boldsymbol{x} \rangle$ or the likelihood function of $\boldsymbol{\theta}$ while the denominator is called the marginal likelihood. Based on our previous assumption, $f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta})$ is the joint density function of the multivariate normal distribution with the mean vector specified by $h^*(\boldsymbol{\theta}^*, \boldsymbol{x})$ and the variance-covariance matrix by $\boldsymbol{\Sigma}$.

As shown in Eq. (3), the posterior distribution is a combination of the prior (determined by the researcher) and the likelihood (determined by the data). The contribution of these two quantities to the posterior is not equal though. With more sensory responses (*i.e.*, $m >> n$), the likelihood is given much more relative weight in calculating the posterior for the unknown stimuli [21]. This may seem trivial but it can have enormous implications when the number of sensors is small as the posterior distribution is highly reliant on how the prior was specified. When the information contained in the likelihood is relatively small due to a limited number of the sensory responses, the prior will play a key role in the posterior for estimating the stimuli and each additional piece of information will have a pronounced impact. Thus, it is necessary to include external information in the form of informative (or subjective) priors. For instance, one might need to consult application-specific experts, meta-analyses, or review studies in the area of interest to obtain informative, accurate priors that can meaningfully contribute to the posterior distribution of $\boldsymbol{X}$. Previous knowledge or information about the environment the sensor array is exposed to can also be incorporated in the prior. Specifying the priors of $\boldsymbol{X}$ based on expert opinions or previous studies can potentially improve the inferential performance since it allows to base results on more information than what is strictly provided in the sensory responses, which is especially helpful with small data sizes.