Journal of Computational Science 11 (2015) 102–111

Contents lists available at ScienceDirect

Journal of Computational Science

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

Proper orthogonal decomposition methods for the analysis of real-time data: Exploring peak clustering in a secondhand smoke exposure intervention

V. Berardi[a,b,*], R. Carretero-González[a], N.E. Klepeis[b], A. Palacios[a], J. Bellettiere[b], S. Hughes[b], S. Obayashi[b], M.F. Hovell[b]

[a] *Nonlinear Dynamical Systems Group, Computational Science Research Center, and Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182-7720, USA*

[b] *Center for Behavioral Epidemiology and Community Health, Graduate School of Public Health, San Diego State University, San Diego, CA 92182-7720, USA*

ARTICLE INFO

Article history:
Received 29 April 2015
Received in revised form 25 August 2015
Accepted 14 October 2015
Available online 19 October 2015

Keywords:
Proper orthogonal decomposition
k-means
Dylos monitor
Real-time measurement
Secondhand smoke

ABSTRACT

This work explores a method for classifying peaks appearing within a data-intensive time-series. We summarize a case study from a clinical trial aimed at reducing secondhand smoke exposure via the installation of air particle monitors in households. Proper orthogonal decomposition (POD) in conjunction with a *k*-means clustering algorithm assigns each data peak to one of two clusters. Aversive feedback from the monitors increased the proportion of short-duration, attenuated peaks from 38.8% to 96.6%. For each cluster, a distribution of parameters from a physics-based model of airborne particles is estimated. Peaks generated from these distributions are correctly identified by POD/clustering with >60% accuracy.

1. Introduction

Real-time and mobile technology for health delivery is becoming increasingly widespread and has the capacity to fundamentally alter the nature of the interaction between patients and health service providers. This technology offers the potential for personalized treatments that can be modified in real-time in response to several variables, namely participants' varying behaviors, environmental contexts, and unique past history [1]. Capitalizing on this opportunity is predicated on the accurate identification of these variables in a variety of dynamic contexts. Our ability to achieve this is limited by the availability of suitable technology to gauge behavior. In an effort to move towards this eventual future, this study explored the clustering of behavioral characteristics from intensive time-series data generated via a secondhand smoke exposure (SHSe) real-time technology intervention.

Project Fresh Air (PFA) is an ongoing randomized intervention trial aimed at reducing SHSe in the homes of smokers via the installation of Dylos DC1700 air particle quality monitors. Each study household contains a child as well as an adult who engages in SHS-generating behavior, typically indoor cigarette smoking. As described in Ref. [2], the monitors are calibrated to detect particles with sizes ranging from 0.5 to 2.5 μm, which is consistent with SHS as well as non-tobacco aerosol sources such as cooking and incense. One monitor is installed in the main smoking room and another is placed in the child's bedroom; measurements from only the main room monitor are included in the ensuing discussion. Every ten seconds, the monitor collects a measurement of the air particle concentration, which is an average of the previous 10 measurements collected at one-second intervals. This data is transmitted to a small computer that, in turn, uploads the data to a website that is accessible to PFA staff in near real-time. The monitors are fit with devices that deliver aversive visual and auditory feedback (yellow/red lights and beeps) that are programmed to engage when air particle concentrations exceed 60 μg/m^3; the aversiveness of the feedback increases [3] if the 120 μg/m^3 threshold is breached. For each home, the duration of the trial is broken into two phases: (1.) *Baseline (BL)* – a washout period during which feedback is not activated, designed to allow for the abatement of participant reactivity to monitor installation and (2.) *Treatment (TX)* – the period during which the feedback is activated.

[*] Corresponding author at: Computational Science Research Center, San Diego State University, San Diego, CA 92182-7720, USA.
E-mail address: vberardi@cbeachsdsu.org (V. Berardi).

http://dx.doi.org/10.1016/j.jocs.2015.10.006
1877-7503/© 2015 Elsevier B.V. All rights reserved.

To reduce SHSe, the PFA intervention aims to modify particle-generating behavior, in particular tobacco smoking. The intervals of the particle time-series data with elevated concentrations, or peaks, serve as proxy measures of this behavior. As such, we seek to abstract behavioral features from peaks in the time-series data. Complicating this task is the lack of information about the identification and number of household occupants associated with a given peak. Additionally, the monitors only detect information about particle size and not chemical composition so confounding sources of smoke particles, such as burning food, are likely present. Ultimately, we aim to associate different peaks with distinct behaviors such as cigarette smoking, food burning, or air venting and to analyze the patterns of these behaviors over time. The approach outlined hereafter represents the establishment of the groundwork on which to accomplish this task.

In Section 2, proper orthogonal decomposition (POD), a blind signal separation (BSS) technique that can be used to identify underlying source signals that are functionally associated with peak characteristics, is described. Section 3 discusses the application of the methodology in Section 2 to a case study from PFA. A cluster analysis of POD coefficients that allows characteristically-similar peaks to be classified together is set forth in Section 4 and the results of this analysis are summarized in Section 5. Section 6 describes the relationship between peak clusters and parameters from a physics-based model of airborne particulates, which enables a physical interpretation of the POD/clustering results. A discussion of findings ensues in Section 7.

2. Extension of proper orthogonal decomposition to peak analysis

BSS is defined as the factoring of a mixed source into previously-unknown, independent components [4]. It has been implemented in a variety of contexts including the analysis of interstellar dust [5], neuroprocessing [6], and audio processing [7]. A popular BSS technique is proper orthogonal decomposition (POD) also known as Karhunen–Loève decomposition [8], principal components analysis [9], singular systems analysis [10], or singular value decomposition [11]. This procedure transforms a set of observations to a new coordinate system in which each dimension is linearly uncorrelated with the others. It is an attractive option to discriminate between peak characteristics since it provides an optimal basis to decompose signals and analytical bounds for the estimate of total "energy" captured by the decomposition [8]. For this study, POD is used to define a projection (decomposition) into a lower dimensional space where different types of peaks that represent similar physical scenarios that triggered elevated particle counts can be identified via clustering analysis.

Consider a sequence of observations represented by scalar functions $u(\mathbf{x}, t_i)$, $i = 1 \ldots M$. Typically t_i represents the ith temporal observation of state variable $\mathbf{x}$. Without loss of generality, the time-average of the sequence, defined by

$$\overline{u}(\mathbf{x}) = \langle u(\mathbf{x}, t_i)\rangle = \frac{1}{M}\sum_{i=1}^{M} u(\mathbf{x}, t_i), \qquad (1)$$

is assumed to be zero (if not, as it is in our case, simply subtract the time-average from all observations). The POD extracts time-independent orthonormal basis functions, $\phi_k(\mathbf{x})$, and time-dependent orthonormal amplitude coefficients, $a_k(t_i)$, such that the reconstruction

$$u(\mathbf{x}, t_i) = \sum_{k=1}^{M} a_k(t_i)\,\phi_k(\mathbf{x}), \quad i = 1, \ldots, M \qquad (2)$$

is *optimal* in the sense that the average least squares truncation error of the POD reconstruction $\varepsilon_m = \langle |u(\mathbf{x}, t_i) - \sum_{k=1}^{m} a_k(t_i)\,\phi_k(\mathbf{x})|^2\rangle$ is minimized for any given number $m \le M$ of basis functions over all possible sets of orthogonal functions. $\langle\cdot\rangle$ denotes an average operation, usually over time; and the functions $\phi_k(\mathbf{x})$ are called *empirical eigenfunctions*, *coherent structures*, or *POD modes*.

The domains $\mathbf{x}$ and t are completely empirical so that there is flexibility to interpreting them according to the characteristics of the data. Often times, POD analysis is performed on a state variable $\mathbf{x}$ assessed at various times t_i [12]. When extended to time-series data, the interpretation can change to i instances of a univariate time-series $\mathbf{x}$, e.g., stock returns for multiple companies over a specified interval [13]. The procedure can also be performed on multivariate time-series [14]. Yet another interpretation is *singular spectrum analysis*, where a univariate time-series is embedded to create a multidimensional state variable $\mathbf{x}$, that is observed at time steps t_i [15]. In our case, we are interested in peak events, i.e., the intervals in the time-series with elevated particle measurements. We assign $u(\mathbf{x}, t_i)$ to the indoor particle concentration measurements of the ith peak. Rather than representing a state variable assessed at some time t_i, $\mathbf{x}$ is a subset of the data corresponding to the ith peak. Thus the collection of peaks can be summarized as the matrix $U = [u(\mathbf{x}, t_1)|u(\mathbf{x}, t_2)| \ldots |u(\mathbf{x}, t_M)]$ where the ith column corresponds to the data from the ith peak event, although the order of the peaks does not affect the analysis.

It can be shown that the eigenfunctions ϕ_k in Eq. (2) are the eigenvectors of the matrix product $(1/M)UU^T$. A popular technique for finding these eigenvectors when the resolution of $\mathbf{x}$ is greater than the number of observations is the *method of snapshots* developed by Sirovich [16]. First the eigenvectors of $(1/M)U^TU$, denoted as $\mathbf{v}_k$, are found. Then the ϕ_k's are calculated by $\Phi = UV$ where $\Phi = [\boldsymbol{\phi}_1|\boldsymbol{\phi}_2| \ldots \boldsymbol{\phi}_M]$ and $V = [\mathbf{v}_1|\mathbf{v}_2| \ldots \mathbf{v}_M]$. Let $\mathbf{a}_i$ represent the reconstruction coefficients associated with the ith peak. These can be calculated by $A = U^T\Phi$, where A is the M-by-M matrix $[\mathbf{a}_1|\mathbf{a}_2| \ldots \mathbf{a}_M]$. Statistically speaking, the eigenvalues λ_k of $(1/M)U^TU$ represent the variance of the data set in the direction of the corresponding POD mode $\phi_k(\mathbf{x})$. In physical terms, if u represents a component of a velocity field, then λ_k measures the amount of kinetic energy captured by the respective POD mode, $\phi_k(\mathbf{x})$. In this sense, the energy measures the contribution of each mode to the overall dynamics. Thus, the total energy captured in the POD is defined as the sum of all eigenvalues: $E = \sum_{k=1}^{M} \lambda_k$, and the relative energy captured by the kth mode is $E_k = \lambda_k/E$.

3. POD of particle concentration time-series

To demonstrate the application of POD to particle concentration data, we considered a single household from PFA, HM180. This home is a single-story, 1 bedroom, 1 bathroom detached house. The monitor was placed at a height of 8 feet in the living room of the home. The household was enrolled in the study for 95 days, with the first 31 days in the *BL* phase and the remainder in the *TX* phase. Approximately 750,000 measurements were collected from the monitor in the main smoking room. As will be discussed in detail in Section 5.2, HM180 was chosen based on its reporting of tobacco smoking events to PFA staff.

When recorded by the Dylos monitor, each particle concentration measurement is assigned an alarm status variable that controls the emission of visual and auditory feedback. We use this variable to define peak events. An event begins when the alarm status indicates an initial breach of 60 μg/m^3; this triggers a yellow light and the first sound. The peak event does not end until the alarm status indicates that the concentration has fallen below 40 μg/m^3 which corresponds with the cessation of all visual and auditory feedback.