Thai Journal of Mathematics : 21–33
Special Issue: Annual Meeting in Mathematics 2017

http://thaijmath.in.cmu.ac.th
ISSN 1686-0209

An Application of Proper Orthogonal Decomposition for Estimating Missing Data of Patients in Different Cause Groups

Norapon Sukuntee and Saifon Chaturantabut[1]

Department of Mathematics and Statistics, Faculty of Science and Technology, Thammsat University, Pathumthani 12120, Thailand
e-mail : zeroryz@hotmail.com (N. Sukuntee)
saifon@mathstat.sci.tu.ac.th (S. Chaturantabut)

Abstract : This work applies the notion of proper orthogonal decomposition (POD) to approximate missing data that represents the amount of in-patients from different cause groups, such as influenza, malaria, HIV and alcoholic liver diseases, in Saraburi provice, Thailand. POD is used to construct a low-dimensional basis that extracts dominant trends of data from existing set of complete samples. The approximation of each missing data is obtained through an extension of POD called gappy POD (GPOD), which employs available data efficiently and optimally in the least-squares sense. Due to the large variation among the numbers of in-patients in different categories, this work applies a normalization using mean and standard deviation to pre-process the data and introduces a possible range of each approximated missing value. The numerical tests demonstrate the accuracy of the estimation from GPOD using different numbers of complete samples. We also investigate the change in accuracy when the number of missing data in each incomplete sample increases.

Keywords : data reconstruction; proper orthogonal decomposition; least-squares method.
2010 Mathematics Subject Classification : 65F15; 65F30.

[1] Corresponding author.

1 Introduction

The information related to patients in each country is an important indicator that can be used to improve health and medical care services in the future. However, in Thailand, some of these previous medical informations are not available consistently every year due to, for example, lack of linkage between hospitals' systems and restricted access to data. As a results, reconstructing unavailable data efficiently and accurately can be an essential part to recover a trend of the country's health situation. In this work, a supervised-learning technique is employed to estimate missing information by testing on the in-patient data from only one of the central provinces in Thailand to avoid the inconsistent problem in data due to, e.g. incomplete reports from local offices. In particular, this work focuses on the number of in-patients arranged by 70 cause groups, according to health service units, Ministry of Public Heath, in Saraburi province, Thailand.

The concept of proper orthogonal decomposition (POD) is introduced in 1937 by Lumley in the context of inhomogeneous structure turbulent flows [1] and stochastic tools in turbulence [2]. POD is also known as, for example, Karhunen-Love decomposition (KLD), principal component analysis (PCA), or singular value decomposition(SVD). POD has been used in many applications, e.g. [3–6]. It can be considered as a supervised learning method, since it can provide an approximation from the basis that extracts the dominant characteristic of the existing data.

An important class of POD applications is based on repairing damaged data and constructing missing or "gappy" data as proposed by Everson and Sirovich [7] in the context of face recognition. The approach using POD for the purpose of data reconstruction is therefore often called **gappy POD** (GPOD). In the application of aerodynamic flow fields, GPOD was formally introduced in [8] and it was later used to calibrate and illustrate air flow past a wing [9]. GPOD has been recently used in many other engineering applications. In chemical engineering, GPOD was applied on the reconstruction of flame kinematic in spark ignition engine [10] and combustion of natural gases [11]. In mechanical engineering, GPOD has been applied to fluid mechanics [12] and it was used to estimate of the spatial distribution of the unknown material properties [13]. In image processing, it can be used to derive a low-dimensional model [14]. In [15], GPOD was used to optimize the operation of wells in water flooding reservoir.

In this work, we apply GPOD to approximate data for the number of in-patients arranged by many cause groups, such as influenza, malaria, HIV, alcoholic liver diseases, thalassaemia, diabetes mellitus, and motorcycle rider injured transport accidents. This type of data is different from the ones used in the previous works (e.g flow or image data) in the sense that there is no smooth continuity between the nearby data samples or their features in adjacent components, which makes the estimation task more difficult and challenging. Therefore, in this paper, besides introducing some basic concepts of POD and GPOD in Section 2, we introduce a simple normalization using mean and standard deviation of the data to avoid the effect of large variation in the components (features) of data,

which are corresponding to the number of in-patients in various categories each year. A general procedure for constructing POD basis that extracts the dominant trend of data is given in Algorithm 1. A modified gappy POD procedure that includes the approximation based on the normalization and a possible range of each approximated missing value is described in Section 2.2 and summarized in Algorithm 2. Numerical experiments are shown in Section 3. In Section 3.1, the accuracy of reconstructing the known data using POD are investigated with normalized and non-normalized data with mean and standard deviation. Section 3.2 demonstrates the accuracy of the estimation from GPOD using different amount of complete samples. It also compares accuracy of the approximations when the number of missing data in each incomplete sample changes. Finally, some concluded remarks and future extension are discussed in Section 4.

2 Methodology

This section provides the fundamental concept of proper orthogonal decomposition (POD) and the mathematical explanation for its extension, called gappy POD method, with some modifications to make it suitable for the data.

2.1 Proper Orthogonal Decomposition (POD)

The aim of POD is to construct a set of basis by extracting features that describe the main characteristics from the system of interest. Let $\{\mathbf{y}_j\}_{j=1}^{n_s} \subset \mathbb{R}^n$ be the set of snapshots with mean zero for each component. I.e. $\bar{y}_i = \frac{1}{n_s}\sum_{j=1}^{n_s} y_{ij} = 0$, for all $i = 1, \ldots, n$. The POD basis $\{\mathbf{v}_i\}_{i=1}^{k}$ can be viewed as an orthonormal basis that minimizes the approximation error in $2-$norm for a given fixed basis rank. Note that the approximation for each snapshot $\mathbf{y}_j$ using projection on an orthogonal basis $\{\mathbf{v}_i\}_{i=1}^{k}$ is given by

$$\mathbf{y}_j \approx \sum_{i=1}^{k} \mathbf{v}_i(\mathbf{v}_i^T\mathbf{y}_j) = \mathbf{V}\mathbf{V}^T\mathbf{y}_j,$$

where $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_k] \in \mathbb{R}^{n\times k}$. When POD is used to generate the basis $\{\mathbf{v}_i\}_{i=1}^{k}$, the resulting basis solves the minimization problem

$$\min_{\{\phi_i\}_{i=1}^{k}} \sum_{j=1}^{n_s} \|\mathbf{y}_j - \sum_{i=1}^{k} \phi_i(\phi_i^T\mathbf{y}_j)\|_2^2, \qquad \phi_i^T\phi_j = \delta_{ij},$$

where $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ of $i = j$. POD basis can be computed by the singular value decomposition (SVD) of solutions or snapshots:$\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{n_s}] \in \mathbb{R}^{n\times n_s}$. The SVD of a rectangular matrix $\mathbf{Y} \in \mathbb{R}^{n\times n_s}$ is given by $\mathbf{Y} = \hat{\mathbf{V}}\Sigma\mathbf{Z}^T$, where r is the rank of $\mathbf{Y}$, $\hat{\mathbf{V}} = [\mathbf{v}_1, \ldots, \mathbf{v}_r] \in \mathbb{R}^{n\times r}$ and $\mathbf{Z} \in \mathbb{R}^{n_s\times r}$ are orthogonal matrices and $\Sigma = diag(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$ with singular values in decreasing order: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. Then the POD basis of rank $k < r$ consists of the first k

columns of $\hat{\mathbf{V}}$. It can be shown [16] that the minimum error of the approximation by using POD basis is given by

$$\sum_{j=1}^{n_s} \|\mathbf{y}_j - \mathbf{V}\mathbf{V}^T\mathbf{y}_j\|_2^2 = \sum_{\ell=k+1}^{r} \sigma_\ell^2, \tag{2.1}$$

which is the sum of the neglected singular values $\sigma_{k+1}, \ldots, \sigma_r$. Note that, besides using SVD, the POD basis can be computed by using the method of snapshots based on eigenvalue decomposition of correlation matrix of the snapshots [7]. The procedure for computing POD basis is shown in Algorithm 1.

Algorithm 1 Algorithm for constructing POD basis

Input: Snapshots $\{\mathbf{y}_j\}_{j=1}^{n_s} \subset \mathbb{R}^n$
Output: POD basis $\mathbf{V}_k$.

1: Create snapshot matrix : $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{n_s}] \in \mathbb{R}^{n \times n_s}$and let $r = rank(\mathbf{Y})$
2: Compute SVD: $\mathbf{Y} = \hat{\mathbf{V}}\Sigma\mathbf{Z}^T$ and choose dimension $k \leq r$
3: POD basis of rank k : $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_k] = \hat{\mathbf{V}}(:, 1:k)$

2.2 Gappy POD

Gappy POD can be used to approximate or reconstruct missing data from the available partial data, that obtained, e.g. from experimental measurements or numerical simulations.

Let $\mathcal{W} := \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{n_s}\} \subset \mathbb{R}^n$ be a set of complete data. I.e. all n components of $\mathbf{w}_j$ are knows, for $j = 1, \ldots, n_s$. Each vector $\mathbf{w}_j = [w_{1j}, \ldots, w_{nj}]^T$, $j = 1, \ldots, n_s$ is generally called a **snapshot** or a **sample** and each component i of every vector in $\mathcal{W}$ consists of the same *feature*. The mean and the unbiased sample variance of each feature i for all snapshots in $\mathcal{W}$ is defined by, respectively,

$$\bar{w}_i = \frac{1}{n_s}\sum_{j=1}^{n_s} w_{ij}, \qquad \text{and} \qquad s_i^2 = \frac{1}{n_s - 1}\sum_{j=1}^{n_s}(w_{ij} - \bar{w}_i), \qquad i = 1, \ldots, n, \tag{2.2}$$

and the corresponding standard deviation is $s_i = \sqrt{s_i^2}$. In general, each snapshot could have extremely different quantities in its components due to the variation among the features as shown in Section 3.1. To avoid this variation effect, each i-th component w_{ij} of snapshot $\mathbf{w}_j$, $j = 1, \ldots, n_s$, will be scaled or normalized as

$$y_{ij} \quad = \quad \frac{w_{ij} - \bar{w}_i}{s_i}, \qquad i = 1, \ldots, n \tag{2.3}$$

to obtain $\mathbf{y}_j = [y_{1j}, \ldots, y_{nj}]^T \in \mathbb{R}^n$. Let $\bar{\mathbf{w}} = [\bar{w}_1, \ldots, \bar{w}_n]^T$ be the mean vector of all features $1, ..., n$. Then $\bar{\mathbf{w}} = \frac{1}{n_s}\sum_{j=1}^{n_s} \mathbf{w}_j \in \mathbb{R}^n$.
Let $\mathbf{D} = diag(1/s_1, 1/s_2, \ldots, 1/s_n) \in \mathbb{R}^{n \times n}$ be the diagonal matrix whose diagonal

entries consist of reciprocal of all standard deviations. Then we can write the *normalized* snapshot $\mathbf{y}_j$ as

$$\mathbf{y}_j = \mathbf{D}(\mathbf{w}_j - \bar{\mathbf{w}}), \qquad j = 1, 2, ..., n_s. \tag{2.4}$$

Define $\mathcal{Y} := \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_s}\} \subset \mathbb{R}^n$. Let $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_s}] \in \mathbb{R}^{n \times n_s}$. Let $\mathbf{V}$ be the POD basis of rank k constructed as described in Algorithm 1 of the previous section.

Suppose $\widehat{\mathbf{w}} \in \mathbb{R}^n$ is a vector of **incomplete** snapshot, $\widehat{\mathbf{w}} \notin \mathcal{W}$. That is, there are some components in $\widehat{\mathbf{w}}$ that are unknown. Define $\widehat{\mathbf{y}} := \mathbf{D}(\widehat{\mathbf{w}} - \bar{\mathbf{w}})$. Then the unknown components in $\widehat{\mathbf{w}}$ and $\widehat{\mathbf{y}}$ are in the same locations, i.e. the same indices. In particular, suppose there are $n_\wp$ known components and $n_m = n - n_\wp$ unknown components. Let $\mathcal{P} := \{\wp_1, \wp_2, \ldots, \wp_{n_\wp}\} \subset \{1, 2, \ldots, n\}$ be the index set of the *known* components in $\widehat{\mathbf{y}}$. Define $\vec{\wp} = [\wp_1, \wp_2, \ldots, \wp_{n_\wp}] \in \mathbf{R}^{n_\wp}$ and $\mathbf{P} = [\mathbf{e}_{\wp_1}, \ldots, \mathbf{e}_{\wp_{n_\wp}}] \in \mathbb{R}^{n \times n_\wp}$, where $\mathbf{e}_{\wp_i} = [0, \ldots, 0, 1, 0, \ldots, 0]^T \in \mathbb{R}^n$ is the $\wp_i$-th column of the identity matrix $\mathbf{I}_n \in \mathbb{R}^{n \times n}$, for $i = 1, \ldots, n_\wp$. Note that, pre-multiplying $\mathbf{P}^T$ is equivalent to extracting the $n_\wp$ rows corresponding to the indices $\wp_1, \ldots, \wp_{n_\wp}$,. Similarly, let $\mathcal{M} = \{m_1, m_2, \ldots, m_{n_m}\} \subset \{1, 2, \ldots, n\}$ be the index set of the *unknown* components in $\widehat{\mathbf{y}}$ and define $\vec{\mathbf{m}} = [m_1, m_2, \ldots, m_{n_\wp}] \in \mathbf{R}^{n_m}$, $\mathbf{M} = [\mathbf{e}_{m_1}, \ldots, \mathbf{e}_{m_{n_m}}] \in \mathbb{R}^{n \times n_m}$. I.e. the known components and the unknown components are given in the following two vectors, respectively:

$$\widehat{\mathbf{y}}_\wp := \mathbf{P}^T\widehat{\mathbf{y}} \;=\; [\widehat{y}_{\wp_1}, \ldots, \widehat{y}_{\wp_{n_\wp}}]^T \in \mathbb{R}^{n_\wp}, \;\; \widehat{\mathbf{y}}_m := \mathbf{M}^T\widehat{\mathbf{y}} \;=\; [\widehat{y}_{m_1}, \ldots, \widehat{y}_{m_{n_m}}]^T \in \mathbb{R}^{n_m}.$$

The goal is to approximate the components of $\widehat{\mathbf{y}}_m$. To do this, we first assume that

$$\widehat{\mathbf{y}} \;\approx\; \mathbf{V}\mathbf{a}$$

for some coefficient vector $\mathbf{a} \in \mathbb{R}^k$, which implies that

$$\begin{aligned} \mathbf{P}^T\widehat{\mathbf{y}} \;&\approx\; (\mathbf{P}^T\mathbf{V})\mathbf{a} \quad \text{or} \quad \widehat{\mathbf{y}}_\wp = \mathbf{V}_\wp\mathbf{a}, \qquad \text{where} \;\; \mathbf{V}_\wp := \mathbf{P}^T\mathbf{V} \in \mathbb{R}^{n_\wp \times k} \\ \text{and} \quad \mathbf{M}^T\widehat{\mathbf{y}} \;&\approx\; (\mathbf{M}^T\mathbf{V})\mathbf{a} \quad \text{or} \quad \widehat{\mathbf{y}}_m = \mathbf{V}_m\mathbf{a}, \;\; \text{where} \;\; \mathbf{V}_m := \mathbf{M}^T\mathbf{V} \in \mathbb{R}^{n_m \times k}. \end{aligned}$$

Since the only available data points are contained in $\widehat{\mathbf{y}}_\wp = \mathbf{P}^T\widehat{\mathbf{y}}$, we can find the vector $\mathbf{a}$ by focusing on the approximation $\widehat{\mathbf{y}}_\wp \approx \mathbf{V}_\wp\mathbf{a}$ and solving for $\mathbf{a}$ from the following least-squares problem:

$$\min_{\mathbf{a} \in \mathbb{R}^k} \|\widehat{\mathbf{y}}_\wp - \mathbf{V}_\wp\mathbf{a}\|_2^2. \tag{2.5}$$

The closed-form solution of the above problem is given by $\mathbf{a} = (\mathbf{V}_\wp^T\mathbf{V}_\wp)^{-1}\mathbf{V}_\wp^T\widehat{\mathbf{y}}_\wp$. By using the solution from (2.5), $\widehat{\mathbf{y}}_m$ is approximated by

$$\widehat{\mathbf{y}}_m \;\approx\; \mathbf{V}_m\mathbf{a} \;=\; \mathbf{V}_m(\mathbf{V}_\wp^T\mathbf{V}_\wp)^{-1}\mathbf{V}_\wp^T\widehat{\mathbf{y}}_\wp. \tag{2.6}$$

Equivalently, an approximation of the unknown component $\widehat{y}_i$ is given by

$$\widehat{y}_i = \sum_{j=1}^{k} v_{ij}a_j, \qquad i \in \mathcal{M}, \tag{2.7}$$

where v_{ij} is the element in row i and column j of the basis matrix $\mathbf{V}$. From the normalization formula (2.3), a direct approximation of the original non-normalized data is given by

$$\widehat{w}_i = s_i\widehat{y}_i + \bar{w}_i, \qquad i \in \mathcal{M} \tag{2.8}$$

which is equivalent to the following matrix form

$$\widehat{\mathbf{w}}_m = \mathbf{D}_m^{-1}\widehat{\mathbf{y}}_m + \bar{\mathbf{w}}_m, \tag{2.9}$$

where $\widehat{\mathbf{w}}_m := \mathbf{M}^T\widehat{\mathbf{w}}$ and $\bar{\mathbf{w}}_m := \mathbf{M}^T\bar{\mathbf{w}}$, $\mathbf{D}_m^{-1} = diag(s_{m_1}, \ldots, s_{m_{n_m}}) \in \mathbb{R}^{n_m \times n_m}$. Alternatively, we can introduce a range of possible approximation by using a similar notion as N **standard deviation of the mean**, $N = 1, 2, 3$, which is given by

$$\bar{w}_i \pm s_i\widehat{y}_i, \qquad i \in \mathcal{M}.$$

In particular, let $L_i = \min\{\bar{w}_i - s_i\widehat{y}_i,\ \bar{w}_i + s_i\widehat{y}_i\}$ and $U_i = \max\{\bar{w}_i - s_i\widehat{y}_i,\ \bar{w}_i + s_i\widehat{y}_i\}$. Then, L_i and U_i are the lower and upper bounds, respectively, of the approximation $\widehat{w}_i$, i.e.

$$\widehat{w}_i \in [L_i, U_i], \qquad i \in \mathcal{M}. \tag{2.10}$$

The steps described above are summarized in Algorithm 2, which will be used to generate numerical results in the next section.

Algorithm 2 Algorithm for approximating missing data

Inputs:
- Complete snapshot set $\{\mathbf{w}_j\}_{j=1}^{n_s} \subset \mathbb{R}^n$ and
- Incomplete data $\widehat{\mathbf{w}} \in \mathbb{R}^n$ with
known entries $\widehat{w}_j$, $j \in \mathcal{P}$ and unknown entries s $\widehat{w}_j$, $j \in \mathcal{M}$

Outputs:
- Approximation: $\widehat{\mathbf{w}}_m = [\widehat{w}_j]$, $j \in \mathcal{M} = \{m_1, m_2, \ldots, m_{n_m}\}$
- Approximated Range $[L_i, U_i]$ for $\widehat{w}_i$, $i \in \mathcal{M}$

1: Compute $\bar{\mathbf{w}} = \frac{1}{m}\sum_{j=1}^{n_s} \mathbf{w}_j \in \mathbb{R}^n$.
2: Compute s_i, $i = 1, ..., n$ from (2.2) and
form $\mathbf{D} = diag(1/s_1, 1/s_2, \ldots, 1/s_n) \in \mathbb{R}^{n\times n}$.
3: Set $\mathbf{y}_j = \mathbf{D}(\mathbf{w}_j - \bar{\mathbf{w}})$, $j = 1, 2, ..., n_s$.
4: Create snapshot matrix : $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{n_s}] \in \mathbb{R}^{n\times n_s}$, $r = rank(\mathbf{Y})$.
5: Compute POD basis $\mathbf{V}$ of rank $k \le r$ for $\mathbf{Y}$ from Algorithm 1.
6: Compute $\widehat{\mathbf{y}}_\wp$: $\widehat{y}_i = \frac{\widehat{w}_i - \bar{w}_i}{s_i}$, $\quad i \in \mathcal{P}$.
7: Find coefficient vector $\mathbf{a}$ from (2.5).
8: Compute:
- Approximation: $\widehat{\mathbf{w}}_m = \mathbf{D}_m^{-1}\mathbf{V}_m\mathbf{a} + \bar{\mathbf{w}}_m$ from (2.6) and (2.9).
- Approximated interval: $[L_i, U_i]$, $i \in \mathcal{M}$, from (2.10)where
$L_i = \min\{\bar{w}_i - s_i\widehat{y}_i,\ \ \bar{w}_i + s_i\widehat{y}_i\}$ and $U_i = \max\{\bar{w}_i - s_i\widehat{y}_i,\ \ \bar{w}_i + s_i\widehat{y}_i\}$.

3 Numerical Results

This section considers the data representing numbers of in-patients with 70 different cause groups, provided by Office of the Permanent Secretary for Public Health, Ministry of Public Health, Saraburi provice, Thailand. This information is compiled by Statistical Forecasting Bureau, National Statistical Office and available on the website:
`http://service.nso.go.th/nso/web/statseries/statseries.html`. Some examples of these 70 cause groups are dengue hemorrhagic fever and other mosquito-borne viral hemorrhagic fever, viral hepatitis, human immunodeficiency virus (HIV) disease, influenza, malaria, thalassemia, alcoholic liver diseases, and motorcycle rider injured transport accidents.

We will first consider the accuracy in reconstructing the complete data set in Section 3.1. Then we will use the Gappy POD approach in Algorithm 2 to approximate the missing data of incomplete sample in Section 3.2.

3.1 Reconstruction Data

To apply the approaches described in the previous sections, we can consider each snapshot or sample as the data in each year that consists of 70 numbers of patients with different cause groups (i.e. 70 features in each data). Due to the large differences in the numbers of patients for 70 diseases, the standard deviations and the means of these data can be extremely difference as shown in Figure 1. To handle this variation, the data set would be normalized by subtracting mean and dividing the standard deviation shown in Step 2 of Algorithm 2, which is not required, in general, when reconstructing data for the flow applications, e.g. [8].

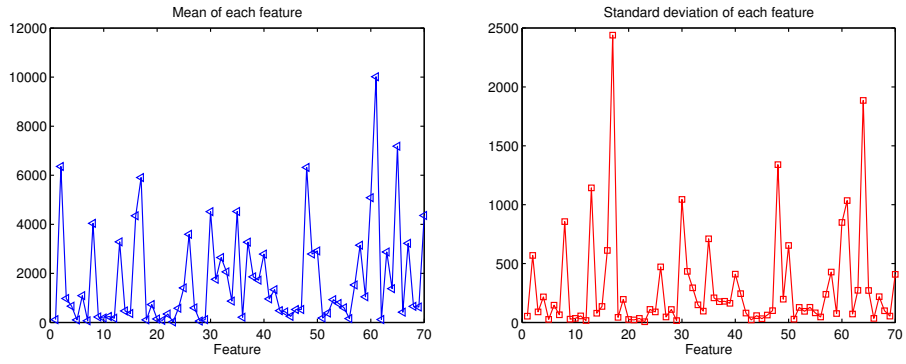

Figure 1: Means and standard deviations of 70 features that represent numbers of patients in 70 different diseases or cause groups.

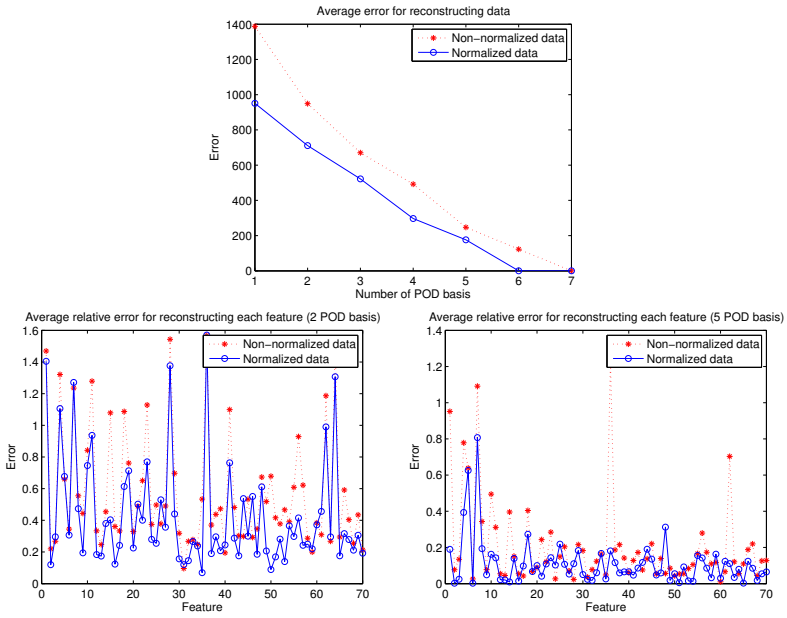

Figure 2: Comparison of average relative errors from nomalized and nomalized data.

To emphasize the importance of normalization, the comparisons of reconstructing errors for data from 2003 to 2010 are provided in Figure 2 for both normalized and non-normalized data. Notice from the first plot of Figure 2 that it is more accurate when the normalized data is used for all cases of dimensions for POD basis. The plots in Figure 3 and the first plot in Figure 2 demonstrate that, as the dimension of POD basis vectors used in the reconstruction increases, the error generally decreases.

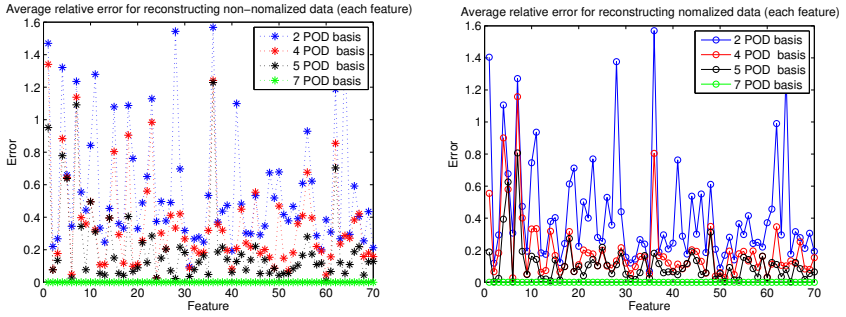

Figure 3: Comparison of average relative errors from nomalized and nomalized data using 2, 4, 5, and 7 POD basis vectors.

In conclusion, the numerical experiment in this section confirms the convergence of the reconstruction accuracy as the number of POD basis vector increases, as well as suggests that using the normalization could improve accuracy of the approximation, which will be used in the next section.

3.2 Approximation of Missing Data

We present two numerical tests in this section. The first one, in Section 3.2.1, uses POD basis constructed from 7 complete samples from the year 2003 to the year 2009 and apply GPOD to construct approximated interval for missing data. The second numerical test, in Section 3.2.2, considers two different POD basis sets: one is constructed from 5 complete samples from 2003-2007, and the other one is constructed from 7 complete samples from 2003-2009 as in the first numerical test. In each of these tests, we will consider different percentages of missing data.

3.2.1 Numerical Test 1

Missing data in 2010 (Cause groups of in-patients)	Approximation (2.9)	Approximated Range (2.10)	True data
1. Ca liver	249.30	$[199.27, 249.30]$	257
2. Mental and behavioral disorders due to psychoactive substance use	757.73	$[699.12, 757.73]$	959
3. Chronic rheumatic heart diseases	97.10	$[97.10, 133.47]$	101
4. Asthma and acute severe asthma	1875.69	$[1569.17, 1875.69]$	1893
5. Diseases of the skin and subcutaneous tissue	2762.88	$[2762.88, 2816.26]$	2805
6. Pregnancy with abortive outcome	1141.56	$[950.16, 1141.56]$	972
7. Poisoning and toxic effect by accidental event self-harm, assault and event of undetermined intent	703.31	$[551.26, 703.31]$	623

Table 1: Comparison of true data (number of patients) with the approximated value (2.9) and the approximated range (2.10) from Algorithm 2 with POD basis of dimension 3.

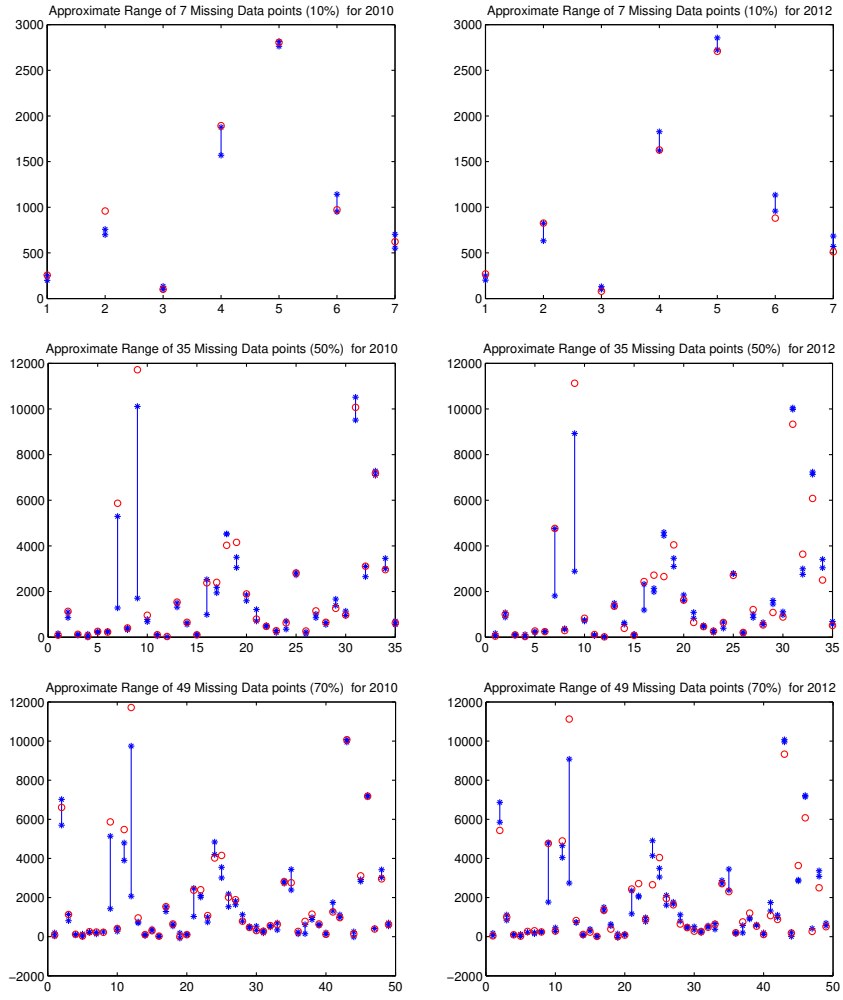

Figure 4: Approximated intervals and the true data values when 10%, 50% , and 70% of 70 features are missing in 2010 and 2012. The red circle dots are the true data values and the blue lines indicate the approximated interval obtained from (2.10) with 3 POD basis vectors.

In this section, we use a set of complete 7 snapshots corresponding to the data in 2003-2009 to construct a POD basis from Algorithm 1. This section first considers the case when there are 10% of 70 components (features) missing in the sample for 2010, which are not included in the complete sample set. In particular, 7 components of the data in 2010 are chosen to be missing randomly. By applying Algorithm 2 with POD basis of rank 3, we obtain the approximation (2.9) and the predicted range (2.10) as shown in Table 1. Notice that almost every true data lies in the approximated range.

Figure 4 also illustrates the approximated ranges and the true data values for missing features for 2010 given in Table 1 in the first plot together with other similar plots for 2010 and 2012 with 10%, 50%, and 70% missing data.

Notice from Figure 4 that, as the number of missing features increases, there are more approximated ranges that fail to include the true values. However, almost all true data values of the missing features lie close to the approximated ranges. Since the scales for these missing features are very different, it might be hard to compare the accuracy of the approximated missing values. To decrease this variation effect, we will next consider average relative error for each approximation, which is shown in Table 2 of the next section.

3.2.2 Numerical Test 2

Two different POD basis sets are used in this section: one is constructed from 5 complete samples from 2003-2007, and the other one is constructed from 7 complete samples from 2003-2009 as in Section 3.2.1. For each of these POD basis sets, we consider average relative errors over 3 years of incomplete samples from 2010 to 2012 given by

$$\mathcal{E} = \frac{1}{n_m} \sum_{i \in \mathcal{M}} \frac{E_i^{2010} + E_i^{2011} + E_i^{2012}}{3\bar{w}_i}, \tag{3.1}$$

where E_i^k is the absolute error between the approximation from (2.9) and the true data value in year k of feature $i \in \mathcal{M}$, for $k = 2010, 2011, 2012$, with index set $\mathcal{M}$ of missing data, and $\bar{w}_i$ is the mean of feature i from the complete sample set $\mathcal{W}$. The average relative errors in Table 2, which are computed by using (3.1), consider 4 different cases of missing data, i.e. 10%, 50%, 70% and 90% missing data. These errors demonstrate that the accuracy could be increased by using more complete samples to construct POD basis. As demonstrated in Figure 4, the results in Table 2 also show that the approximation errors increase when there are more missing data points in the incomplete sample.

Number of missing data: n_m	Average Relative Error (3.1) (POD: 5 complete samples – years 2003-2007)	Average Relative Error (3.1) (POD: 7 complete samples – years 2003-2009)
7 (10%)	0.1426	0.1338
35 (50%)	0.2322	0.2170
49 (70%)	0.2736	0.2592
63 (90%)	0.9386	0.4119

Table 2

Table 2: Average relative error (3.1) of three incomplete samples from 2010 to 2012 when using Gappy POD approximation with 3 POD basis vectors from two cases of 5 and 7 complete samples from 2003-2007 and 2003-2009, respectively, in Algorithm 1.

4 Conclusion

This work has shown an application of GPOD on the data representing number of in-patients arranged by 70 cause groups in Saraburi provice, Thailand. A simple normalization using mean and standard deviation of the data has been introduced and shown to be efficient for avoiding the effect of large variation in the components of sample data. The numerical results demonstrate that GPOD can be used effectively to approximate data when partial components or features are missing. Almost all approximated intervals given in Algorithm 2 can capture the true data values accurately with different amount of missing components, e.g.10%, 50%, 70%. These results also suggest that the accuracy can be improved by using more complete samples to construct the POD basis.

All numerical experiments in the previous section have illustrated the possibility of using GPOD with minor modification to recover missing or unavailable features, even though the data does not process smooth continuity among nearby samples or features. This nature of data set is different from most existing applications of GPOD, for example, in flow field or image reconstruction. GPOD approach has also shown the potential to predict trends on a larger scale for this type of data. Theoretical analysis of GPOD approach can be considered in the future to provide a rigorous error bound for the approximation of missing data.

Acknowledgements : The authors would like to thank the referees for the comments and suggestions on this manuscript. This work is supported by The Thailand Research Fund (TRF) grant for new researcher: Grant No. TRG5880216.

References

[1] J.L. Lumley, The Structure of Inhomogeneous Turbulent Flows, in Atmospheric Turbulence and Radio Wave Propagation (A.M. Yaglom and V. I. Tararsky, eds.), Nauka, Moscow, 1967.

[2] J.L. Lumley, Stochastic Tools in Turbulence, Academic Press, New York, 1970.

[3] G. Berkooz, P. Holmes, J.L. Lumley, The proper orthogonal decomposition in the analysis of turbulent flows, Annual Rev. Fluid Mech (1993) 539-575.

[4] R. Gurka, A. Liberzon, G. Hetsroni, POD of vorticity fields: A method for spatial characterization of coherent structures, International Journal of Heat and Fluid Flow 27 (3) (2006) 416-423.

[5] F. Lanata, A.D. Grosso, Damage detection and localization for continuous static monitoring of structures using a proper orthogonal decomposition of signals, Smart Materials and Structures 15 (6) (2006) 1811.

[6] E. Schenone, Reduced Order Models, Forward and Inverse Problems in Cardiac Electrophysiology, Theses, Universit e Pierre et Marie Curie - Paris VI, November 2014.

[7] L. Sirovich, Turbulence and the dynamics of coherent structures. i. coherent structures, Quart. Appl. Math. 45 (3) (1987) 561-571.

[8] T. Bui-Thanh, M. Damodaran, K. Willcox, Aerodynamic data reconstruction and inverse design using proper orthogonal decomposition, AIAA. 42 (8) (2004) 1505-1516.

[9] A.I. Moreno, A.A. Jarzabek, J.M. Perales, J.M. Vega, Aerodynamic database reconstruction via gappy high order singular value decomposition, Aerospace Science and Technology 52 (2016) 115-128.

[10] K. Bizon, G. Continillo, S.S. Merola, B.M. Vaglieco, Reconstruction of flame kinematics and analysis of cycle variation in a spark ignition engine by means of proper orthogonal decomposition, Computer Aided Chemical Engineering 26 (2009) 1039-1043.

[11] O. Choi, M.C. Lee, Investigation into the combustion instability of synthetic natural gases using high speed flame images and their proper orthogonal decomposition, International Journal of Hydrogen Energy 41 (45) (2016) 20731-20743.

[12] E. Bouhoubeiny, P. Druault, Note on the POD-based time interpolation from successive PIV images, Comptes Rendus Mcanique 337 (11) (2009) 776-780.

[13] M. Wang, D. Dutta, K.Kim, J.C. Brigham, A computationally efficient approach for inverse material characterization combining gappy {POD} with direct inversion, Computer Methods in Applied Mechanics and Engineering 286 (2015) 373-393.

[14] J. Lei, J.H. Qiu, S. Liu, K. Willcox, Dynamic reconstruction algorithm for electrical capacitance tomography based on the proper orthogonal decomposition, Applied Mathematical Modelling 39 (22) (2015) 6925-6940.

[15] X.H. Sun, M.H. Xu, Optimal control of water flooding reservoir using proper orthogonal decomposition, Journal of Computational and Applied Mathematics 320 (2017) 120-137.

[16] S. Volkwein, Proper Orthogonal Decomposition: Applications in Optimization and Control, CEA-EDFINRIA Numerical Analysis Summer School, 2007.

(Received 17 April 2017)
(Accepted 30 June 2017)