AIAA JOURNAL
Vol. 48, No. 6, June 2010

Unifying Perspective for Gappy Proper Orthogonal Decomposition and Probabilistic Principal Component Analysis

Kyunghoon Lee* and Dimitri N. Mavris†
Georgia Institute of Technology, Atlanta, Georgia 30332-0150

DOI: 10.2514/1.45750

In aerospace engineering, various problems such as restoring impaired experimental flow data can be handled by gappy proper orthogonal decomposition. Similar to gappy proper orthogonal decomposition, probabilistic principal component analysis can approximate missing data with the help of an expectation-maximization algorithm, yielding an expectation-maximization algorithm for probabilistic principal component analysis (expectation-maximization principal component analysis). Although both gappy proper orthogonal decomposition and expectation-maximization principal component analysis address the same missing-data-estimation problem, their antithetical formulation perspectives hinder their direct comparison; the development of the former is deterministic, whereas that of the latter is probabilistic. To effectively differentiate both methods, this research provides a unifying least-squares perspective to qualitatively dissect them within a unified least-squares framework. By virtue of the unifying least-squares perspective, gappy proper orthogonal decomposition and the expectation-maximization principal component analysis turn out to be similar in that they are twofold: basis and least-squares coefficient evaluations. On the other hand, they are dissimilar because the expectation-maximization principal component analysis, unlike gappy proper orthogonal decomposition, dispenses with either a gappy norm or a proper orthogonal decomposition basis. To illustrate the theoretical analysis of both methods, numerical experiments using simple and complex data sets quantitatively examine their performance in terms of convergence rates and computational cost. Finally, comprehensive comparisons, including theoretical and numerical aspects, establish that the expectation-maximization principal component analysis is simpler and thereby more efficient than gappy proper orthogonal decomposition.

Nomenclature

$\mathbf{b}$	=	least-squares coefficient of gappy proper orthogonal decomposition
$\mathbf{C}$	=	model covariance matrix
$\mathbf{c}$	=	generalized least-squares coefficient
d	=	dimension of an observed variable
$\mathbf{I}$	=	identity matrix
k	=	number of iterations
$\mathcal{L}$	=	log-likelihood function
$\mathcal{L}_C$	=	complete data log-likelihood function
$\mathcal{N}$	=	Gaussian probability distribution
N	=	snapshot ensemble size
$\mathbf{n}$	=	mask vector
p	=	probability density function
$\mathbf{Q}$	=	orthogonal matrix
q	=	dimension of a latent variable, i.e., model selection
R	=	averaged estimation error
$\mathbb{R}^n$	=	n-dimensional real number space
r	=	estimation error
$\mathbf{S}$	=	sample covariance matrix
$\mathbf{T}$	=	collection of error-accounted observed variables
$\mathbf{t}$	=	error-accounted observed variable
$\mathbf{V}$	=	eigenvector matrix, i.e., a proper orthogonal decomposition basis
$\mathbf{V}_e$	=	guessed $\mathbf{V}_q$ obtained from $\tilde{\mathbf{Y}}^{(0)}$
$\mathbf{V}_q$	=	matrix of the first q eigenvectors
$\mathbf{W}$	=	factor loadings
$\mathbf{X}$	=	collection of latent variables
$\mathbf{x}$	=	latent variable
$\mathbf{Y}$	=	collection of observed variables
$\mathbf{y}$	=	observed variable
$\mathbf{0}$	=	zero matrix
$\mathbf{1}_N$	=	vector with N ones
$\bar{\cdot}$	=	sample mean
$\circ$	=	Hadamard product, i.e., pointwise multiplication
$\mathring{\cdot}$	=	data with missing values
$\dot{\cdot}$	=	mean-centered data
$\langle\cdot\rangle$	=	expectation
$\tilde{\cdot}$	=	estimation
α	=	general norm
$\boldsymbol{\epsilon}$	=	error variable
$\boldsymbol{\Lambda}$	=	eigenvalue matrix
λ	=	eigenvalue
$\boldsymbol{\mu}$	=	mean vector
σ^2	=	variance
$\boldsymbol{\Phi}$	=	basis
$\boldsymbol{\Omega}$	=	diagonal matrix

Subscripts

j	=	jth vector
L^2	=	L^2 norm
ML	=	maximum likelihood estimate
n	=	gappy norm

Superscript

k	=	kth iteration

Presented as Paper 3899 at the 39th AIAA Fluid Dynamics Conference, San Antonio, TX, 22–25 June 2009; received 1 June 2009; revision received 2 December 2009; accepted for publication 5 January 2010.

*Ph.D. Candidate, Graduate Research Assistant, Aerospace Systems Design Laboratory.

†Professor, Aerospace Systems Design Laboratory. Lifetime Member AIAA.

I. Introduction

FOR identifying flow characteristics and simplifying high-dimensional flow data, proper orthogonal decomposition (POD), also known as principal component analysis (PCA) or the Karhunen–Loève transform, has been conducive to distilling an orthogonal basis from high-fidelity flow data. As POD is impotent even in the slightest absence of data, Everson and Sirovich [1] devised gappy POD, which restores missing data in observations by solving a least-squares problem defined with a gappy norm relying on a POD basis. Initially, gappy POD was applied to repairing marred images of a human face, but its application has spread to aerospace engineering; missing-data estimation is so general that it can epitomize diverse problems as long as appropriate missing-data forms can be devised for them. For instance, Bui-Thanh et al. [2] adopted gappy POD for not only missing aerodynamic data reconstruction but also inverse airfoil design by formulating it as a missing-data-estimation problem. Likewise, Bui-Thanh [3] took advantage of gappy POD for parametric flowfield prediction, and Willcox [4] used it for unsteady flow reconstruction and effective sensor placement. For variable-fidelity analysis, Robinson et al. [5] exploited gappy POD to evaluate mapping from low-fidelity to high-fidelity analysis. In addition, Venturi and Karniadakis [6] investigated gappy POD as a data-assimilation technique and compared it to other reconstruction methods such as local kriging and local linear interpolation. Similarly, Murray and Ukeiley [7,8] and Murray and Seinery [9] employed gappy POD to estimate obscured experimental flow data (common in flow experiments) using particle image velocimetry.

Apart from gappy POD, the same missing-data estimation problem can be tackled by probabilistic principal component analysis (PPCA), which stems from a different theoretical background. Tipping and Bishop [10] formulated PPCA to impart a density model to PCA so as to afford it a probabilistic interpretation. Based on probability and statistics theories, PPCA yields a Gaussian probability distribution for given observations, with its probability parameters yet to be found. Since maximum likelihood estimates (MLEs) for PPCA parameters cannot be found by the method of maximum likelihood if observations have missing data, PPCA has to rely on other statistical inference techniques. For this reason, an expectation-maximization (EM) algorithm that assumes missingness of data is invoked for PPCA, which generates an EM algorithm for PPCA (EM-PCA). By virtue of the EM algorithm, the EM-PCA can naturally deal with missing-data estimation, as can gappy POD, while estimating PPCA parameters. The EM-PCA has been widely used in image processing and pattern recognition fields, and lately in an aerospace realm, Lee et al. [11] explored its potential for aerodynamic data reduction and reconstruction.

Intriguingly, despite the antithetical formulation perspectives of gappy POD and the EM-PCA, they are equally capable of restoring missing data, which raises the following primary questions:

1) What are the similarities and the disparities between the two missing-data estimation algorithms?

2) Which algorithm is more efficient?

3) The previous question being determined, how much more competitive is it than the other?

To facilitate answering these questions, this research proposes a unifying least-squares perspective, which unveils that seemingly irrelevant gappy POD and EM-PCA formulations indeed pertain to solving a generalized least-squares problem. With the help of the least-squares perspective, both gappy POD and the EM-PCA are found to consist of the same two rudimentary steps: basis generation and a least-squares coefficient evaluation. However, they do not share a common basis and norm such that the EM-PCA obviates the gappy norm and the POD basis, which are two indispensable ingredients for gappy POD, using an L^2 norm and a factor loadings instead. After all, these disparities between gappy POD and the EM-PCA will determine their algorithmic characteristics that will eventually lead to their performance differentials.

In summary, for the theoretical and numerical comparisons of gappy POD and the EM-PCA, this paper addresses the aforementioned questions with the proposed least-squares perspective. To begin with, the gappy POD formulation is articulated followed by the development of PPCA and the EM-PCA. After the theory behind each method is described, the two estimation algorithms will be analytically dissected with the unifying least-squares perspective. The theoretical analysis of both methods will shed light on their computational efficiency, and subsequent numerical experiments will validate their projected numerical characteristics based on their theoretical attributes. In conclusion, this paper will ascertain which algorithm between gappy POD and the EM-PCA efficiently estimates missing flow data by integrating the qualitative as well as quantitative comparison results.

II. Theory

For simplicity of exposition, a snapshot $\mathbf{y}_j \in \mathbb{R}^d$ and a snapshot ensemble, i.e., compiled snapshots, $\mathbf{Y} = \{\mathbf{y}_j\}_{j=1}^N \in \mathbb{R}^{d\times N}$ are mean-subtracted such that $\dot{\mathbf{y}}_j = \mathbf{y}_j - \bar{\mathbf{y}}$ and $\dot{\mathbf{Y}} = \mathbf{Y} - \bar{\mathbf{y}}\mathbf{1}_N^T$, respectively, where $\bar{\mathbf{y}} \in \mathbb{R}^d$ is a sample mean given by

$$\bar{\mathbf{y}} = \frac{1}{N}\sum_{j=1}^{N} \mathbf{y}_j$$

and $\mathbf{1}_N \in \mathbb{R}^N$ is a column vector of N ones defined as $\mathbf{1}_N = (1, \ldots, 1)^T$. For notational convenience, the mean-centered notations $\dot{\mathbf{y}}_j$ and $\dot{\mathbf{Y}}$ will henceforth be referred to as $\mathbf{y}_j$ and $\mathbf{Y}$, respectively.

A. Gappy Proper Orthogonal Decomposition

Without any missing data, an arbitrary snapshot $\mathbf{y}_j$ that belongs to a snapshot ensemble $\mathbf{Y}$ can be estimated as a linear combination of the first q POD basis $\mathbf{V}_q = \{\mathbf{v}_j\}_{j=1}^q \in \mathbb{R}^{d\times q}$ such that $\mathbf{y}_j \approx \tilde{\mathbf{y}}_j = \mathbf{V}_q\mathbf{b}_j$. For the best approximation with $\mathbf{V}_q$, a modal coefficient $\mathbf{b}_j \in \mathbb{R}^q$ is found by minimizing a squared error as follows:

$$\min \|\mathbf{y}_j - \mathbf{V}_q\mathbf{b}_j\|_{L^2}^2 \quad \text{subject to } \mathbf{b}_j \tag{1}$$

which yields the optimal coefficient $\mathbf{b}_j = (\mathbf{V}_q^T\mathbf{V}_q)^{-1}\mathbf{V}_q^T\mathbf{y}_j$ for the given POD basis $\mathbf{V}_q$. Similarly, in the presence of gappy data, the same least-squares approach in Eq. (1) can benefit restoring missing data in an incomplete snapshot $\mathring{\mathbf{y}}_j$ by the form of $\mathring{\mathbf{y}}_j \approx \mathbf{V}_q\mathbf{b}_j$, as shown in Eq. (2), provided that the POD basis $\mathbf{V}_q$ is known.

$$\min \|\mathring{\mathbf{y}}_j - \mathbf{V}_q\mathbf{b}_j\|_{L^2}^2 \quad \text{subject to } \mathbf{b}_j \tag{2}$$

However, because of unknown missing elements in $\mathring{\mathbf{y}}_j$, an L^2 norm in Eq. (2) cannot be evaluated correctly. To resolve this issue due to missing data, Everson and Sirovich [1] came up with a gappy norm defined with a gappy inner product $(\cdot,\cdot)_n$ on $\mathbb{R}^d$ such that

$$\|\mathbf{y}_j\|_n^2 := (\mathbf{y}_j, \mathbf{y}_j)_n = (\mathbf{n}_j \circ \mathring{\mathbf{y}}_j, \mathbf{n}_j \circ \mathring{\mathbf{y}}_j)_{L^2} = \|\mathbf{n}_j \circ \mathring{\mathbf{y}}_j\|_{L^2}^2 \tag{3}$$

where $\circ$ denotes a Hadamard product [[12], page 30], i.e., pointwise multiplication, and $\mathbf{n}_j \in \mathbb{R}^d$ is a mask vector corresponding to an incomplete snapshot $\mathring{\mathbf{y}}_j$ to screen out missing data in it. A mask vector $\mathbf{n}_j$ for $\mathring{\mathbf{y}}_j$ is defined by

$$n_{ij} = \begin{cases} 0 & \text{if } \mathring{y}_{ij} \text{ is missing} \\ 1 & \text{if } \mathring{y}_{ij} \text{ is known} \end{cases} \quad \text{for} \quad i = 1, \ldots, d \tag{4}$$

Note that masking by $\mathbf{n}_j$ implies assigning a sample mean $\bar{\mathbf{y}}$ to each missing-data element of $\mathring{\mathbf{y}}_j$, for a snapshot is treated as mean-centered herein, for convenience. With the help of the gappy norm, the previous squared residual in Eq. (2) can be rephrased as

$$r_j^2 = \|\mathbf{n}_j \circ (\mathring{\mathbf{y}}_j - \mathbf{V}_q\mathbf{b}_j)\|_{L^2}^2 = \|\mathbf{y}_j - \mathbf{V}_q\mathbf{b}_j\|_n^2 \tag{5}$$

and the least-squares problem for gappy POD is

$$\min \|\mathbf{y}_j - \mathbf{V}_q \mathbf{b}_j\|_n^2 \quad \text{subject to } \mathbf{b}_j \tag{6}$$

A coefficient $\mathbf{b}_j$ satisfying Eq. (6) can be found by annihilating the first derivative of a squared residual r_j^2 such that

$$\begin{aligned} r_j^2 &= \|\mathbf{y}_j - \mathbf{V}_q \mathbf{b}_j\|_n^2 \\ &= \left(\mathbf{n}_j \circ \left(\mathbf{y}_j - \sum_{i=1}^{q} b_{ij}\mathbf{v}_i\right), \mathbf{n}_j \circ \left(\mathbf{y}_j - \sum_{i=1}^{q} b_{ij}\mathbf{v}_i\right)\right)_{L^2} \\ &= (\mathbf{n}_j \circ \mathbf{y}_j)^T(\mathbf{n}_j \circ \mathbf{y}_j) - 2\sum_{i=1}^{q} b_{ij}(\mathbf{n}_j \circ \mathbf{y}_j)^T(\mathbf{n}_j \circ \mathbf{v}_i) \\ &\quad + \sum_{i=1}^{q}\sum_{k=1}^{q} b_{ij}b_{kj}(\mathbf{n}_j \circ \mathbf{v}_i)^T(\mathbf{n}_j \circ \mathbf{v}_k) \end{aligned} \tag{7}$$

with respect to $\mathbf{b}_j$ as follows:

$$\begin{aligned} \left.\frac{\partial r_j^2}{\partial b_{ij}}\right|_{\mathbf{y}_j, \mathbf{v}_i} &= -2(\mathbf{n}_j \circ \mathbf{y}_j)^T(\mathbf{n}_j \circ \mathbf{v}_i) + 2\sum_{k=1}^{q} b_{kj}(\mathbf{n}_j \circ \mathbf{v}_i)^T(\mathbf{n}_j \circ \mathbf{v}_k) \\ &= 0 \end{aligned} \tag{8}$$

which determines the ith element of $\mathbf{b}_j$ to be

$$b_{ij} = \left(\sum_{k=1}^{q}(\mathbf{n}_j \circ \mathbf{v}_i)^T(\mathbf{n}_j \circ \mathbf{v}_k)\right)^{-1}(\mathbf{n}_j \circ \mathbf{y}_j)^T(\mathbf{n}_j \circ \mathbf{v}_i) \tag{9}$$

where $\mathbf{v}_i$ is the ith vector of $\mathbf{V}_q$. As a result, the least-squares problem in Eq. (6) boils down to a system of q linear equations formulated as

$$M_{ij} b_{ij} = f_{ij}$$

where

$$M_{ij} = \sum_{k=1}^{q}(\mathbf{n}_j \circ \mathbf{v}_i)^T(\mathbf{n}_j \circ \mathbf{v}_k), \qquad f_{ij} = (\mathbf{n}_j \circ \mathbf{y}_j)^T(\mathbf{n}_j \circ \mathbf{v}_i)$$

for $i = 1, \ldots, q$ to reconstruct missing data in $\mathring{\mathbf{y}}_j$. Note that the least-squares coefficient of gappy POD in Eq. (9) has an identical form to the ordinary least-squares coefficient in Eq. (1) such that

$$b_{ij} = \left(\sum_{k=1}^{q}(\mathbf{v}_i^T\mathbf{v}_k)\right)^{-1}(\mathbf{y}_j^T\mathbf{v}_i)$$

for a without-missing-data case, except that every vector in Eq. (9) is masked by $\mathbf{n}_j$. Once $\mathbf{b}_j$ is obtained from Eq. (9), missing components in $\mathring{\mathbf{y}}_j$ are replaced with estimates as follows:

$$\mathring{y}_{ij} = \begin{cases} \sum_{k=1}^{q} b_{kj} v_{ik} & \text{if } n_{ij} = 0 \\ y_{ij} & \text{if } n_{ij} = 1 \end{cases} \quad \text{for} \quad i = 1, \ldots, d$$

which changes only missing data, keeping known data as they are. Before missing-data estimation, gappy POD presupposes that the POD basis $\mathbf{V}_q$ is available; however, the true $\mathbf{V}_q$ is not known a priori if gappy data are the only data at hand. Therefore, for missing-data restoration, gappy POD has to repeat the basis and coefficient evaluations; the former derives an estimated POD basis $\tilde{\mathbf{V}}_q$ from an intermediately recovered snapshot ensemble $\tilde{\mathbf{Y}}$, and the latter rectifies the estimated snapshot ensemble $\tilde{\mathbf{Y}}$ by restoring missing data using the previously obtained $\tilde{\mathbf{V}}_q$.

B. Probabilistic Principal Component Analysis

1. Probability Model for PPCA

a. Latent Variable Model. A PPCA formulation by Tipping and Bishop [10] starts with a latent variable model that presumes a latent variable[‡] $\mathbf{x}_j \in \mathbb{R}^q$ for an observed variable $\mathbf{y}_j \in \mathbb{R}^d$, where $d \gg q$. With a latent variable $\mathbf{x}_j$, an observed variable $\mathbf{y}_j$ can be expressed as

$$\mathbf{y}_j(\mathbf{x}_j; \mathbf{W}) = \mathbf{W}\mathbf{x}_j$$

where $\mathbf{W} \in \mathbb{R}^{d\times q}$ is a factor-loading matrix representing a linear mapping $\mathbf{W}$: $\mathbb{R}^q \mapsto \mathbb{R}^d$ between $\mathbf{x}_j$ and $\mathbf{y}_j$. After an observation error $\boldsymbol{\epsilon} \in \mathbb{R}^d$, independent of a latent variable $\mathbf{x}_j$, is accounted for $\mathbf{y}_j$, a linear latent variable model is defined as

$$\mathbf{t}_j(\mathbf{x}_j; \mathbf{W}, \boldsymbol{\epsilon}) = \mathbf{W}\mathbf{x}_j + \boldsymbol{\epsilon} \tag{10}$$

where $\mathbf{t}_j \in \mathbb{R}^d$ is the error-accounted observation of $\mathbf{y}_j$. Note that the latent variable model in Eq. (10) conveys the idea of dimensionality reduction because a high-dimensional observation $\mathbf{t}_j$ can be delineated by a low-dimensional latent variable $\mathbf{x}_j$ through the mapping $\mathbf{W}$.

b. Probability Model. For the construction of the probability density model of $\mathbf{t}_j$, a unit isotropic Gaussian distribution is assumed for a latent variable $\mathbf{x}_j$ such that $\mathbf{x}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and an isotropic Gaussian distribution is assumed for the error $\boldsymbol{\epsilon}$ such that $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. With these assumptions, a probability density model for $\mathbf{t}_j$ is determined to be a Gaussian distribution such that $\mathbf{t}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, where $\mathbf{C}$ is a model covariance defined as $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$. From the probability distribution of $\boldsymbol{\epsilon}$ such that

$$p(\boldsymbol{\epsilon}; \sigma^2) = (2\pi\sigma^2)^{-d/2}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}\right)$$

the probability distribution of $\mathbf{t}_j$ given $\mathbf{x}_j$ is found as

$$p(\mathbf{t}_j|\mathbf{x}_j; \mathbf{W}, \sigma^2) = (2\pi\sigma^2)^{-d/2}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{t}_j - \mathbf{W}\mathbf{x}_j\|^2\right) \tag{11}$$

using $\boldsymbol{\epsilon} = \mathbf{t}_j - \mathbf{W}\mathbf{x}_j$ in Eq. (10) With the conditional distribution $p(\mathbf{t}_j|\mathbf{x}_j)$ in Eq. (11) and the assumed prior probability of $\mathbf{x}_j$, given by

$$p(\mathbf{x}_j) = (2\pi)^{-q/2}\exp(-\tfrac{1}{2}\mathbf{x}_j^T\mathbf{x}_j)$$

the marginal probability of $\mathbf{t}_j$ is evaluated as follows:

$$\begin{aligned} p(\mathbf{t}_j; \mathbf{W}, \sigma^2) &= \int p(\mathbf{t}_j|\mathbf{x}_j)p(\mathbf{x}_j)\,\mathrm{d}\mathbf{x}_j \\ &= (2\pi)^{-d/2}|\mathbf{C}|^{-1/2}\exp\left(-\frac{1}{2}\mathbf{t}_j^T\mathbf{C}^{-1}\mathbf{t}_j\right) \end{aligned} \tag{12}$$

In addition, the posterior probability of $\mathbf{x}_j$ given $\mathbf{t}_j$ is found by Bayes's rule such that

$$\begin{aligned} p(\mathbf{x}_j|\mathbf{t}_j) &= (2\pi)^{-q/2}|\sigma^{-2}\mathbf{M}|^{1/2} \\ &\quad \times \exp(-\tfrac{1}{2}(\mathbf{x}_j - \mathbf{M}^{-1}\mathbf{W}^T\mathbf{t}_j)^T(\sigma^{-2}\mathbf{M})(\mathbf{x}_j - \mathbf{M}^{-1}\mathbf{W}^T\mathbf{t}_j)) \end{aligned} \tag{13}$$

where a matrix $\mathbf{M}$ is given by $\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}$.

2. Maximum Likelihood Estimators of PPCA

From the marginal probability of $\mathbf{t}_j$ in Eq. (12), a log-likelihood function is evaluated as

$$\begin{aligned} \mathcal{L}(\mathbf{W}, \sigma^2) &= \sum_{j=1}^{N} \ell_n\, p(\mathbf{t}_j; \mathbf{W}, \sigma^2) \\ &= -\frac{Nd}{2}\ell_n(2\pi) - \frac{N}{2}\ell_n|\mathbf{C}| - \frac{N}{2}\operatorname{tr}(\mathbf{C}^{-1}\mathbf{S}) \end{aligned} \tag{14}$$

where $\mathbf{S}$ is a sample covariance matrix defined by $\mathbf{S} = (1/N)\mathbf{T}\mathbf{T}^T$. For the derivation of PPCA parameter MLEs, such as $\mathbf{W}_{\mathrm{ML}}$ and σ^2_{ML}, the method of maximum likelihood yields the following equations:

[‡]A latent variable is unobservable, so it can be inferred only from an observed variable.

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = 0 \Rightarrow (\mathbf{C}^{-1}\mathbf{S} - \mathbf{I})\mathbf{C}^{-1}\mathbf{W} = 0$$
$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = 0 \Rightarrow \mathbf{I} - \mathbf{C}^{-1}\mathbf{S} = 0 \tag{15}$$

albeit the equations in Eq. (15) cannot analytically determine $\mathbf{W}_{\mathrm{ML}}$ and σ^2_{ML}. Nevertheless, Eq. (15) helps interpret $\mathbf{W}_{\mathrm{ML}}$ and σ^2_{ML} given that the singular value decomposition (SVD) of $\mathbf{W}$ is known as $\mathbf{W} = \mathbf{Q}_1\mathbf{\Omega}\mathbf{Q}_2^T$, where $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are orthogonal matrices and $\mathbf{\Omega}$ is a diagonal matrix. After feeding the SVD of $\mathbf{W}$ into Eq. (15) Tipping and Bishop [10] derived $\mathbf{W}_{\mathrm{ML}}$ and σ^2_{ML} in the following forms:

$$\sigma^2_{\mathrm{ML}} = \frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j, \quad \text{and} \quad \mathbf{W}_{\mathrm{ML}} = \mathbf{V}_q(\mathbf{\Lambda}_q - \sigma^2\mathbf{I}_q)^{1/2}\mathbf{Q}_2^T \tag{16}$$

where $\mathbf{V}_q$ is a matrix of q eigenvectors of $\mathbf{S}$ equivalent to the first q POD basis, $\mathbf{\Lambda}_q$ is a diagonal matrix listing corresponding q eigenvalues of $\mathbf{S}$ in diagonal, and $\mathbf{Q}_2$ is an orthogonal matrix expressing an arbitrary rotation. In Eq. (16), σ^2_{ML} conveys the average of abandoned $d-q$ eigenvalues of $\mathbf{S}$, which indicates an error by projecting d-dimensional data onto a q-dimensional subspace. Likewise, $\mathbf{W}_{\mathrm{ML}}$ in Eq. (16) implies scaled and rotated eigenvectors $\mathbf{V}_q$ by a diagonal matrix $(\mathbf{\Lambda}_q - \sigma^2\mathbf{I}_q)^{1/2}$ and by an orthogonal matrix $\mathbf{Q}_2$, respectively. Note that $\mathbf{W}_{\mathrm{ML}}$ spans a q-dimensional subspace as does $\mathbf{V}_q$, but it is *not* orthogonal.

3. EM Algorithm for PPCA

Although a sample covariance matrix $\mathbf{S}$ is conducive indirectly evaluating the PPCA parameter MLEs in Eq. (16) as Tipping and Bishop showed [10], an EM algorithm is indispensable for PPCA to find its parameter MLEs insofar as observations have missing data. Dempster et al. [13] developed the EM algorithm adumbrating several feasible applications of it, and Rubin and Thayer [14] later articulated it for a factor analysis model pertinent to PPCA. Afterward, Tipping and Bishop [10] capitalized on the EM algorithm formulating the EM-PCA to derive PPCA parameter MLEs. For parameter estimation, the EM algorithm iteratively yields parameter MLEs alternating two steps: an expectation step (E-step) and a maximization step (M-step). The E-step estimates unknown variables given current parameter estimates, and the subsequent M-step corrects the parameter estimates given the estimated variables in the previous E-step so as to maximize the expectation of a log-likelihood function. Literally, the EM algorithm requires the presence of hidden or missing data, yet applications of the EM algorithm are not limited, since missingness is a virtual device to exploit the EM algorithm. In general, the EM algorithm can always reach a local maximum of a likelihood function [13] and, particularly for PPCA, the EM-PCA can locate the global maximum of a likelihood function [10].

For the derivation of the EM-PCA, an observed variable $\mathbf{t}_j$ and a latent variable $\mathbf{x}_j$ that is hidden are coalesced to form a complete data set $(\mathbf{t}_j, \mathbf{x}_j)$. The probability distribution of the complete data set can be formulated as a joint distribution of $\mathbf{t}_j$ and $\mathbf{x}_j$ such that

$$\begin{aligned} p(\mathbf{t}_j, \mathbf{x}_j) &= p(\mathbf{t}_j|\mathbf{x}_j)p(\mathbf{x}_j) \\ &= (2\pi\sigma^2)^{-d/2}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{t}_j - \mathbf{W}\mathbf{x}_j\|^2\right)(2\pi)^{-q/2} \\ &\times \exp\left(-\frac{1}{2}\|\mathbf{x}_j\|^2\right) \end{aligned} \tag{17}$$

Given the joint probability distribution in Eq. (17), both E-step and M-step of the EM-PCA are developed as follows:

a. Expectation Step. A latent variable $\mathbf{x}_j$ and an error-accounted observed variable $\mathbf{t}_j$ with missing data are the two unknown variables to be estimated. First, the posterior probability of $\mathbf{x}_j$ is given by

$$p(\mathbf{x}_j|\mathbf{t}_j) = \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^T\mathbf{t}_j, \sigma^2\mathbf{M}^{-1})$$

from Eq. (13), which yields the expectation of $\mathbf{x}_j$ as $\langle\mathbf{x}_j\rangle = \mathbf{M}^{-1}\mathbf{W}^T\mathbf{t}_j$. Likewise, the expectation of $\mathbf{t}_j$ is $\langle\mathbf{t}_j\rangle = \mathbf{W}\mathbf{x}_j$ from the conditional probability of $\mathbf{t}_j$ in Eq. (11). For each collection of $\mathbf{x}_j$ and $\mathbf{t}_j$, denoted as $\mathbf{X} = \{\mathbf{x}_j\}_{j=1}^N \in \mathbb{R}^{q\times N}$ and $\mathbf{T} = \{\mathbf{t}_j\}_{j=1}^N \in \mathbb{R}^{d\times N}$, respectively, their expected values are

$$\langle\mathbf{X}\rangle = \mathbf{M}^{-1}\mathbf{W}^T\mathbf{T}, \qquad \langle\mathbf{T}\rangle = \mathbf{W}\mathbf{X}$$

Normally, an E-step only evaluates the expectation of a hidden variable $\langle\mathbf{x}_j\rangle$, and when it deals with both $\langle\mathbf{x}_j\rangle$ and $\langle\mathbf{t}_j\rangle$, it is called a generalized E-step [15].

b. Maximization Step. The log-likelihood function of the complete data set is constructed as

$$\begin{aligned} \mathcal{L}_C(\mathbf{W}, \sigma^2) &= \sum_{j=1}^{N}\ell_n\, p(\mathbf{t}_j, \mathbf{x}_j; \mathbf{W}, \sigma^2) \\ &= -\frac{N}{2}(d+q)\,\ell_n(2\pi) - \frac{Nd}{2}\ell_n(\sigma^2) \\ &-\frac{1}{2\sigma^2}\sum_{j=1}^{N}\|\mathbf{t}_j - \mathbf{W}\mathbf{x}_j\|^2 - \frac{1}{2}\sum_{j=1}^{N}\|\mathbf{x}_j\|^2 \end{aligned}$$

from the joint probability in Eq. (17), and its expectation is evaluated as

$$\begin{aligned} \langle\mathcal{L}_C\rangle &= -\frac{N}{2}(d+q)\,\ell_n(2\pi) - \frac{Nd}{2}\ell_n(\sigma^2) \\ &-\frac{1}{2\sigma^2}\,\mathrm{tr}(\mathbf{T}^T\mathbf{T} - 2\langle\mathbf{X}\rangle^T\mathbf{W}^T\mathbf{T} + \mathbf{W}^T\mathbf{W}\langle\mathbf{X}\mathbf{X}^T\rangle) - \frac{1}{2}\,\mathrm{tr}(\langle\mathbf{X}\mathbf{X}^T\rangle) \end{aligned} \tag{18}$$

where

$$\langle\mathbf{X}\rangle = \mathbf{M}^{-1}\mathbf{W}^T\mathbf{T}, \qquad \langle\mathbf{X}\mathbf{X}^T\rangle = \sigma^2\mathbf{M}^{-1} + \langle\mathbf{X}\rangle\langle\mathbf{X}\rangle^T$$

Finally, parameter estimates that maximize $\langle\mathcal{L}_C\rangle$ in Eq. (18) can be found from the first derivatives of $\langle\mathcal{L}_C\rangle$ with respect to $\mathbf{W}$ and σ^2 as follows:

$$\begin{aligned} \frac{\partial\langle\mathcal{L}_C\rangle}{\partial\mathbf{W}} = 0 &\Rightarrow \tilde{\mathbf{W}} = \mathbf{T}\langle\mathbf{X}\rangle^T\langle\mathbf{X}\mathbf{X}^T\rangle^{-1} \\ \frac{\partial\langle\mathcal{L}_C\rangle}{\partial\sigma^2} = 0 &\Rightarrow \tilde{\sigma}^2 = \frac{1}{Nd}\,\mathrm{tr}(\mathbf{T}^T\mathbf{T} - 2\langle\mathbf{X}\rangle^T\mathbf{W}^T\mathbf{T} + \mathbf{W}^T\mathbf{W}\langle\mathbf{X}\mathbf{X}^T\rangle) \end{aligned} \tag{19}$$

c. EM-PCA Algorithm. In summary, under a zero-noise limit such that $\lim\sigma^2 \to 0$, the EM-PCA for missing-data estimation is as follows.

Generalized E-step:

$$\langle\mathbf{X}\rangle = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{T} \tag{20a}$$

$$\langle\mathbf{T}\rangle = \mathbf{W}\mathbf{X} \tag{20b}$$

M-step:

$$\tilde{\mathbf{W}} = \mathbf{T}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} \tag{20c}$$

Note that matrix multiplications and inversions comprise the equations of the EM-PCA in Eq. (20), and each matrix inversion deals with a q by q matrix that is the smallest matrix in Eq. (20).

III. Formulation of a Unifying Least-Squares Perspective

A. Gappy POD Recast in Forms of Matrix Multiplication

The least-squares problem for gappy POD in Eq. (6), translated from the gappy norm to the L^2 norm, is

$$\min\|\mathbf{n}_j \circ (\mathring{\mathbf{y}}_j - \mathbf{V}_q\mathbf{b}_j)\|_{L^2}^2 \quad \text{subject to } \mathbf{b}_j$$

involving the Hadamard product, which is not as transparent as ordinary matrix multiplication. To facilitate a comparative study, this

research proposes to recast the Hadamard product into matrix multiplication such that

$$\mathbf{n}_j \circ \mathbf{y}_j = \mathbf{N}_j \mathbf{y}_j \tag{21}$$

by introducing a diagonal matrix $\mathbf{N}_j \in \mathbb{R}^{d\times d}$ for $\mathbf{n}_j \in \mathbb{R}^d$, which lists $\mathbf{n}_j$ in its diagonal, i.e., $\mathbf{N}_j = \text{diag}(\mathbf{n}_j)$. Note that $\mathbf{N}_j$ is symmetric, for it is diagonal, and because of zeros and ones in the diagonal, it is positive semidefinite and singular. Moreover, $\mathbf{N}_j$ is a projection since $\mathbf{N}_j^2 = \mathbf{N}_j$. With the help of the relationship in Eq. (21), the gappy norm defined in Eq. (3) can be expressed as

$$\begin{aligned}\|\mathbf{y}_j\|_n^2 &= (\mathbf{y}_j, \mathbf{y}_j)_n = (\mathbf{n}_j \circ \mathring{\mathbf{y}}_j, \mathbf{n}_j \circ \mathring{\mathbf{y}}_j)_{L^2} = (\mathbf{N}_j\mathbf{y}_j, \mathbf{N}_j\mathbf{y}_j)_{L^2} \\ &= \mathbf{y}_j^T \mathbf{N}_j \mathbf{y}_j \end{aligned} \tag{22}$$

which reveals that a squared gappy norm is equivalent to a weighted inner product of $\mathbf{y}_j$ with either zero or one weight.

With the transformation shown in Eq. (22), the squared estimation residual of gappy POD, shown in Eq. (7), can be rephrased in matrix multiplications:

$$\begin{aligned} r_j^2 &= \|\mathbf{y}_j - \mathbf{V}_q\mathbf{b}_j\|_n^2 = (\mathbf{n}_j \circ (\mathring{\mathbf{y}}_j - \mathbf{V}_q\mathbf{b}_j), \mathbf{n}_j \circ (\mathring{\mathbf{y}}_j - \mathbf{V}_q\mathbf{b}_j))_{L^2} \\ &= (\mathbf{N}_j(\mathbf{y}_j - \mathbf{V}_q\mathbf{b}_j), \mathbf{N}_j(\mathbf{y}_j - \mathbf{V}_q\mathbf{b}_j))_{L^2} \\ &= \mathbf{y}_j^T\mathbf{N}_j\mathbf{y}_j - 2\mathbf{y}_j^T\mathbf{N}_j\mathbf{V}_q\mathbf{b}_j + \mathbf{b}_j^T(\mathbf{V}_q^T\mathbf{N}_j\mathbf{V}_q)\mathbf{b}_j \end{aligned}$$

As with the previous derivation process of gappy POD, the stationary point of $\mathbf{b}_j$ can be found after taking a derivative of r_j^2 with respect to $\mathbf{b}_j$ and requiring it to vanish:

$$\left.\frac{\partial r_j^2}{\partial \mathbf{b}_j}\right|_{\mathbf{y}_j,\mathbf{V}_q} = -2(\mathbf{N}_j\mathbf{V}_q)^T\mathbf{y}_j + 2\mathbf{b}_j^T(\mathbf{V}_q^T\mathbf{N}_j\mathbf{V}_q) = 0$$

which reduces to a system of q linear equations such that

$$(\mathbf{V}_q^T\mathbf{N}_j\mathbf{V}_q)\mathbf{b}_j = (\mathbf{N}_j\mathbf{V}_q)^T\mathbf{y}_j$$

and, finally, the optimal coefficient $\mathbf{b}_j$ is determined by

$$\mathbf{b}_j = (\mathbf{V}_q^T\mathbf{N}_j\mathbf{V}_q)^{-1}(\mathbf{N}_j\mathbf{V}_q)^T\mathbf{y}_j \tag{23}$$

which corresponds to b_{ij} in Eq. (9) derived in Sec. II.A. Note that the matrix multiplied by $\mathbf{y}_j$ in Eq. (23), i.e., $(\mathbf{V}_q^T\mathbf{N}_j\mathbf{V}_q)^{-1}(\mathbf{N}_j\mathbf{V}_q)^T$, must change in accordance with $\mathring{\mathbf{y}}_j$ because $\mathbf{N}_j$ is unique to each $\mathring{\mathbf{y}}_j$. As a result, gappy POD requires as many evaluations of $(\mathbf{V}_q^T\mathbf{N}_j\mathbf{V}_q)^{-1}(\mathbf{N}_j\mathbf{V}_q)^T$ as the number of $\mathring{\mathbf{y}}_j$.

B. EM-PCA as an Iterative Optimizer

Unlike gappy POD, which directly tackles a least-squares problem in a deterministic way, the EM-PCA does not address an explicit form of a least-squares problem; instead, the EM-PCA is designed to maximize the expected log-likelihood in Eq. (18) by alternating the E-step and the M-step to find probability parameter estimates. During iterations, the EM-PCA repeats both E-step and M-step in such a way that the E-step computes unknown variables while keeping parameter estimates fixed and, similarly, the subsequent M-step evaluates the parameter estimates while holding the unknown variables constant. Interestingly, this EM-PCA process is equivalent to iteratively solving a least-squares problem, because the EM algorithm belongs to bound optimization methods that are known to carry out fixed-point iterations for optimization [16]. After all, provided that observations are free of measurement errors, the EM-PCA in Eq. (20) is actually identical to minimizing an averaged squared residual R^2, defined as

$$R^2 = \frac{1}{N}\sum_{j=1}^{N}\|\mathbf{y}_j - \mathbf{W}\mathbf{x}_j\|_{L^2}^2 = \frac{1}{N}\,\text{tr}(\|\mathbf{Y} - \mathbf{W}\mathbf{X}\|_{L^2}^2) \tag{24}$$

in a fixed-point iteration fashion. Indeed, the same equations as those of the EM-PCA, listed in Eq. (20), can be achieved after R^2 in Eq. (24) is differentiated with respect to each variable, and then the first derivatives are equated to zero as follows:

$$\left.\frac{\partial R^2}{\partial \mathbf{X}}\right|_{\mathbf{Y},\mathbf{W}} = 0 \Rightarrow \mathbf{X} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{Y}$$

$$\left.\frac{\partial R^2}{\partial \mathbf{Y}}\right|_{\mathbf{X},\mathbf{W}} = 0 \Rightarrow \mathbf{Y} = \mathbf{W}\mathbf{X}$$

$$\left.\frac{\partial R^2}{\partial \mathbf{W}}\right|_{\mathbf{X},\mathbf{Y}} = 0 \Rightarrow \mathbf{W} = \mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}$$

which ascertains that the EM-PCA implicitly minimizes R^2 in Eq. (24) in an iterative manner. Hence, similar to the least-squares problem of gappy POD in Eq. (6), the EM-PCA solves a de facto least-squares problem such that

$$\min \|\mathbf{y}_j - \mathbf{W}\mathbf{x}_j\|_{L^2}^2 \quad \text{subject to } \mathbf{x}_j \tag{25}$$

to restore missing data in an observed variable $\mathbf{y}_j$. For a coefficient $\mathbf{x}_j$, a squared residual r_j^2 is evaluated as

$$\begin{aligned} r_j^2 &= \|\mathbf{y}_j - \mathbf{W}\mathbf{x}_j\|_{L^2}^2 = (\mathbf{y}_j - \mathbf{W}\mathbf{x}_j, \mathbf{y}_j - \mathbf{W}\mathbf{x}_j)_{L^2} \\ &= \mathbf{y}_j^T\mathbf{y}_j - 2\mathbf{y}_j^T\mathbf{W}\mathbf{x}_j + \mathbf{x}_j^T(\mathbf{W}^T\mathbf{W})\mathbf{x}_j \end{aligned}$$

and like the earlier derivation of $\mathbf{b}_j$ for gappy POD, the optimal least-squares coefficient $\mathbf{x}_j$ of the EM-PCA can be found by the following stationary equation:

$$\left.\frac{\partial r_j^2}{\partial \mathbf{x}_j}\right|_{\mathbf{y}_j,\mathbf{W}} = -2(\mathbf{y}_j^T\mathbf{W}) + 2(\mathbf{W}^T\mathbf{W})\mathbf{x}_j = 0$$

which yields a coefficient $\mathbf{x}_j$ as

$$\mathbf{x}_j = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{y}_j \tag{26}$$

which corresponds to $\langle \mathbf{x}_j \rangle$ of the E-step, shown in Eq. (20a). Note that $\mathbf{x}_j$ in Eq. (26), as opposed to $\mathbf{b}_j$ in Eq. (23), has a constant multiplying matrix $(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T$ to $\mathbf{y}_j$, regardless of a gappy snapshot $\mathring{\mathbf{y}}_j$. Thus, the least-squares coefficient evaluation of the EM-PCA is more efficient than that of gappy POD.

C. Generalized Least-Squares Problem

As elucidated previously, gappy POD solves a least-squares problem explicitly, and, interestingly, the EM-PCA does so implicitly. For a methodical comparison of both gappy POD and the EM-PCA, each least-squares problem in Eqs. (6) and (25), on which gappy POD and the EM-PCA hinges, respectively, can be generalized as

$$\min \|\mathbf{y}_j - \mathbf{\Phi}\mathbf{c}_j\|_\alpha^2 \quad \text{subject to } \mathbf{\Phi} \quad \text{and} \quad \mathbf{c}_j \tag{27}$$

where $\mathbf{\Phi}$ is a basis, α is a norm, and $\mathbf{c}_j$ is a coefficient for $\mathbf{\Phi}$. Therefore, in view of a least-squares perspective, the two missing-data reconstruction methods share the generalized least-squares problem formulated in Eq. (27), which requires the evaluation of a basis $\mathbf{\Phi}$ and a least-squares coefficient $\mathbf{c}_j$. Their difference, however, lies in their choices of a basis $\mathbf{\Phi}$ for a subspace projection and a norm α for a squared residual evaluation, which results in a disparate coefficient $\mathbf{c}_j$ that reflects their algorithmic characteristics to address missing-data estimation.

To summarize the similarities and the disparities of gappy POD and the EM-PCA, Table 1 contrasts each step of their formulations. For the similarities, they both address a least-squares problem that reduces to a twofold algorithm: basis and least-squares coefficient evaluations. Despite their processes in common, they differ in each step due to their disparities in a basis $\mathbf{\Phi}$ and a norm α. In detail, to evaluate a basis, gappy POD exploits POD for a POD basis $\mathbf{V}_q$, which is orthogonal, whereas the EM-PCA relies on its M-step for a factor-loading matrix $\mathbf{W}$, which is nonorthogonal. Similarly, to evaluate a least-squares coefficient, due to their norm difference, both gappy

Table 1 Each step of gappy POD and the EM-PCA

	Gappy POD		EM-PCA	
	Evaluation	Value	Evaluation	Value
Initialization	——	$\mathbf{Y}^{(0)}$	——	$\mathbf{Y}^{(0)}$ and $\mathbf{W}^{(0)}$
Basis $\mathbf{\Phi}$	POD (standard or snapshot)	$\mathbf{V}_q$	M-step	$\mathbf{W} = \mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}$
Coefficient $\mathbf{c}_j$	Weighted least-squares problem	$\mathbf{b}_j = (\mathbf{V}_q^T\mathbf{N}_j\mathbf{V}_q)^{-1}(\mathbf{N}_j\mathbf{V}_q)^T\mathbf{y}_j$	E-step	$\mathbf{x}_j = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{y}_j$

POD and the EM-PCA end up with dissimilar least-squares problems: the former being a weighted least-squares problem and the latter being an ordinary least-squares problem. Note that Table 1 conveys that each step of the EM-PCA is more efficient than that of gappy POD for the following reasons:

1) For a basis, POD invokes SVD, which is more expensive than matrix multiplications and inversions in the M-step.

2) For a coefficient, $\mathbf{b}_j$ requires a new evaluation for each $\mathring{\mathbf{y}}_j$, for $\mathbf{N}_j$ is specific to $\mathring{\mathbf{y}}_j$, but $\mathbf{x}_j$ can reuse matrices multiplied to $\mathbf{y}_j$ because they are constant, regardless of any $\mathring{\mathbf{y}}_j$.

As a result, the theoretical dissection of gappy POD and the EM-PCA, recapitulated in Table 1, infers that the EM-PCA is preferable to gappy POD in regard to computational cost.

IV. Numerical Demonstration

For numerical illustration, both gappy POD and the EM-PCA are applied to two exemplary data sets that are artificially marred: simple sine wave data for a preliminary test and airfoil pressure coefficient $\mathbf{C}_p$ data from an Euler computational fluid dynamics (CFD) analysis as a real case. Although the expected log-likelihood $\langle \mathcal{L}_C \rangle$ in Eq. (18) is useful for monitoring the convergence of the EM-PCA, the normalized root-mean-squared residual (RMSR) of an estimated snapshot $\tilde{\mathbf{y}}_j$, defined by

$$\frac{\mathrm{RMSR}^{(k)}}{\mathrm{RMSR}^{(1)}} < 10^{-6} \tag{28}$$

where

$$\mathrm{RMSR}^{(k)} = \sqrt{\frac{1}{\mathrm{d}N}\sum_{j=1}^{N} \|\tilde{\mathbf{y}}_j^{(k)} - \tilde{\mathbf{y}}_j^{(k-1)}\|_{L^2}^2}$$

is employed as a convergence criterion to fairly compare both reconstruction methods. Moreover, the root-mean-squared error (RMSE) of repaired data is measured to quantify estimation errors reduced by the two methods; an RMSE is defined similarly to an RMSR such that it evaluates a differential between the true and estimated values. For the numerical investigation herein, two versions for each gappy POD and the EM-PCA are cross-examined, which results in a total of four implementations:

1) GPOD $\boldsymbol{\mu}$ invariant is a gappy POD that holds a sample mean constant during iterations such that it calculates a sample mean to determine a mean-centered snapshot ensemble beforehand to start iterations. Thus, both the sample mean and the mean-centered snapshot ensemble are kept constant throughout iterations.

2) GPOD $\boldsymbol{\mu}$ variant is a gappy POD that calculates a sample mean as well as a snapshot ensemble at every iteration. Thus, both the sample mean and the mean-centered snapshot ensemble are updated per iteration.

3) EM-PCA `rand` init. is an EM-PCA that takes random initialization for $\mathbf{W}$.

4) EM-PCA $\mathbf{V}_e$ init. is an EM-PCA that initializes $\mathbf{W}$ with a guessed POD basis $\mathbf{V}_e$ that gappy POD uses to initiate its first iteration. Here, $\mathbf{V}_e$ is obtained by applying POD to an estimated snapshot ensemble $\tilde{Y}$ whose missing data are accordingly filled with sample means.

Note that the last two EM-PCA implementations keep a sample mean constant like the GPOD $\boldsymbol{\mu}$ invariant, and the EM-PCA $\mathbf{V}_e$ init. will be used only for numerical performance tests, because it is basically identical to the EM-PCA `rand` init., except the initialization of $\mathbf{W}$. In this paper, all numerical experiments are carried out in an MATLAB R2007b environment, operating on an Intel Pentium dual-core 2.8 GHz processor with 1 GB memory.

A. Sine Wave Data

A collection of simple sinusoidal data is generated using the following equation:

$$\mathbf{y}_j(\mathbf{x}) = \sin(\mathbf{x} + 2\pi/Nj)$$

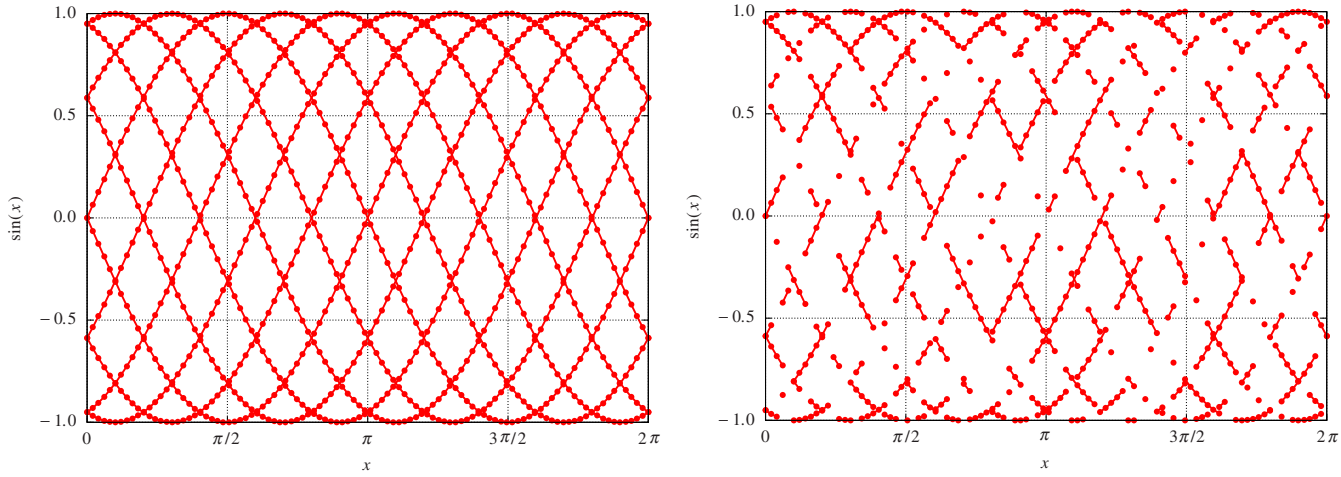
where N is a snapshot ensemble size with successive phase shifts, and then some of the data are randomly thrown away to produce an incomplete data set. This data set is devised as an abstraction to real aerodynamic data characterized by complex irregular waves, and its simplicity is conducive to interpreting the behavior of each algorithm in accordance with the theoretical analysis, presented in Sec. III. As an illustration of the sine data, Fig. 1a depicts the intact 100 by 10 sine wave data set, and Fig. 1b shows the gappy sine wave data set, missing 30.5% of the total data by attempting to arbitrarily exclude 30% of the data along a row direction.

1. Validation

As shown at the bottom of Table 2, snapshot POD [17] is used to evaluate the eigenvalues of the intact data set, which indicate that the first two POD modes are dominant. For the incomplete data set, Table 2 illustrates that the same eigenvalues as those by snapshot POD can be obtained by the four algorithms. In addition to the validation of eigenvalues, Figs. 2a and 2b delineate that all of the tested algorithms can perfectly restore missing data as well as POD modes, respectively. The first two POD modes in Fig. 2b imply that a combination of sine and cosine functions can describe the entire sine wave date set, which is easily anticipated from the well-known Fourier series analysis. Despite excellent validation results, the four algorithms can accurately recover eigenvalues, eigenvectors, and missing data only at a particular q number of modes; otherwise, they exhibit poor reconstruction performance due to underfitting or overfitting. Note that since the $\boldsymbol{\mu}$ variant implementations can more agilely adjust local changes than their $\boldsymbol{\mu}$ invariant counterparts, the GPOD $\boldsymbol{\mu}$ variant requires one less q value such that $q = 2$, unlike the GPOD $\boldsymbol{\mu}$ invariant and the other two EM-PCA implementations, which require $q = 3$, for the most accurate results.

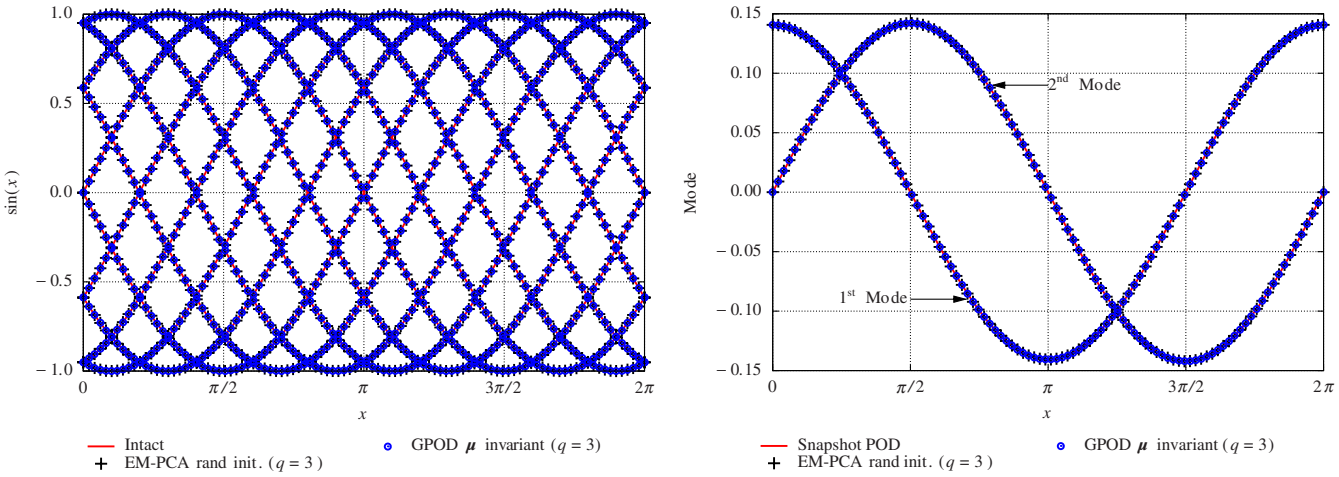
2. Discussion

To reveal the numerical characteristics of each algorithm in detail, Fig. 3 shows the convergence histories, measured by a normalized RMSR and an RMSE in Figs. 3a and 3b, respectively. In Fig. 3a, the two gappy POD implementations display smooth RMSR drops, whereas the two EM-PCA implementations exhibit sluggish convergence behavior at the beginning of iterations. In particular, the EM-PCA `rand` init. shows fluctuating RMSRs in the early iterations due to random initialization, but its RMSR quickly settles down as it marches through iterations. In addition, the RMSR convergence plots in Figs. 3a and 3b delineate the RMSE histories of each reconstruction method to visualize actual error decreases. In Fig. 3b, throughout the iterations, the two gappy POD algorithms outperform their counterparts, i.e., the two EM-PCA implementations, in recovering missing data. The excellence of gappy POD seems to stem from the compound effects of the POD basis and the gappy norm. Between the EM-PCA $\mathbf{V}_e$ init. and the EM-PCA `rand` init.,



a) Complete data b) Incomplete data missing 30.5%

Fig. 1 Complete and incomplete sine wave data.



a) Data b) Modes

Fig. 2 Restored data and modes of the sine wave data missing 30.5%.

the former demonstrates the faster convergence in terms of both normalized RMSR and RMSE due to the informed initialization of $\mathbf{W}$ with $\mathbf{V}_e$.

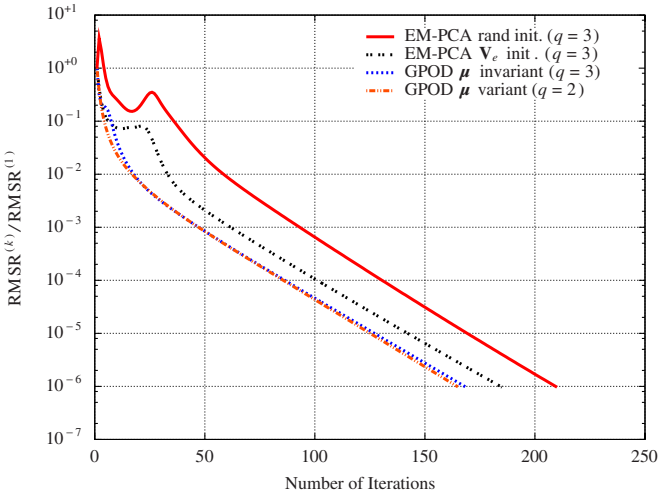
To contrast the numerical performance of both gappy POD and the EM-PCA, Fig. 4 delineates the computational cost of all the tested algorithms measured with MATLAB `tic` and `toc` functions. Note that because of random initialization, the computational time of the EM-PCA `rand` init. is based on the averaged time of 100 runs to mitigate the effect of randomness. To begin with, Fig. 4a plots the total time spent by each tested algorithm along with the total iteration numbers taken under the same convergence criterion in Eq. (28). Figure 4a shows that the EM-PCA implementations are overall more efficient than the gappy POD implementations, despite their higher iteration numbers. In Fig. 4b, total time shown in Fig. 4a is decomposed into time segments spent at each step: basis generation and least-squares coefficients evaluation. Note that each step of the two EM-PCA implementations takes less time than that of gappy POD, especially in the coefficient evaluation step, which explains why the EM-PCA is faster than gappy POD. This noticeable numerical advantage of the EM-PCA over gappy POD is due to the formulations of the EM-PCA being simpler than those of gappy POD for both basis and coefficient evaluations, as analyzed in Sec. III.

B. Euler CFD Analysis Data

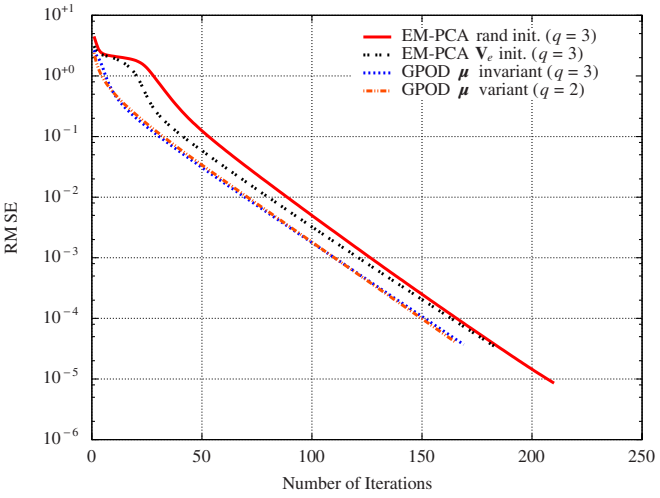
For high-fidelity aerodynamic data generation, a generic numerical compressible airflow solver (GENCAS) [18] is used to

Table 2 Restored eigenvalue spectrum of the sine wave data missing 30.5%

Algorithm	q	Basis initialization	λ_1	λ_2	λ_3
Gappy POD					
μ invariant	3	$\mathbf{V}^{(0)} = \mathbf{V}_e$	5.049993e − 01	4.950007e − 01	4.022614e − 12
μ variant	2	$\mathbf{V}^{(0)} = \mathbf{V}_e$	5.049996e − 01	4.950004e − 01	N/A
EM-PCA					
μ invariant	3	$\mathbf{W}^{(0)} = \mathbf{V}_e$	5.049994e − 01	4.950006e − 01	2.439643e − 12
μ invariant	3	$\mathbf{W}^{(0)} =$ `rand`	5.049998E − 01	4.950002e − 01	1.597123e − 13
Snapshot POD	——	——	5.050000e − 01	4.950000e − 01	7.681359e − 17

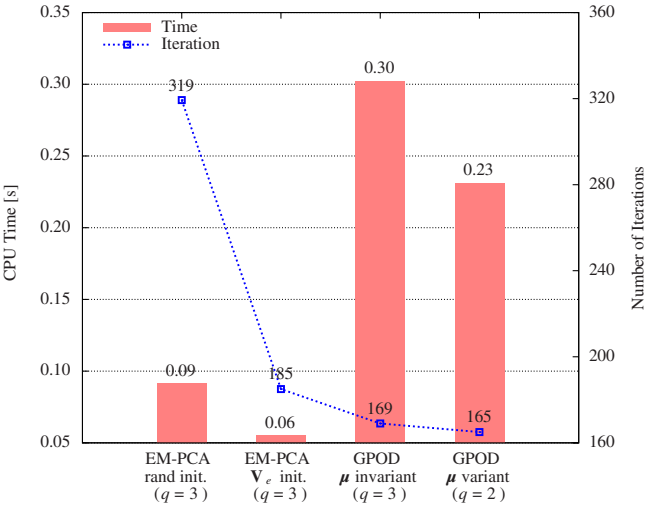


a) RMSR$^{(k)}$/RMSR$^{(1)}$ of $\tilde{Y}$

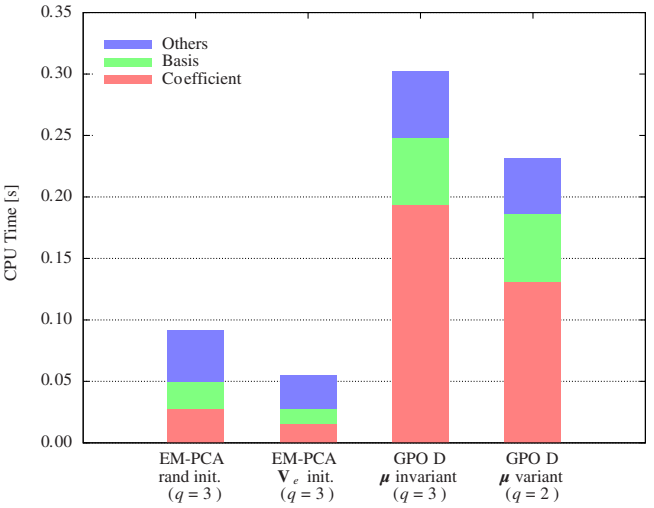


b) RMSE of $\tilde{Y}$

Fig. 3 Convergence histories of the sine wave data missing 30.5%.



a) Total time versus total iteration



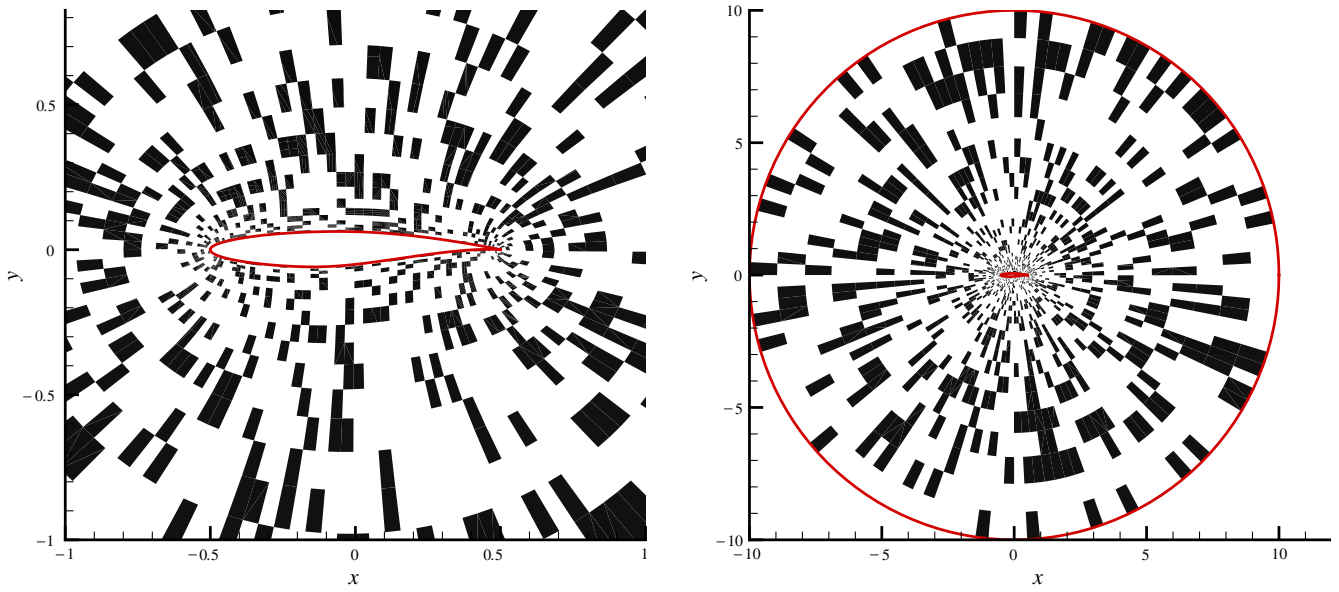
b) Total time breakdown

Fig. 4 Computational time of the sine wave data missing 30.5%.

analyze the flowfield of the RAE 2822 airfoil by solving Euler equations. The airfoil flow analysis with GENCAS employs Roe's flux difference splitting for spatial discretization using second-order MUSCL reconstruction with a minmod limiter and an implicit lower–upper symmetric Gauss–Seidel scheme for local time-marching. As a convergence criterion, the GENCAS uses either a maximum number of iterations or a minimum of the sum of normalized RMSRs of flow variables; the former is set to 50,000 and the latter is set to 10^{-6} for this Euler CFD airfoil analysis. To efficiently produce airfoil analysis snapshots, this research generated 100 samples using a maximum-entropy space-filling design with JMP software by changing two flow parameters: Mach number (0.6–0.8) and angle of attack (1–3 deg). Afterward, steady-state pressure coefficient C_P data are compiled, and then 30% of the snapshots are arbitrarily discarded along a row direction, resulting in a 30%-missing-data set. As an illustration, the first snapshot of the incomplete airfoil C_p data set is depicted in Fig. 5, in which dark cells denote missing-data areas. Note that the gappy airfoil C_p data set is 30% absent with respect to the total number of grid points, not to an entire grid domain; hence, the missing-data percentage does not literally pertain to actual spatial missingness.

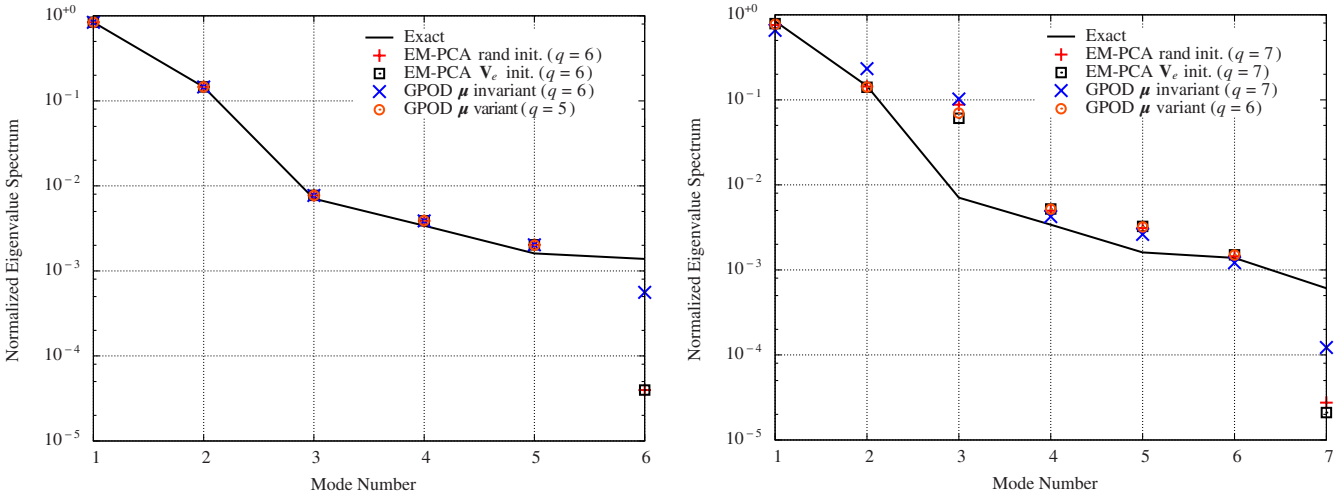
1. *Validation*

Similar to the previous validation process for the sine wave data set, eigenvalues, eigenvectors, and restored missing data of the incomplete airfoil C_p data set are compared with their corresponding true values of the intact airfoil C_p data set. For instance, Fig. 6 delineates true eigenvalues by snapshot POD with estimated eigenvalues by the four tested algorithms at two different q number of modes: $q = 6$ in Fig. 6a and $q = 7$ in Fig. 6a. As shown in Fig. 6a for $q = 6$, dominant eigenvalues estimated by the four tested algorithms, closely match the exact values. In contrast, restored eigenvalues at $q = 7$ in Fig. 6b exhibit deviations from the exact values as q increases to more than 6, leading to degenerate missing-data reconstruction; the RMSE history at $q = 7$ shows that all the implementations struggle for convergence, generating more reconstruction errors rather than reducing them. According to the true eigenvalue spectrum, $q = 6$ accounts for 99.88% of the variations, and $q = 7$ does so for 99.94% of the variations, but adding the



a) Zoom-in **b) Zoom-out**

Fig. 5 First snapshot of the Euler airfoil C_p data missing 30%.



a) $q = 6$ **b) $q = 7$**

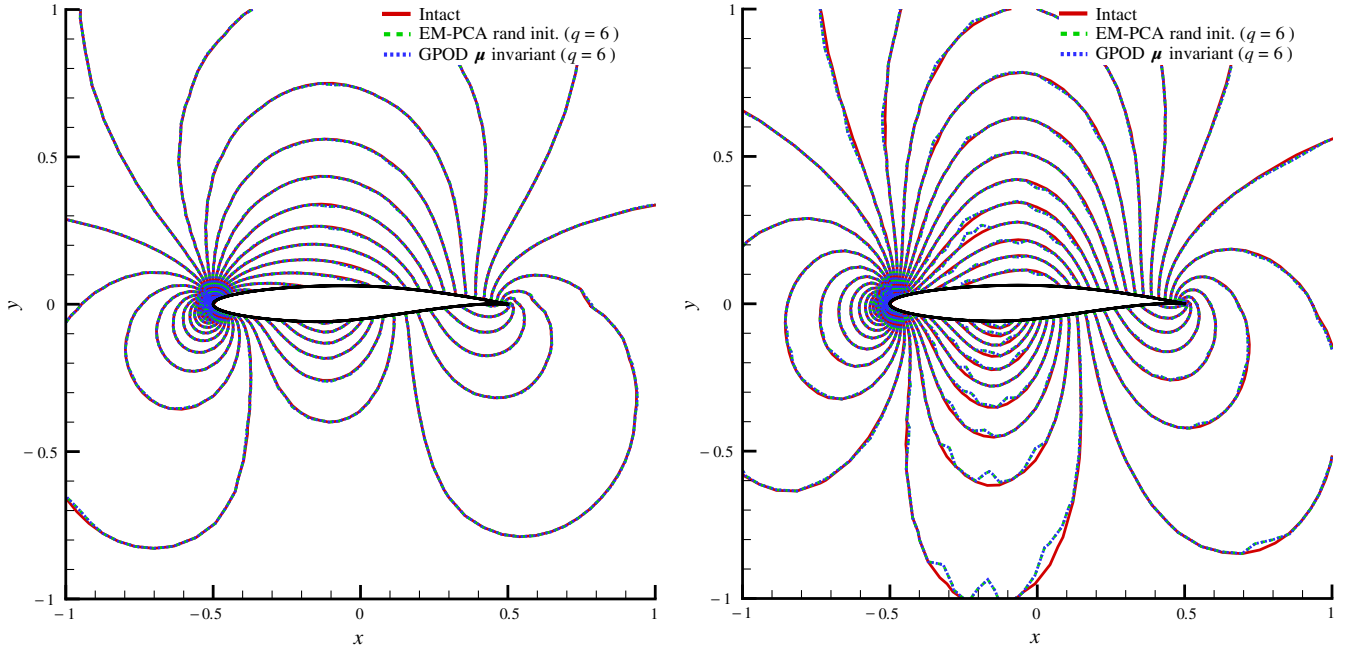
Fig. 6 Eigenvalue spectrum of the restored Euler airfoil C_p data missing 30%.

seventh mode, which has a minuscule 0.06% contribution, noticeably affects the quality of estimation. Therefore, this validation study with the Euler airfoil C_p data hereinafter uses a total of six modes for the four implementations of gappy POD and the EM-PCA. Note that the GPOD $\boldsymbol{\mu}$ variant uses five modes instead of six modes, as it used one less mode in the previous validation study with the sine wave data.

For the validation of the tested implementations, Figs. 7 and 8 delineate estimated data and modes, respectively. In Fig. 7, the contours of the first and the fifth snapshots are compared with those of the restored C_p data by the implementations of gappy POD and the EM-PCA. Since both methods share the identical convergence criterion in Eq. (28), they show the same level of reconstruction quality. For instance, for the first C_p data snapshot in Fig. 7a, which is moderately nonlinear, the two methods are equally satisfactory, and similarly, for the fifth C_p data snapshot in Fig. 7b, they both suffer from higher degrees of nonlinearity. In addition to the restored C_p data, Fig. 8 illustrates the contours of the first two significant modes along with the corresponding modes of reconstructed C_p data for which the relative contribution is 98.54% altogether. Figure 8a shows that the tested implementations estimate the first mode, which delineates 83.97% of the variations, in relatively good agreement with its exact counterpart, and Fig. 8b depicts that the second mode, which accounts for 14.57% of the variations, displays relative disparities between estimation and true values. In general, as a relative modal contribution declines, the quality of an estimated mode also deteriorates, though the estimated mode roughly follows the trend of the original mode.

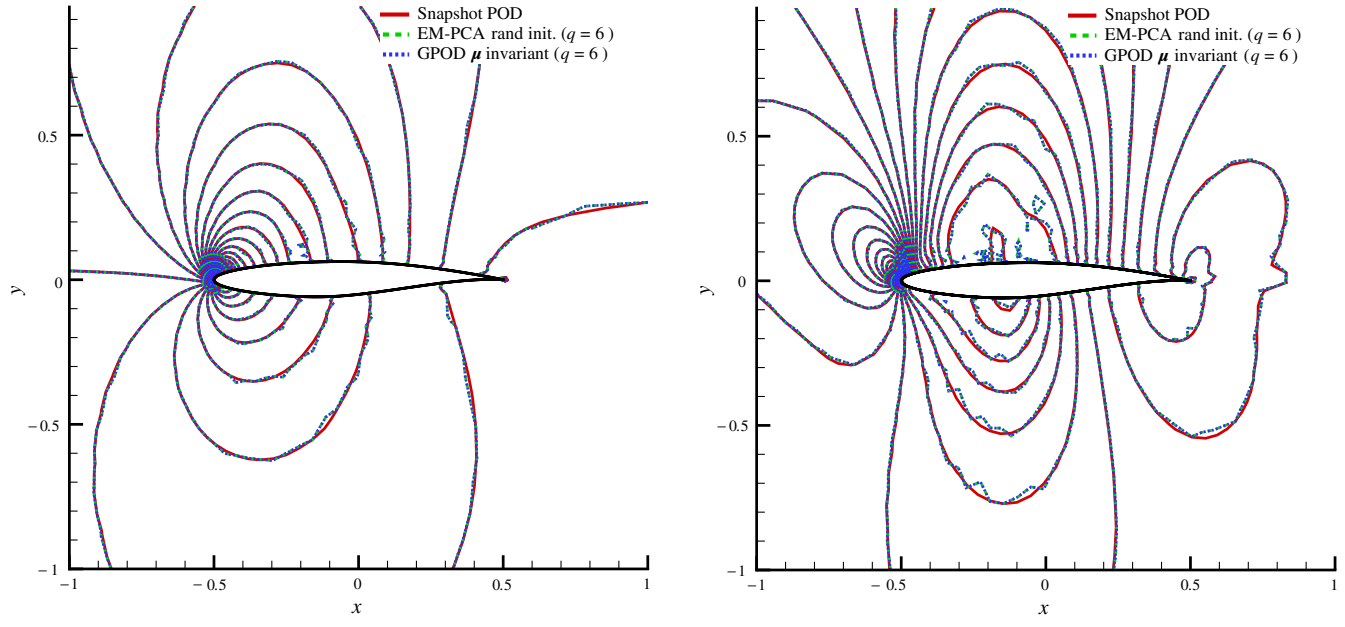
2. *Discussion*

To illustrate the numerical characteristics of each method for the gappy Euler airfoil C_p data, Fig. 9 compares the convergence characteristics of the four tested implementations in terms of a normalized RMSR and an RMSE. Similar to the case of the sine wave data in Fig. 3a, the normalized RMSR convergence histories,



a) 1st snapshot **b) 50th snapshot**

Fig. 7 Contours of the restored Euler airfoil C_p data missing 30%.

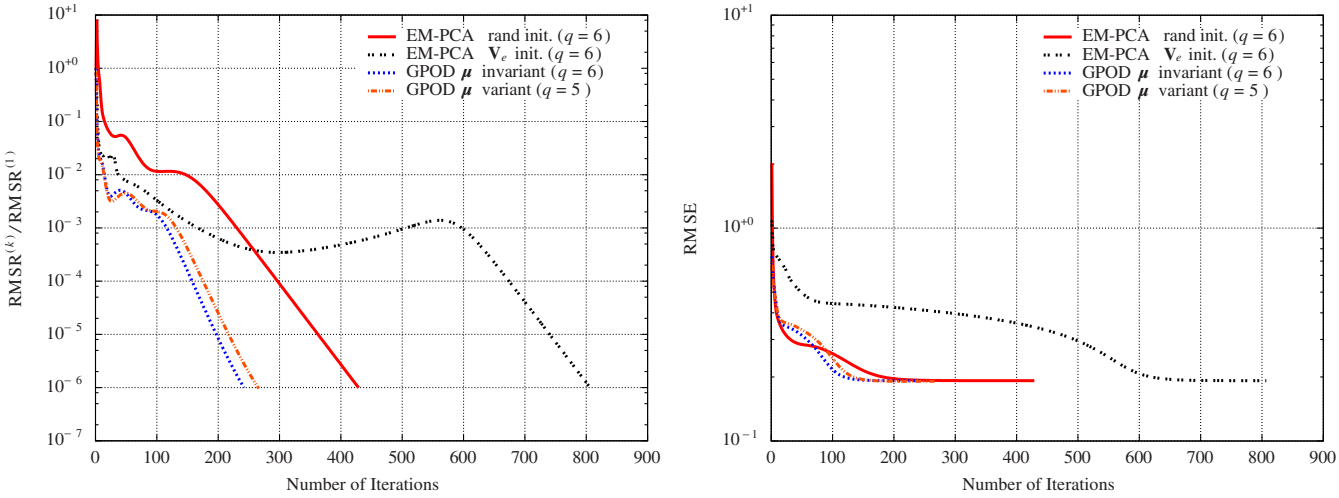


a) 1st mode **b) 2nd mode**

Fig. 8 Mode contours of the restored Euler airfoil C_p data missing 30%.

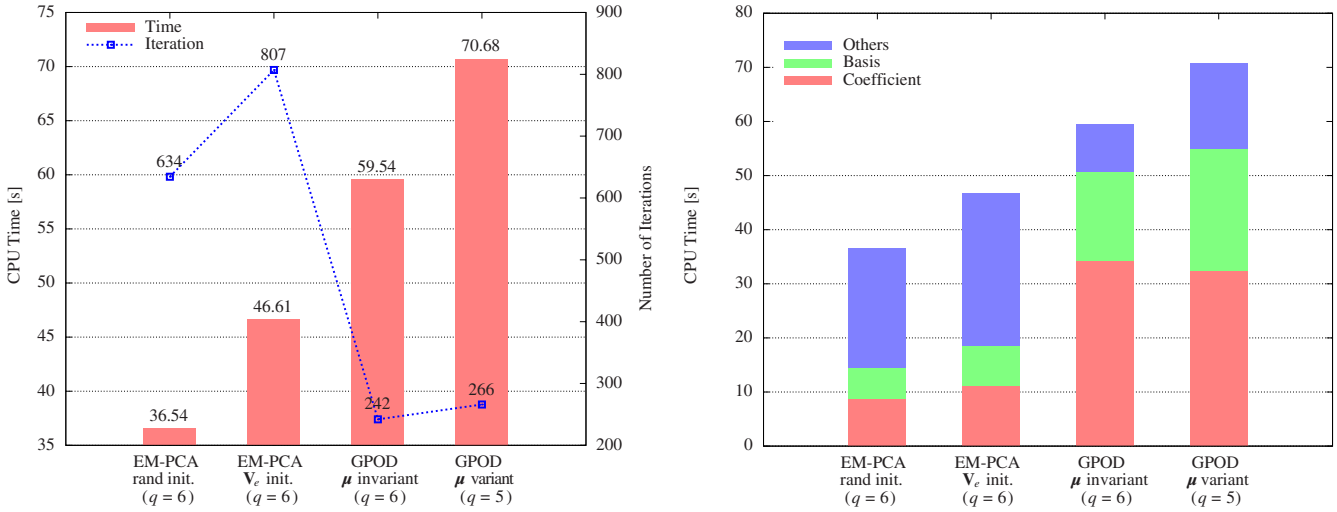
shown in Fig. 9a, convey that both gappy POD implementations converge faster than the two EM-PCA implementations under the same convergence criterion in Eq. (28). In addition, unlike the convergence histories of the sine wave data in Fig. 3a, the slowest turns out to be the EM-PCA $\mathbf{V}_e$ init., which relies on an estimated POD basis $\mathbf{V}_e$ for the better initialization of $\mathbf{W}$. Thus, $\mathbf{V}_e$ is not always a good choice to initialize $\mathbf{W}$, especially for such a data set that is inherently complicated as well as absent of a large portion of its data. In spite of the various convergence performances in Fig. 9a, Fig. 9b delineates that all four algorithms reduce the true estimation errors, measured by the RMSEs, to the same level because of the same convergence criterion. Note that due to higher intrinsic complexity involved in the Euler airfoil C_p data, the RMSE decreases in Fig. 9b are not as substantial as those of the sine wave data, depicted in Fig. 3b.

In addition to the previous investigation on the convergence behavior of the four tested implementations, Fig. 10 delineates their computational performance measured with the Euler CFD airfoil C_p data. In terms of overall iterations, although the two EM-PCA implementations have larger iteration numbers than the two gappy POD implementations, Fig. 10a shows that the former is more efficient than the latter, which is also observed earlier in Fig. 4a for the sine wave data. Note that even the slowest among the tested algorithms, the EM-PCA $\mathbf{V}_e$ init., takes less time than the two fast-



a) RMSR$^{(k)}$/RMSR$^{(1)}$ of $\widetilde{\mathbf{Y}}$ **b) RMSE of $\widetilde{\mathbf{Y}}$**

Fig. 9 Convergence histories of the Euler airfoil C_p data missing 30%.



a) Total time versus total iteration **b) Total time breakdown**

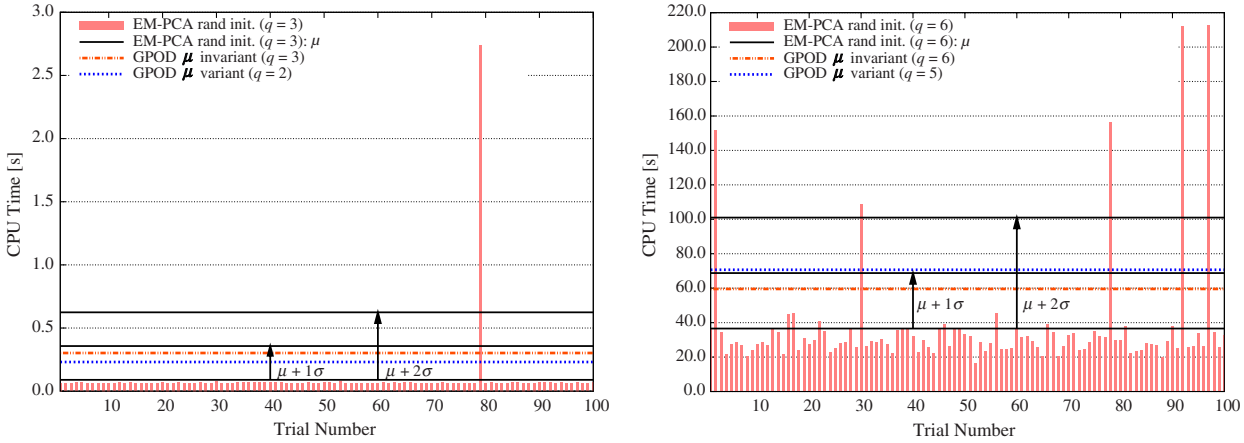
Fig. 10 Computational time of the Euler airfoil C_p data missing 30%.

converging gappy POD algorithms in Fig. 10a. For detailed efficiency analysis, the total time shown in Fig. 10a breaks down into time segments spent at each step of basis and least-squares coefficient evaluation in Fig. 10b. The decomposed total time depicted in Fig. 10b establishes that the two EM-PCA implementations spend less time at both steps, especially at evaluating a least-squares coefficient, which is similar to the previous case in Fig. 4b with the sine wave data. All in all, the computational advantage of the EM-PCA over gappy POD, shown in Fig. 10, again substantiates the predicted numerical performance of both methods based on their formulations in Sec. III.

V. Conclusions

To integrate gappy POD and the EM-PCA within a common formulation framework, this paper reformulates gappy POD in terms of matrix multiplications and construes the EM-PCA as an iterative optimizer performing fixed-point iterations. As a result, gappy POD and EM-PCA address a weighted least-squares problem and an ordinary least-squares problem, respectively, both of which can be easily analyzed from the unifying least-squares perspective. By virtue of the unifying least-squares perspective, the two antithetically originated algorithms can be comprehended as least-squares methods formulated with different bases and norms, highlighting the similarities and the disparities between the two methods. According to the theoretical dissection of both methods, their algorithmic characteristics predetermines their numerical efficiency such that the EM-PCA is more competitive than gappy POD, due to its simpler formulation. The following numerical experiments with the two sample data sets corroborate the deduction about their numerical efficiency based on their theoretical analysis. Despite the computational advantage of the EM-PCA over gappy POD, convergence histories indicate that gappy POD converges faster than the EM-PCA because the formulation of gappy POD is more suited for reducing estimation residuals: the optimal subspace projection by the orthogonal POD basis and a stringent residual evaluation by the gappy norm.

In conclusion, the EM-PCA can effectively address possible applications of gappy POD in aerospace engineering. Moreover, unlike gappy POD, the EM-PCA can account for measurement errors



a) Sine wave data missing 30.5% **b) Euler airfoil C_p data missing 30%**

Fig. A1 Total time variations of the EM-PCA `rand` init.

in restoring experimental flow data because of its intrinsic error model. With regard to accounting for a random effect in data restoration, gappy POD could take advantage of a stochastic POD basis [19], conceivably leading to the formulation of gappy stochastic POD. For future research, further study is warranted to delve into the effects of the different bases and norms on missing-data estimation, since the rudimentary basis and norm differences are crucial factors that distinguish both gappy POD and the EM-PCA. As a final remark on both methods, regardless of whether gappy POD or the EM-PCA is used, numerical tests show that the selection of an optimal number of modes is of paramount importance to the accurate estimation of missing data, which necessitates another future study on determining the number of modes a priori.

Appendix A: Supplementary Investigation on the Computational Efficiency of the EM-PCA

For the thorough numerical performance analysis of the EM-PCA over gappy POD, Fig. A1 illustrates the comparisons between the computational times of the EM-PCA `rand` init. for 100 trials and those of the two gappy POD implementations: the GPOD $\boldsymbol{\mu}$ invariant and the GPOD $\boldsymbol{\mu}$ variant. Moreover, Fig. A1 delineates the ranges of one and two sample standard deviations from the sample mean of the EM-PCA `rand` init., denoted as $\mu + 1\sigma$ and $\mu + 2\sigma$, respectively. Overall, the EM-PCA `rand` init. runs much faster than the two gappy POD implementations; however, because of a few ill-performing cases, the order of magnitude of a sample standard deviation is relatively large. Therefore, the observations in Fig. A1 cannot statistically substantiate that the EM-PCA `rand` init. is computationally more efficient than the two gappy POD implementations. Even though the EM-PCA `rand` init. exhibits performance fluctuations, Figs. 4 and 10 show that the other implementation of the EM-PCA with no random initialization, the EM-PCA $\mathbf{V}_e$ init., still performs computationally better than the two gappy POD implementations. All in all, the EM-PCA $\mathbf{V}_e$ init. is recommended over the EM-PCA `rand` init. for a computational advantage in a conservative sense.

Acknowledgments

The authors would like to thank the two anonymous referees for their valuable comments and suggestions.

References

[1] Everson, R., and Sirovich, L., "Karhunen–Loève Procedure for Gappy Data," *Journal of the Optical Society of America A (Optics, Image Science and Vision)*, Vol. 12, No. 8, Aug. 1995, pp. 1657–1664. doi:10.1364/JOSAA.12.001657

[2] Bui-Thanh, T., Damodaran, M., and Willcox, K., "Aerodynamic Data Reconstruction and Inverse Design Using Proper Orthogonal Decomposition," *AIAA Journal*, Vol. 42, No. 8, Aug. 2004, pp. 1505–1516. doi:10.2514/1.2159

[3] Bui-Thanh, T., "Proper Orthogonal Decomposition Extensions and Their Applications in Steady Aero-Dynamics," M.S. Thesis, Dept. of Aeronautics and Astronautics, Massachusetts Inst. of Technology, Cambridge, MA, 2003.

[4] Willcox, K., "Unsteady Flow Sensing and Estimation Via the Gappy Proper Orthogonal Decomposition," *Computers and Fluids*, Vol. 35, No. 2, Feb. 2006, pp. 208–226. doi:10.1016/j.compfluid.2004.11.006

[5] Robinson, T. D., Eldred, M. S., Willcox, K. E., and Haimes, R., "Strategies for Multifidelity Optimization with Variable Dimensional Hierarchical Models," 47th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, AIAA Paper 2006-1819, 7th, Newport, RI, May 2006.

[6] Venturi, D., and Karniadakis, G. E., "Gappy Data And Reconstruction Procedures for Flow Past a Cylinder," *Journal of Fluid Mechanics*, Vol. 519, 2004, pp. 315–336. doi:10.1017/S0022112004001338

[7] Murray, N. E., and Ukeiley, L. S., "flowfield Dynamics in Open Cavity Flows," 12th AIAA/CEAS Aeroacoustics Conference, AIAA Paper 2006-2428, Cambridge, MA, May 2006.

[8] Murray, N. E., and Ukeiley, L. S., "An Application of Gappy POD: for Subsonic Cavity Flow PIV Data," *Experiments in Fluids*, Vol. 42, No. 1, 2007, pp. 79–91. doi:10.1007/s00348-006-0221-y

[9] Murray, N. E., and Seinery, J. M., "The Effects of Gappy POD on Higher-Order Turbulence Quantities," 46th AIAA Aerospace Sciences Meeting and Exhibit, AIAA Paper 2008-241, Reno, NV, Jan. 2008.

[10] Tipping, M. E., and Bishop, C. M., "Probabilistic Principal Component Analysis," *Journal of the Royal Statistical Society Series B (Methodological)*, Vol. 61, No. 3, 1999, pp. 611–622. doi:10.1111/1467-9868.00196

[11] Lee, K., Rallabhandi, S. K., and Mavris, D. N., "Aerodynamic Data Reconstruction via Probabilistic Principal Component Analysis," 46th AIAA Aerospace Sciences Meeting and Exhibit, AIAA Paper 2008-899, Reno, NV, Jan. 2008.

[12] Shi, J., *Stream of Variation Modeling and Analysis for Multistage Manufacturing Processes*, 1st ed., CRC Press, Boca Raton, FL, Dec. 2006.

[13] Dempster, A. P., Laird, N. M., and Rubin, D. B., "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society Series B (Methodological)*, Vol. 39, No. 1, 1977, pp. 1–38; also available online at http://www.jstor.org/stable/2984875 [retrieved Feb. 2010].

[14] Rubin, D. B., and Thayer, D. T., "EM Algorithms for ML Factor Analysis," *Psychometrika*, Vol. 47, No. 1, March 1982, pp. 69–76. doi:10.1007/BF02293851

[15] Roweis, S., "EM Algorithms for PCA and SPCA," *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, MA, 1998, pp. 626–632.

[16] Huang, H.-S., Yang, B.-H., and Hsu, C.-N., "Triple Jump Acceleration for the EM Algorithm," IEEE International Conference on Data Mining, Inst. of Electrical and Electronics Engineers, Piscataway, NJ, 2005, pp. 649–652.

[17] Sirovich, L., "Turbulence and the Dynamics of Coherent Structures," *Quarterly of Applied Mathematics*, Vol. 45, Oct. 1987, pp. 561–571, 573–590.

[18] Min, B. Y., Lee, W., Englar, R., and Sankar, L. N., "Numerical Investigation of Circulation Control Airfoils," 46th AIAA Aerospace Sciences Meeting and Exhibit, AIAA Paper 2008-0329, Reno, NV, Jan. 2008.

[19] Venturi, D., Wan, X., and Karniadakis, G. E., "Stochastic Low-Dimensional Modelling of a Random Laminar Wake Past a Circular Cylinder," *Journal of Fluid Mechanics*, Vol. 606, No. 1, 2008, pp. 339–367.
doi:10.1017/S0022112008001821

K. Willcox
Associate Editor