Computers and Mathematics with Applications 70 (2015) 2566–2574

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

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

Reduced basis decomposition: A certified and fast lossy data compression algorithm

Yanlai Chen

Department of Mathematics, University of Massachusetts Dartmouth, 285 Old Westport Road, North Dartmouth, MA 02747, USA

ARTICLE INFO

Article history:
Received 20 March 2015
Received in revised form 14 September 2015
Accepted 20 September 2015
Available online 14 October 2015

Keywords:
Data mining
Lossy compression
Reduced basis method
Singular value decomposition
Greedy algorithm

ABSTRACT

Dimension reduction is often needed in the area of data mining. The goal of these methods is to map the given high-dimensional data into a low-dimensional space preserving certain properties of the initial data. There are two kinds of techniques for this purpose. The first, projective methods, builds an explicit linear projection from the high-dimensional space to the low-dimensional one. On the other hand, the nonlinear methods utilizes nonlinear and implicit mapping between the two spaces. In both cases, the methods considered in literature have usually relied on computationally very intensive matrix factorizations, frequently the Singular Value Decomposition (SVD). The computational burden of SVD quickly renders these dimension reduction methods infeasible thanks to the ever-increasing sizes of the practical datasets.

In this paper, we present a new decomposition strategy, Reduced Basis Decomposition (RBD), which is inspired by the Reduced Basis Method (RBM). Given X the high-dimensional data, the method approximates it by $Y\,T (\approx X)$ with Y being the low-dimensional surrogate and T the transformation matrix. Y is obtained through a greedy algorithm thus extremely efficient. In fact, it is significantly faster than SVD with comparable accuracy. T can be computed on the fly. Moreover, unlike many compression algorithms, it easily finds the mapping for an arbitrary "out-of-sample" vector and it comes with an "error indicator" certifying the accuracy of the compression. Numerical results are shown validating these claims.

1. Introduction

Dimension reduction is ubiquitous in many areas ranging from pattern recognition, clustering, classification, to fast numerical simulation of complicated physical phenomena. The fundamental question to address is how to approximate a n-dimensional space by a d-dimensional one with $d \ll n$. Specifically, we are given a set of high-dimensional data

$$X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{m\times n}, \tag{1}$$

and the goal is to find its low-dimensional approximation

$$Y = [y_1, y_2, \ldots, y_d] \in \mathbb{R}^{m\times d} \tag{2}$$

with reasonable accuracy.

There are two types of dimension reduction methods. The first category consists of "projective" ones. These are the linear methods that are *global* in nature, and that explicitly transform the data matrix X into a low-dimensional one by $Y = X\,T$.

E-mail address: yanlai.chen@umassd.edu.
URL: http://www.faculty.umassd.edu/yanlai.chen/.

http://dx.doi.org/10.1016/j.camwa.2015.09.023

The leading examples are the Principal Component Analysis (PCA) and its variants. The methods in the second category act locally and are inherently nonlinear. For each sample in the high-dimensional space (e.g. each column of X), they directly find their low-dimensional approximations by preserving certain locality or affinity between nearby points.

In this paper, inspired by the reduced basis method (RBM), we propose a linear method called "Reduced Basis Decomposition (RBD)". It is much faster than PCA/SVD-based techniques. Moreover, its low-dimensional vectors are equipped with error estimator indicating how close they are approximating the high-dimensional data. RBM is a relative recent approach to speed up the numerical simulation of parametric Partial Differential Equations (PDEs) [1–5]. It utilizes an Offline–Online computational decomposition strategy to produce surrogate solution (of dimension N) in a time that is of orders of magnitude shorter than what is needed by the underlying numerical solver of dimension $\mathcal{N} \gg N$ (called *truth* solver hereafter). The RBM relies on a projection onto a low dimensional space spanned by truth approximations at an optimally sampled set of parameter values [6–10]. This low-dimensional manifold is generated by a greedy algorithm making use of a rigorous *a posteriori* error bounds for the field variable and associated functional outputs of interest which also guarantees the fidelity of the surrogate solution in approximating the truth approximation.

The RBD method acts in a similar fashion. Given the data matrix X as in (1), it iteratively builds up Y (2) whose column space approximates that of X. It starts with a randomly selected column of X (or a user input if existent). At each step where we have k vectors $\{y_1, \ldots, y_k\}$, the next vector y_{k+1} is found by scanning the columns of X and locating the one whose error of projection into the current space $\text{span}\{y_1, \ldots, y_k\}$ is the largest. This process is continued until the maximum projection/compression error is small enough or until the limit on the size of the reduced space is reached. An important feature is an offline–online decomposition that allows the computation of the compression error, and thus the cost of locating y_{k+1}, to be independent of (the potentially large) m.

This paper is organized as follows. In Section 2, we review the background material, mainly the RBM. Section 3 describes the reduced basis decomposition algorithm and discusses its properties. Numerical validations are presented in Section 4, and finally some concluding remarks are offered in Section 5.

2. Background

The reduced basis method was developed for use with finite element methods to numerically solve PDEs. We assume, for simplicity, that the problems (usually parametric PDE) to simulate are written in the weak form: find $u(\mu)$ in an Hilbert space X such that $a(u(\mu), v; \mu) = f(v; \mu),\ \forall v \in X$ where μ is an input parameter. These simulations need to be performed for many values of μ chosen in a given parameter set $\mathcal{D}$. In this problem a and f are bilinear and linear forms, respectively, associated to the PDE (with $a^{\mathcal{N}}$ and $f^{\mathcal{N}}$ denoting their numerical counterparts). We assume that there is a numerical method to solve this problem and the solution $u^{\mathcal{N}}$, called the "truth approximation" or "snapshot", is accurate enough for all $\mu \in \mathcal{D}$.

The fundamental observation utilized by RBM is that the parameter dependent solution $u^{\mathcal{N}}(\mu)$ is not simply an arbitrary member of the infinite-dimensional space associated with the PDE. Instead, the solution manifold $\mathcal{M} = \{u^{\mathcal{N}}(\mu),\ \mu \in \mathcal{D}\}$ can typically be well approximated by a low-dimensional vector space. The idea is then to propose an approximation of $\mathcal{M}$ by $W^N = \text{span}\{u^{\mathcal{N}}(\mu_1),\ \ldots,\ u^{\mathcal{N}}(\mu_N)\}$ where, $u^{\mathcal{N}}(\mu_1),\ \ldots,\ u^{\mathcal{N}}(\mu_N)$ are N ($\ll \mathcal{N}$) pre-computed truth approximations corresponding to the parameters $\{\mu_1, \ldots, \mu_N\}$ judiciously selected according to a sampling strategy [10]. For a given μ, we now solve in W^N for the reduced solution $u^{(N)}(\mu)$. The online computation is $\mathcal{N}$-independent, thanks to the assumption that the (bi)linear forms are affine[1] and the fact that they can be approximated by affine (bi)linear forms when they are nonaffine [11,12]. Hence, the online part is very efficient. In order to be able to "optimally" find the N parameters and to assure the fidelity of the reduced basis solution $u^{(N)}(\mu)$ to approximate the truth solution $u^{\mathcal{N}}(\mu)$, we need an *a posteriori* error estimator $\Delta_N(\mu)$ which involves the residual $r(v, \mu) = f^{\mathcal{N}}(v; \mu) - a^{\mathcal{N}}(u^N(\mu), v; \mu)$ and stability information of the bilinear form [13,14,2,3,15]. With this estimator, we can describe briefly the classical **greedy algorithm** used to find the N parameters $\mu_1, \ldots, \mu_N$ and the space W^N. We first randomly select one parameter value and compute the associated truth approximation. Next, we scan the entire (discrete) parameter space and for each parameter in this space compute its RB approximation $u^{(N=1)}$ and the error estimator $\Delta_1(\mu)$. The next parameter value we select, μ_2, is the one corresponding to the largest error estimator. We then compute the truth approximation and thus have a new basis set consisting of two elements. This process is repeated until the maximum of the error estimators is sufficiently small.

The reduced basis method typically has exponential convergence with respect to the number of pre-computed solutions [16–18]. This means that the number of pre-computed solutions can be small, thus the computational cost reduced significantly, for the reduced basis solution to approximate the finite element solution reasonably well. The author and his collaborators showed [19] that it works well even for a complicated geometric electromagnetic scattering problem that efficiently reveals a very sensitive angle dependence (the object being stealthy with a particular configuration).

3. Reduced basis decomposition

In this section, we detail our proposed methodology by stating the algorithm, studying the error evaluation, and pinpointing the computational cost.

[1] $a(w, v; \mu) \equiv \sum_{q=1}^{Q_a} \Theta_a^q(\mu)\, a^q(w, v),\ \forall\, w, v \in X^{\mathcal{N}},\ f(v; \mu) \equiv \sum_{q=1}^{Q_f} \Theta_f^q(\mu) f^q(v),\ \forall\, v \in X^{\mathcal{N}}$.