Appl. Comput. Harmon. Anal. 36 (2014) 302–315

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Cover-based bounds on the numerical rank of Gaussian kernels

Amit Bermanis [a], Guy Wolf [b], Amir Averbuch [b,*]

[a] Department of Applied Mathematics, School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel
[b] School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

ARTICLE INFO

Article history:
Received 25 July 2012
Received in revised form 20 January 2013
Accepted 27 May 2013
Available online 31 May 2013
Communicated by M.V. Wickerhauser

Keywords:
Spectral analysis
Kernel method
Numerical rank
Gaussian convolution operator

ABSTRACT

A popular approach for analyzing high-dimensional datasets is to perform dimensionality reduction by applying non-parametric affinity kernels. Usually, it is assumed that the represented affinities are related to an underlying low-dimensional manifold from which the data is sampled. This approach works under the assumption that, due to the low-dimensionality of the underlying manifold, the kernel has a low numerical rank. Essentially, this means that the kernel can be represented by a small set of numerically-significant eigenvalues and their corresponding eigenvectors.

We present an upper bound for the numerical rank of Gaussian convolution operators, which are commonly used as kernels by spectral manifold-learning methods. The achieved bound is based on the underlying geometry that is provided by the manifold from which the dataset is assumed to be sampled. The bound can be used to determine the number of significant eigenvalues/eigenvectors that are needed for spectral analysis purposes. Furthermore, the results in this paper provide a relation between the underlying geometry of the manifold (or dataset) and the numerical rank of its Gaussian affinities.

The term cover-based bound is used because the computations of this bound are done by using a finite set of small constant-volume boxes that cover the underlying manifold (or the dataset). We present bounds for finite Gaussian kernel matrices as well as for the continuous Gaussian convolution operator. We explore and demonstrate the relations between the bounds that are achieved for finite and continuous cases. The cover-oriented methodology is also used to provide a relation between the geodesic length of a curve and the numerical rank of Gaussian kernel of datasets that are sampled from it.

1. Introduction

The rapid development of data collection techniques together with high availability of data and storage space introduce increasingly big high-dimensional datasets that fit data analysis tasks. In many cases the quantity of data does not reflect on its quality. Usually, it contains many redundancies that do not add important information over a limited set of representatives. Furthermore, more often than not, the distribution of samples (also called data points) is significantly affected by the sampling techniques that are used. These problems affect both the massive size of the sampled datasets and their high-dimensionality, which in turn prevent classical statistical methods from being effective tools to analyze these datasets due to the "curse of dimensionality" phenomenon.

[*] Corresponding author. Fax: +972 3 6422020.
E-mail address: amir@math.tau.ac.il (A. Averbuch).

1063-5203/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.acha.2013.05.004

Due to the vast number of observable quantities that can be measured/sensed and used as parameters or features, the raw representation of the data is usually high-dimensional. Recent dimensionality reduction methods use manifolds to cope with this problem. Under this manifold existence assumption, a dataset is assumed to be sampled from an Euclidean submanifold that has a relatively small intrinsic dimension. The ambient high-dimensional Euclidean space of the manifold is defined by the raw parameters (or features) of the dataset. These parameters are mapped via non-linear functions to low-dimensional coordinates of the manifold, which represent the independent factors that control the behaviors of the analyzed phenomenon.

Several methods have been suggested to provide a low-dimensional representation of data points by preserving the intrinsic structure of their underlying manifold. Kernel methods such as k-PCA [13,17], LLE [16], Isomaps [19], Laplacian Eigenmaps [2], Hessian Eigenmaps [9], Local Tangent Space Alignment [22,23] and Diffusion Maps [5] have been used for this task. These methods extend the classical PCA [11,10] and MDS [8,12] methods that project the data on a low-dimensional hyperplane that preserves most of the variance in the dataset. Kernel methods substitute the linear relations (i.e., inner-products) that are preserved by PCA and MDS with a kernel construction that introduces the synonymous notion of similarity, proximity, or affinity between data points. Spectral analysis of this kernel is used to obtain an embedding of the data points into an Euclidean space while preserving the kernel's qualities, which are based on non-linear local qualities of the underlying manifold.

Beside the high-dimensionality of the data, its size (i.e., number of sampled data points) is usually very big. The massive size of the dataset is mostly due to the ease of obtaining data points. For example, most systems nowadays collect detailed logs of every action, event and operation that occur with high frequency over long periods of time. However, most of the collected data points are redundant, either because they are near-duplicates of other already-measured data points, or because their properties can be interpolated by suitable subsets of representatives. Therefore, a combination of subsampling and out-of-sample extension techniques can alleviate performance issues that massive datasets entail, and provide a more suitable representation of the analyzed data. Optimally, such a representation would not be affected by the availability of the data or by a sampling method but only rely on the behavior of the observed and analyzed phenomena.

The kernel approach, which is used for dimensionality reduction, has been applied for the described out-of-sample extension tasks. A classical kernel-based technique is the Nyström extension [14,1]. More recent methods are Geometric Harmonics [6] and the Multiscale Extension in [3]. These methods use the spectral decomposition of the kernel (i.e., its eigenvalues and eigenvectors) as a basis of its range. The eigenfunctions are shown to be easily extended to new data points, thus any function in its range, which can be expressed as a linear combination of these eigenfunctions, is also easily extended. Functions that are not in the range of the kernel are extended by projecting them on the kernel's range and using the resulting function (and extension) as an approximation of the original function.

Kernel methods work under the assumption that the used kernel has a small set of significant eigenvalues that should be considered for the analysis, and the rest are negligible in the sense that they are numerically zero. This can be phrased as a low numerical rank assumption, where the numerical rank is the number of numerically nonzero eigenvalues or singular values (see Definition 2.1 for an explicit formulation). While in practice this assumption is usually satisfied, most papers do not present rigorous mathematical support (beyond intuition) for it.

In this paper, we present upper bounds for the numerical rank of affinity kernels. We focus on Gaussian kernels, which are popular in many spectral kernel methods (e.g. [5,2]). Such an upper bound was achieved in [3] based on a bounding box volume of the analyzed dataset in the observable ambient space. We refine this bound by considering the underlying geometry that is provided by the underlying manifold from which the dataset is assumed to be sampled. Instead of using a single large bounding box, we use a finite set of small constant-volume boxes that cover the dataset (or its underlying manifold), and use the minimal cover to provide a cover-based bound. When the constant size of the boxes is large enough to cover the whole dataset with one box, this bound converges to the one in [3]. Thus, it is at least as tight as this already established one.

The paper has the following structure. The problem setup and a previously-established bound are described in Section 2. The refined cover-based bounds are established in Section 3. Section 4 demonstrates various nuances and concepts of cover-based bounds, as well as their theoretical application for proving relations between the geodesic length of curves and the numerical rank of datasets that are sampled from these curves.

2. Problem setup

Let $\mathcal{M}$ be a low-dimensional compact manifold that lies in the high-dimensional ambient space $\mathbb{R}^m$ that has an Euclidean metric $\|\cdot\|$. In addition, let β be the Borel σ-algebra on $\mathcal{M}$ and let μ be a probability measure on $(\mathcal{M},\beta)$. Finally, let $M\subseteq\mathbb{R}^m$ be a set of n data points (i.e., $n=|M|$) sampled from the manifold $\mathcal{M}$.

Define the affinity between two data points $x,y\in M$ to be $g_\varepsilon(x,y)=e^{-\|x-y\|^2/\varepsilon}$ where ε is a positive parameter. Let G_ε^M be an $n\times n$ affinity kernel between the data points in M, where each row and each column of G_ε^M corresponds to a single data point in the dataset M, and each cell contains the affinity $g_\varepsilon(x,y)$ between the row's data point $x\in M$ and the column's data point $y\in M$.

The matrix G_ε^M is called the Gaussian kernel over the dataset M. This kernel introduces the notion of affinities and local neighborhoods of data points in the dataset M (or on the manifold $\mathcal{M}$) due to the exponential decay of it's values in