Multi-View Kernel-based Data Analysis

Amir Averbuch and Moshe Salhov  
School of Computer Science  
Tel Aviv University, IL

Ofir Lindenbaum  
School of Engineering  
Tel Aviv University, IL

Yoel Shkolnisky  
School of Mathematical Science  
Tel Aviv University, IL

Avi Silberschatz  
Department of Computer Science  
Yale University, New Haven, CT 06520-8285

Abstract—The input data features set for many data driven tasks is high-dimensional while the intrinsic dimension of the data is low. Data analysis methods aim to uncover the underlying low dimensional structure imposed by the low dimensional hidden parameters by utilizing distance metrics that considers the set of attributes as a single monolithic set. However, the transformation of the low dimensional phenomena into the measured high dimensional observations might distort the distance metric. This distortion can affect the desired estimated low dimensional geometric structure. In this paper, we suggest to utilize the redundancy in the feature domain by partitioning the features into multiple subsets that are called views. The proposed method utilize the agreement also called consensus between different views to extract valuable geometric information that unifies multiple views about the intrinsic relationships among several different observations. This unification enhances the information that a single view or a simple concatenations of views provides.

I. INTRODUCTION

Kernel methods constitute a wide class of algorithms for non-parametric data analysis of massive high dimensional data. Typically, a limited set of underlying factors generates high dimensional observable parameters via a non-linear mapping.

Multi-Dimensional scaling (MDS) [6], [10] has become a basis for many kernel methods. MDS is used to analyze and visualize data when only pairwise similarities or dissimilarities between data points are observable. The given similarity information is depicted as pairwise distances in a low-dimensional space.

Kernel methods are based on an affinity construction that encapsulates the relations (distances, similarities or correlations) among multidimensional data points. Spectral analysis of this kernel provides an efficient representation of the data that enables its analysis. The non-parametric nature of these methods enables us to uncover hidden structures in the data.

Methods such as Isomap [16], LLE [14], Laplacian eigenmaps [1], Hessian eigenmaps [8] and local tangent space alignment [18], [19] extend the MDS paradigm by considering the manifold assumption. Under this assumption, the data is assumed to be sampled from an intrinsic low dimensional manifold that captures the dependencies between observable parameters. The corresponding spectral-based embedded spaces computed by these methods identify the geometry of the manifold that incorporates the underlying factors in the data.

Similarity assessments between members in datasets is a crucial task for the analysis of any dataset. Important and popular kernel methods such as the methods discussed above utilize similarity metrics that is based on the $l_2$ norm between features. For example, the widely used Gaussian kernel is based on a scaled $l_2$ norm between multidimensional data points. In many cases, the given dataset includes redundant features that relate to the underlying factors via an unknown transformation.

Since the unknown transformation function may have zero derivatives, similarity between transformed data points can be a distorted version of the similarity between the underlying factors. Furthermore, the transformation may take place in the presence noise, which adds additional distortion to the similarity assessment. The utilization of this distorted similarity for kernel based data analysis may fail to uncover the desired geometry for the analysis.

In this paper, we propose a method that aims to compensate for the distortions induced by transformation to similarity assessment. The suggested method considers subsets of features, where each subset is defined as a view. They are used for computing similarity assessment for each view and the agreement between all the computed similarities. Furthermore, we utilize this agreement to estimate the corresponding inaccessible pairwise similarities of the underlying factors.

Learning from several views has motivated various studies that have focused on classification and clustering that are based on the spectral characteristics of multiple datasets. Among these studies are Bilinear Model [9] and Canonical Correlation Analysis [4]. These methods are effective for clustering but neither provide a low dimensional geometry nor a structure for each view. An approach similar to Canonical Correlation Analysis [4], which seeks a linear transformation that maximizes the correlation among the views, is described in [2]. Data modeling by a bipartite graph is given in [7]. Then, based on the ‘minimum-disagreement’ algorithm, [7] partitions the dataset. Recently, a few kernel-based methods have proposed a model of co-regularizing kernels in both views [11]. It is achieved by searching for an orthogonal transformation, which maximizes the diagonal terms of the kernel matrices obtained.
II. PROBLEM FORMULATION

In the following, \( |||| \) denotes the standard Euclidean vector norm and \( ||| \) denotes the Frobenius matrix norm. Vectors are denoted by **bold** letters and vector components are denoted by a superscript \( '[i] \).

Let \( M \) be a low-dimensional manifold that lies in the high-dimensional ambient Euclidean space \( \mathbb{R}^m \) and let \( d \ll m \) be its intrinsic dimension. Let \( M \subset M \) be a dataset of \( |M| = n \) multidimensional data points that were sampled from \( M \). For data analysis tasks in general and signal processing tasks in particular, each extracted/measured feature vector \( x_i \in M \) is assumed to have a corresponding vector \( \theta_i \in \mathbb{R}^d \) of inaccessible controlling parameters whose \( r \)th component is \( \theta_{ir} \), \( i = 1, \ldots, n \).

Kernel methods analyze datasets such as \( M \) by exploiting its geometry. For \( M \) whose data points were sampled as in [5], [12], [14], [16]. The computed kernel describes a measure of data points pair-wise similarity. The Euclidean norm-based similarity metric between two data points \( x_i, x_j \in M \) is given by

\[
K_e(x_i, x_j) = h \left( \frac{||x_i - x_j||^2}{\varepsilon} \right), \quad i, j = 1, \ldots, n \quad (II.1)
\]

where \( \varepsilon \) is the kernel width and \( h: \mathbb{R} \to \mathbb{R} \) is a function designed such that the kernel matrix is symmetric and positive semi-definite. However, instead of using the measured features for the similarity assessment in Eq. II.1, we want to replace the Euclidean norm \( ||x_i - x_j|| \) with the Euclidean norm \( ||\theta_i - \theta_j|| \). \( \theta_i, \theta_j \in \mathbb{R}^d \) that corresponds directly to the similarity between the inaccessible corresponding controlling parameters instances.

In this paper, we approximate the Euclidean distance \( ||\theta_i - \theta_j|| \) (or the corresponding Mahalanobis distance) by utilizing the redundancy in the feature space. In order to quantify the relation between the approximated distance and the actual one, we introduce the notion of a View, which is given in Definition II.1.

**Definition II.1 (A View).** Let \( I_l \) be a subset of \( m_l \) feature indexes that is selected from the set of \( m \) features given in dataset \( M \). \( M_l \) is a view of \( M \) if for every \( x_i \in M, 1 \leq i \leq n \) the vector \( \tilde{x}_{i,l} \in M_l \) is the subset of \( x_i \) that corresponds to the set of indexes in \( I_l \).

Under the multi-view formulation, by selecting \( \zeta \) subsets of features from the features of \( M \) we generate \( \xi \) views of \( M, 1 \leq l \leq \zeta \). Let \( x_i \) be the concatenation of the \( \xi \) views such that \( x_i = \cup_{l=1}^{\zeta} \tilde{x}_{i,l} \) where \( \tilde{x}_{i,l} \triangleq [x_{i,l,1}, \ldots, x_{i,l,m_{l,l}}] \). We further assume that the \( l \)th view is the outcome of the function \( f_l: \mathbb{R}^d \times \mathbb{R}^{k_l} \to \mathbb{R}^{m_l} \) such that \( \tilde{x}_{i,l} = f_l(\theta_i) \).

Our goal in this work is to find a function \( G : \mathbb{R}^\zeta \to [0,1] \) such that the desired kernel similarity \( K_e(\theta_i, \theta_j), i, j = 1, \ldots, n \) is approximated such that

\[
K_e(\theta_i, \theta_j) \approx G(K_{\xi_1}(\tilde{x}_{i,1}, \tilde{x}_{j,1}), \ldots, K_{\xi_{\zeta}}(\tilde{x}_{i,\zeta}, \tilde{x}_{j,\zeta})) \quad (II.2)
\]

where \( G \) is adapted to the characteristics of the \( l \)th view and \( K_e \) is defined in Eq. II.1. We call this kernel the multi-view kernel.

We aim to find the underlying intrinsic geometry of \( M \). By identifying the consensus between the \( \xi \) views, we are able to approximate a kernel that is related to the inaccessible controlling parameters \( \theta_i \).

III. DATASET WITH AN ACCESSIBLE COVARIANCE MATRIX

In this section, we propose to generate multi-view kernel \( K_e \) and a function \( G \) from Eq. II.2 for a dataset \( M \) that has an accessible covariance matrix at each data point in \( M \). The covariance matrix is computed locally. An example of such a dataset is a time series dataset. Furthermore, we assume that the given dataset enables us to compute the covariance matrix. Our goal is to approximate the kernel affinities between data points based on multiple views of the transformed intrinsic space. Following Section II, we assume that the \( l \)th view is the outcome of an almost everywhere differentiable function \( f_l: \mathbb{R}^d \times \mathbb{R}^{k_l} \to \mathbb{R}^{m_l} \) such that \( \tilde{x}_{i,l} = f_l(\theta_i, \psi_{l,i}) \). Hence, a data point of each view is strictly a function of the intrinsic parameters in the consensus \( f_l: \mathbb{R}^d \to \mathbb{R}^{m_l} \) and \( \tilde{x}_{i,l} = f_l(\theta_i) \).

The dimension \( m_l \) of the \( l \)-view is assumed to be higher than the dimension \( d \) of the intrinsic parametric space. Furthermore, the covariance matrix rank is equal to the intrinsic dimension \( d \). Therefore, we use the Moore-Penrose pseudoinverse to compute the Mahalanobis distance for \( i, j = 1, \ldots, n, l = 1, \ldots, \zeta \), as

\[
d_m(\tilde{x}_{i,l}, \tilde{x}_{j,l}) = \frac{1}{2} (\tilde{x}_{i,l} - \tilde{x}_{j,l})^T A_{l}^{\dagger}(\tilde{x}_{i,l} - \tilde{x}_{j,l}) \quad (III.1)
\]

where

\[
A_{l}^{\dagger} \triangleq C_{\tilde{x}_{i,l}} + C_{\tilde{x}_{j,l}}.
\]

The Mahalanobis distance enables us to compare data points in the intrinsic space by comparing data points in the ambient space as Lemma III.1 suggests.
Lemma III.1. Let \( f \) be a bi-Lipschitz transformation such that \( X_i = f_i(\theta_i) \) and \( X_j = f_j(\theta_j) \) two data points from \( M_i \). Then,
\[
d_m(X_i, X_j) = d_m(\theta_i, \theta_j) + O(\phi^2),
\] (III.2)
where \( \phi \triangleq ||\theta_i - \theta_j|| \).

The estimation of the intrinsic dimensionality \( d \) of a dataset has gained considerable importance recently. Proposed estimation methods utilize local distances and angles [3]. In this section we utilize local PCA to estimate \( d \) [15]. This method is more naturally integrated in the proposed algorithm, Algorithm III.1.

Algorithm III.1 approximates the intrinsic dimension \( d \) and provides an improved affinity measure that can be used for various kernel-based methods such as DM to reveal the underlined manifold. The algorithm identify deviation from the required assumptions by estimating the rank of the relevant covariances, which are used for the distance computation.

Algorithm III.1: Multi-view affinity measure approximation

**Input:** Data points: \( x_1, \ldots, x_n \in \mathbb{R}^m \) divided into \( \zeta \) views and associated covariances \( C_{\tilde{X}_i, l} \) \( 1 \leq i \leq n \), \( 1 \leq l \leq \zeta \), \( \varepsilon \) and threshold \( \gamma \).

**Output:** Approximated \( K_r(\theta_i, \theta_j) \).

1. Compute \( \kappa(i, l) \) as the number of singular values of \( C_{\tilde{X}_i, l} \) that are larger than \( \gamma \).
2. Compute \( \kappa_m \) as the median of \( \kappa(i, l) \), \( 1 \leq i \leq n \), \( 1 \leq l \leq \zeta \).
3. For \( 1 \leq i, j \leq n, 1 \leq l \leq \zeta \), such that \( \kappa(i, l) \geq \kappa_m \) and \( \kappa(j, l) \geq \kappa_m \) do
4. Calculate \( d_m(\tilde{X}_{i, l}, \tilde{X}_{j, l}) \) using Eq. III.1.
5. Set \( K_r(\tilde{X}_{i, l}, \tilde{X}_{j, l}) = \exp(-d_m(\tilde{X}_{i, l}, \tilde{X}_{j, l})/\varepsilon) \).
6. Set \( K_r(\theta_i, \theta_j) = G(K_r(\tilde{X}_{1, l}, \tilde{X}_{1, l}), \ldots, K_r(\tilde{X}_{\zeta, l}, \tilde{X}_{j, l})) \) where \( G() \) is the maximum operation.
7. End for

IV. EXPERIMENTAL RESULTS

This section describes two examples that demonstrate how the multi-view approaches are used. The presented example describes a dataset embedding that consists of several views with an accessible covariance matrix. In this case, we show the advantage of the multi-view-based embedding from Section III. It is compared to the embedding of the corresponding single-view dataset that its features are the concatenation of all the feature subsets from all the views.

\( \zeta = 10 \) views are generated to evaluate the performance of Algorithm III.1. All the views are 3-dimensional that are based on one underlying angular parameter denoted by \( \theta_i \in \mathbb{R}, i = 1, \ldots, n \). The 10 views are generated by the application of the following function
\[
\tilde{x}_{i, l} = \begin{bmatrix} \tilde{x}_{1, l}^{1} \\ \tilde{x}_{2, l}^{1} \\ \tilde{x}_{3, l}^{1} \end{bmatrix} = \begin{bmatrix} f_1^2(\theta_i) \\ f_2^2(\theta_i) \\ f_3^2(\theta_i) \end{bmatrix} = \\
\begin{bmatrix} (4/3) \cdot \cos(\theta_i + Z_1^l) - (1/3) \cdot \cos(4(\theta_i + Z_2^l)) \\ (4/3) \cdot \sin(\theta_i + Z_1^l) - (1/3) \cdot \sin(4(\theta_i + Z_2^l)) \end{bmatrix} \sin(0.8 \cdot \mod((\theta_i + Z_3^l), 2\pi))
\] (IV.1)
where \( Z_1^l, Z_2^l, Z_3^l, 1 \leq l \leq 10 \), are random variables drawn from a uniform distribution on the interval \([0, 2\pi]\). As demonstrated in Fig. IV.1, each view is a deformation of an open flower-shaped manifold. A similar deformation can occur in various real-life applications where the measured data is the output from some non-linear phenomena. In this experiment, we demonstrate the ability of a multi-view approach to overcome such deformations.

![Fig. IV.1. The three manifolds generated from the use of Eq. IV.1.](image)

To extract the underlying parameter \( \theta_i \), we first compute the DM for each view. The two leading coordinates of the extracted embedding are denoted by \( \phi_1 \) and \( \phi_2 \). They are presented in Fig. IV.2. All the extracted manifolds are horseshoe-shapes that consist of a large gap created by the deformation (Eq. IV.1) in the third coordinate of the sampled data.

Next, the 10 available views are concatenated to a single view \( x_i = [\tilde{x}_{i, 1}, \ldots, \tilde{x}_{i, \zeta}] \), \( i = 1, \ldots, n \). The Mahalanobis distance is computed using Eq. III.1 and DM is applied to
the resulted kernel. The first two leading coordinates of the kernel-based DM are presented in Fig. IV.2. The large gap in each of the extracted manifolds is a deformation caused by the third coordinate of the transformation functions $f_l$ (Eq. IV.1). Furthermore, the output from the application of DM to the concatenation of the 10 views is presented in Fig. IV.3(a). The embedded manifold is even more distorted as the gaps in the embedding suggests. Hence, the standard procedure concatenating the given set of features without considering the specific distortion each subset of features may introduce into the embedding might amplify the distortion in the resulted embedding.

Finally, we apply Algorithm III.1 to all 10 views and compute the two leading DM coordinates. The outputs are presented in Fig. IV.3(b). The algorithm overcomes the deformations from all the gaps by considering only the non-deformed small distances as the outcome of the rank restriction and minimization procedure in Algorithm III.1. The result is the circle shaped manifold that completely agrees with the corresponding intrinsic controlling angle parameter $\theta_i$.

V. Conclusions

This paper presents a kernel construction scheme that designs kernels by approximating the similarity between intrinsic parameters that are common to multiple subset of features in the presence of non-linear transformation. The presented method utilizes the relation between the Jacobian of each view and the corresponding Mahalanobis distance when a local covariance can be approximated. This relation enables the approximation of the affinities between intrinsic controlling parameters by considering the Mahalanobis distance of each view. The constructed kernel can further be normalized and decomposed to find an embedding of the data. In order to demonstrate the effectiveness of the proposed method, we analyzed a synthetic dataset. This analysis showed that the correct affinities can be approximated in the presence of an unknown non-linear transformation. Furthermore, in cases where the features are the output of a transformed intrinsic parameters with an associated non-full rank Jacobian then the concatenation of the entire set of features results in a deformed manifold. In this case, the proposed multi-view scheme overcomes the problematic Jacobian and out-
puts a non-deformed manifold. The proposed methodology involves a single spectral decomposition while increasing the number of smaller Covariance-based Mahalanobis distances computations per affinity. Hence, the growth in computation complexity is negligible.

ACKNOWLEDGMENT

This research was partially supported by the Israeli Ministry of Science & Technology (Grants No. 3-9096, 3-10898), US-Israel Binational Science Foundation (BSF 2012282), Blavatnik Computer Science Research Fund and Blavatnik ICRC Funds.

REFERENCES