

Variational Graph Embedding for Globally and Locally Consistent Feature Extraction

Shuang-Hong Yang^{1,3}, Hongyuan Zha¹, S. Kevin Zhou², and Bao-Gang Hu³

¹ College of Computing, Georgia Institute of Technology, USA

² Integrate Data Systems, Siemens Corporate Research, USA

³ NLPR & LIAMA, Chinese Academy of Sciences, China

shyang@gatech.edu, zha@cc.gatech.edu

shaohua.zhou@siemens.com, hubg@nlpr.ia.ac.cn

Abstract. Existing feature extraction methods explore either global statistical or local geometric information underlying the data. In this paper, we propose a general framework to learn features that account for both types of information based on variational optimization of non-parametric learning criteria. Using mutual information and Bayes error rate as example criteria, we show that high-quality features can be learned from a variational graph embedding procedure, which is solved through an iterative EM-style algorithm where the E-Step learns a variational affinity graph and the M-Step in turn embeds this graph by spectral analysis. The resulting feature learner has several appealing properties such as *maximum discrimination*, *maximum-relevance-minimum-redundancy* and *locality-preserving*. Experiments on benchmark face recognition data sets confirm the effectiveness of our proposed algorithms.

1 Introduction

Feature extraction, the preprocessing step aimed at learning a small set of highly predictive features out of a large amount of possibly noisy or redundant raw input variables, plays a fundamental role in the success of many learning tasks where high dimensionality arises as a big challenge [8,5].

Over the past decades, a large number of algorithms have been proposed, mostly based on using either *global statistical* or *local geometric* structures underlying the data. Classical techniques, such as the Principal Component Analysis (PCA) and the Fisher Discriminant Analysis (FDA), make use of some well-defined statistical measures (e.g., variance, entropy, linear correlation, cross-correlogram, Fisher information, etc.) to evaluate the usefulness of features. Since these measures are usually defined based on the overall properties of the data set, hence, such statistical approaches are often powerful to retain the global structures of the data space, but usually perform poorly when their underlying assumptions (i.e., the optimal condition of statistical measures) are violated. For example, FDA performs poorly for multi-modal data, because Fisher's criterion is optimal only if the data in each class are sampled from a Gaussian distribution[7].

In contrast, a host of algorithms were recently established by using the local geometric information to restore the submanifold structure from which the data are sampled. Examples include ISOMAP [21], Locally Linear Embedding [17], Laplacian Eigenmap [1], Locality Preserving Projection [9]. Such geometric methods, although are powerful to retain the geometric structure revealed by the training data, neglect the overall properties of the data that are fundamental to the success of extracting predictive features. For example, most geometric methods neglect the supervision (label) information of the data and therefore lack of discriminant power.

In this paper, we show that, by optimizing certain nonparametric learning criteria, both the global (statistical) and local (geometric) structures revealed by the training data can be used to build high-quality features. As case studies, we mainly consider two learning criteria, Mutual Information (MI) and Bayes Error Rate (BER). Both MI and BER are theoretically optimal for feature learning but computationally intractable in practice. Our approach, however, is able to encode these criteria into well-defined data graphs and in turn reduce the task into variational graph-embedding problem. The proposed approach is an iterative EM-style algorithm where the E-Step learns a variational affinity graph and the M-Step in turn embeds this graph by spectral analysis. More importantly, the learned graphs are capable to simultaneously capture the supervision information, the global statistical property and the local geometric structure of the data, leading to feature extractors sharing the advantages of both local geometric methods and global statistical methods while mitigating their drawbacks.

1.1 Related Work

MI is a popular criterion in machine learning. Unlike other statistical measures, such as variance, correlation or Fisher's criterion, which only account for up to second order moments, MI can capture the complete dependence between features and the target concept (i.e., label) [14,16]. A practical prohibition of using MI is the computational cost of entropy estimation which involves numerical integration of high dimensional data. Conventional approaches usually resort to histogram or discretization based methods to obtain estimations of MI [2,26], which is computationally intensive when we are dealing with high-dimensional data. In contrast, our approach encodes maximization of MI as variational graph embedding, a much easier problem to solve. In addition, our approach is significantly different from the existing MI-based feature extraction algorithms such as [12,22,10], all of which are formalized as nonlinear non-convex optimization problems and do not account for the local properties of the data, as a result, cannot capture the geometric structure of the underlying manifold, which is, however, fundamental to feature learning as being demonstrated by recent researches [21,1].

BER has also been employed to learn features. However, due to the unavailability of the underlying generative distribution, almost all the existing

algorithms optimize BER indirectly [18,4], e.g., by maximizing the *Average Pairwise Divergence* or minimizing the *Union Bhattacharyya Error Bounds* or *Bhattacharyya Distance*. And Gaussian assumption usually has to be made to make these approach tractable, which strongly limits the applicability and performance of those methods. Again, our approach shows advantages over this class of methods in computational efficiency as well as the capability to use local information to restore geometric structures.

The importance of utilizing both global and local information for better feature learning has been increasingly recognized. Algorithms that are globally-and-locally consistent were reported to significantly outperform both global algorithms and local algorithms. For example, Weinberger et al [23] proposed an approach to learn a kernel matrix for nonlinear feature projections by maximizing a global statistic measure (i.e. variance of the projected data) subject to local geometric constraints (e.g., preserving the angles and distances between nearest neighbors). Sugiyama [20] proposed a localized FDA algorithm by optimizing a combination of the *Fisher's criterion* [7] and the *locality preserving cost* [9]. In this paper, we propose a principled way to derive such globally-and-locally consistent approaches for feature learning.

The last few years have witnessed a surge of interests in graph-based learning (GBL). Typically, a GBL learner is established by: (1) conveying basic assumptions and heuristical intuitions into pairwise similarity of and/or constraints over the training instances; (2) constructing a affinity graph based on the defined similarity; and (3) building a learner by spectral analysis of the graph. For instance, in dimensionality reduction, Yan et al [24] and Zhao & Liu [29] presented generalized formalization frameworks for graph-based feature extraction and attribute selection respectively. Usually, the spectral graph theory [6] is employed to provide justifications for GBL. However, it provides no guidance on how to construct the graph, which is of central importance to the success of the GBL algorithms since the performance is extremely sensitive to both graph structures and edge weight settings. As a consequence, one has to resort to heuristics to establish graph for GBL. In contrast, we show in this paper that affinity graphs can be learned by optimizing theoretically sound learning measures. In particular, we show that some nonparametric criteria lead to graphs which naturally encode both the global statistical and the local geometric structures of the data.

1.2 Our Contribution

The main contribution of this paper are three folds:

- Firstly, we propose two effective feature extraction algorithms and test them on real-world tasks.
- Secondly, the graphs learned by our method, which encodes both global (statistical) and local (geometric) structures of the data, can be used in a wide variety of graph-based learning tasks, e.g., semi-supervised learning, metric learning, etc.
- Finally, the approach we use to learn graphs, that is, *nonparametric learning measure estimation* and *variational approximation of kernel terms*, can

be applied to other learning criteria to learn predictive graphs/kernels/similarity functions.

2 Optimal Feature Learning Criteria

Suppose we are given a set of input vectors $\{\mathbf{x}_n\}_{n=1}^N$ along with the corresponding labels $\{y_n\}_{n=1}^N$ drawn *i.i.d* from an unknown distribution $p(\mathbf{x}, y)$, where $\mathbf{x}_n \in \mathcal{X} \subset \mathbb{R}^D$ is a training instance and $y_n \in \mathcal{Y} = \{1, \dots, C\}$ is its label, N , D and C denote the training set size, the input space dimensionality and the total number of categories, respectively. The goal of feature extraction is to construct a set of M ($M \ll D$) most predictive features, i.e., to find a preprocessing of data $\mathbf{z} = \tau(\mathbf{x})$, where $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^M$, $\tau: \mathcal{X} \rightarrow \mathcal{Z}$ and $\tau \in \mathcal{H}$ with \mathcal{H} being a hypothesis space. Let the goodness of τ be measured by the feature evaluation criterion $\mathfrak{J}(\cdot)$. Then the problem of feature extraction can be formalized as:

$$\tau = \arg \max_{\tau} \mathfrak{J}(\tau).$$

Theoretically, two optimal criteria can be identified for feature learning. The first one [13] is based on information theory, which attempts to minimize the amount of information loss incurred in the process of dimensionality reduction, i.e.: $\min_{\tau \in \mathcal{H}} KL\{p(y|\mathbf{z})||p(y|\mathbf{x})\}$. Ideally, if the KL-divergence between these two posteriors reaches zero, we can recover the optimal Bayes classifier based on the data in the reduced dimensional space \mathcal{Z} . This criterion is equivalent to maximizing the mutual information, i.e., the expected information gain about y from observing \mathbf{z} :

$$\max_{\tau \in \mathcal{H}} I(\mathbf{z}, y) = \int_{\mathbf{z}, y} p(\mathbf{z}, y) \log \frac{p(\mathbf{z}, y)}{p(\mathbf{z})p(y)} d\mathbf{z}dy, \quad (1)$$

The second optimal criterion [25] considers classification directly and naturally reflects the Bayes error rate \mathcal{Y} in the reduced dimensional space \mathcal{Z} , i.e.:

$$\min_{\tau \in \mathcal{H}} \mathcal{Y}(\tau) = \inf_h E_{\mathbf{x}}[err(h|\mathbf{z})] = E_{\mathbf{z}}[1 - \max_{c=1, \dots, C} P(c|\mathbf{z})], \quad (2)$$

where $E_{\mathbf{x}}\{err(h|\mathbf{z})\}$ is the generalization error of a decision rule h in the reduced-dimensional space.

However, a critical issue of learning features by directly optimizing such optimal learning criteria is that they involve unknown generative distributions. In this paper, we attempt to establish feature extractors by efficiently optimizing these criteria. We achieve this goal by two steps: (1) By nonparametric estimation of the criteria, we reduce the task to kernel-based optimization problems; (2) Based on variational approximation of each kernel term, the task can be compactly formalized as variational graph-embedding problems, which can be solved by graph spectral analysis.

We will mainly focus on MI in the next section and discuss BER in Section 4 since the derivation procedures are quite similar.

3 Graph-Based Feature Learning by Maximizing Mutual Information

In this section, we establish feature extractors by maximize the mutual information between the features and the class label. Compared with other criteria that only account for up to second-order statistics, an appealing property of MI is that it makes use of higher-order statistics and is able to capture the complete nonlinear dependence between the features and the class label [14,16]. However, the computation of MI involves integration of the unknown generative distributions. Conventional approaches usually resort to histogram or discretization based methods to obtain estimations of MI [2,26]. Such approaches are not only highly ill-formed but also computationally intractable when dealing with extremely high-dimensional data. In this section, we will show that MaxMI-features can be learned efficiently by variational graph embedding.

3.1 Nonparametric Quadratic MI

We use a nonparametric method to estimate MI. One of the advantages of using nonparametric estimators is that they can effectively capture the properties (e.g., multimodality) of the underlying distribution. We adopt an efficient estimator that was proposed by Principe et al [16] and exploited recently by Torkkola [22] to learn features. First, instead of using standard MI, we use the quadratic MI:

$$I_2(\mathbf{z}, y) = \int_{\mathbf{z}, y} (p(\mathbf{z}, y) - p(\mathbf{z})p(y))^2 d\mathbf{z}dy, \quad (3)$$

which has been justified both theoretically and experimentally by many previous works, e.g., [16,22,10].

A kernel density estimator is then employed to estimate the density function involved in Eq.(3). Particularly, consider the isotropic Gaussian kernel:

$$k(\mathbf{x}, \mathbf{x}') = g(\mathbf{x} - \mathbf{x}', \sigma^2 U),$$

where σ is the standard deviation, U denotes the unit matrix, and $g(\boldsymbol{\mu}, \Sigma)$ is the Gaussian distribution function with mean $\boldsymbol{\mu}$ and covariance Σ . An interesting property of this kernel is:

$$\langle k(\mathbf{x}, \mathbf{x}'), k(\mathbf{x}, \mathbf{x}'') \rangle = \int_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}')k(\mathbf{x}, \mathbf{x}'')d\mathbf{x} = k(\mathbf{x}', \mathbf{x}'').$$

It turns out that, by using this property and the quadratic form of Eq.(3), we are able to eliminate the integration in MI.

Given the training data, $p(\mathbf{x})$ can be estimated as $\hat{p}(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N k(\mathbf{x}, \mathbf{x}_n)$. Plugging this into the objective $\mathfrak{J}(\mathbf{z}) = I_2(\mathbf{z}, y)$, we obtain,

$$\begin{aligned}
\mathfrak{J}(\mathbf{z}) &= \hat{I}_2(\mathbf{z}, y) = \mathfrak{J}_{\mathbf{z}y} + \mathfrak{J}_{\mathbf{z}} \times \mathfrak{J}_y, \\
\mathfrak{J}_{\mathbf{z}y} &= \frac{1}{N^2} \sum_{c=1}^C \sum_{i \in \mathcal{Y}_c} \sum_{j \in \mathcal{Y}_c} (\sum k(\mathbf{z}_i, \mathbf{z}_j) - \frac{2N_c}{N} \sum_{n=1}^N k(\mathbf{z}_i, \mathbf{z}_n)), \\
\mathfrak{J}_{\mathbf{z}} &= \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N k(\mathbf{z}_i, \mathbf{z}_j), \\
\mathfrak{J}_y &= \sum_{c=1}^C \frac{N_c^2}{N^2},
\end{aligned} \tag{4}$$

where \mathcal{Y}_c denotes the index set of examples in class c and N_c the number of instances in it, $\sum_{c=1}^C N_c = N$.

3.2 MaxMI by Variational Graph Embedding

In this section, we show that maximization of MI can be formalized compactly as a variational graph embedding problem with a generic graph learning procedure being derived. We begin with rewriting the objective Eq.(4).

Theorem 1. *Maximizing the nonparametric quadratic MI is equivalent to the following optimization problem:*

$$\begin{aligned}
\max \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} k(\mathbf{z}_i, \mathbf{z}_j), \\
\gamma_{ij} = I(y_i = y_j) + \sum_{c=1}^C \frac{N_c^2}{N^2} - \hat{P}_{y_i} - \hat{P}_{y_j},
\end{aligned} \tag{5}$$

where $I(\cdot)$ denotes the indicator function, and $\hat{P}_{y_i} = \hat{P}(c = y_i) = \frac{N_c}{N}$ is the proportion of instances sharing the same label with y_i .

Proof. Let $I_{ic} = I(y_i = c)$, $\sum_c I_{ic} I_{jc} = I_{ij} = I(y_i = y_j)$, we have:

$$\begin{aligned}
\mathfrak{J}(\mathbf{z}) &= \mathfrak{J}_{\mathbf{z}y} + \mathfrak{J}_{\mathbf{z}} \times \mathfrak{J}_y \\
&= \frac{1}{N^2} \sum_{c=1}^C \sum_{i=1}^N \sum_{j=1}^N (I_{ic} I_{jc} - I_{ic} \hat{P}_{y_i} - I_{jc} \hat{P}_{y_j}) k(\mathbf{z}_i, \mathbf{z}_j) + \frac{1}{N^2} \sum_{c=1}^C \frac{N_c^2}{N^2} \sum_{i=1}^N \sum_{j=1}^N k(\mathbf{z}_i, \mathbf{z}_j) \\
&= \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (I_{ij} + \sum_c \frac{N_c^2}{N^2} - \hat{P}_{y_i} - \hat{P}_{y_j}) k(\mathbf{z}_i, \mathbf{z}_j) \\
&= \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} k(\mathbf{z}_i, \mathbf{z}_j). \quad \square
\end{aligned}$$

The optimization in Eq.(5) is still computationally difficult due to its expression as a big summation of the nonconvex Gaussian density function $k(\cdot, \cdot)$. To address this problem, we adopt the variational optimization method [11] to approximate each kernel term with its variational lower bound. Since exponential function is convex, a simple lower bound can be easily obtained by making first-order Taylor expansion:

$$\exp\left(\frac{\|\mathbf{z}_i - \mathbf{z}_j\|^2}{-\delta^2}\right) \geq \lambda_{ij} \|\mathbf{z}_i - \mathbf{z}_j\|^2 / \delta^2 - \lambda_{ij} + \lambda_{ij} \log(-\lambda_{ij}), \quad (6)$$

where we have introduced variational parameters λ_{ij} . Then for given λ_{ij} , integrate Eq.(6) into the objective Eq.(5), the task is reduced to a linear graph embedding problem [24]:

$$\min \sum_{i=1}^N \sum_{j=1}^N w_{ij} \|\mathbf{z}_i - \mathbf{z}_j\|^2, \quad (7)$$

where $W = (w_{ij}) \in \mathbb{R}^{N \times N}$ is the adjacency matrix of the derived data graph with edge wights $w_{ij} = -\gamma_{ij} \lambda_{ij}$.

However, since the variational parameters are coupled with the feature extractors, the variational optimization is a EM-style algorithm, where the E-step corresponds to learning a graph W by optimizing the variational parameters in Eq.(6), and the M-step in turn learns feature extractors by solving the graph embedding problem Eq.(7). These two steps are repeated alternatively until convergence.

The E-Step has an analytical solution because the variational lower bound Eq.(6) is exact iff $\lambda_{ij} = -\exp(-\|\mathbf{z}_i - \mathbf{z}_j\|^2 / \delta^2)$. In the following, we will mainly discuss the M-Step.

Let $\mathcal{G} = \{\mathbf{X}, W\}$ be the undirected weighted graph with $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ being the vertex set and W being the adjacency matrix. $D = \text{diag}[W\mathbf{1}]$ denotes the degree matrix of \mathcal{G} , $\mathbf{1} = [1, 1, \dots, 1]^\top$, $L = D - W$ denote the Laplacian matrix of \mathcal{G} . The M-Step, i.e., graph-embedding, learns a set of features by maximizing their consistency with the graph \mathcal{G} :

$$\min_{\tau \in \mathcal{H}} \sum_{i=1}^N \sum_{j=1}^N w_{ij} \|\mathbf{z}_i - \mathbf{z}_j\|^2 = Z^\top LZ, \quad (8)$$

where $Z = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n]^\top$. To obtain practical feature extractors, we consider two special forms.

Linear Case. Assume each feature is obtained as a linear projection of the input attributes, i.e., $\mathbf{z} = T^\top \mathbf{x}$, $T = [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_M] \in \mathbb{R}^{D \times M}$ is a transformation matrix, we have

$$\begin{aligned} \min & \sum_{i=1}^N \sum_{j=1}^N w_{ij} \|T^\top \mathbf{x}_i - T^\top \mathbf{x}_j\|^2 = \text{tr}(T^\top X^\top L X T), \\ \text{s.t.} & : \mathbf{t}_m^\top \mathbf{t}_m = 1, \quad \forall m \in \{1, 2, \dots, M\}, \end{aligned} \quad (9)$$

where $X = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^\top$ is the input data matrix, $\text{tr}(\cdot)$ denotes the trace of a matrix. This is a constrained quadratic program. By using the Lagrangian technique, we can easily prove that the optimal projection vectors are given by the eigenvectors of $X^\top L X$ corresponding to the bottom M eigenvalues¹. Since W is symmetric, the resulting features are naturally orthogonal projections, i.e., $T^\top T = U$. The algorithm is referred to as Mutual Information Embedding (MIE).

Nonlinear Case. The linear MIE might fail to discover nonlinear structures underlying the data. We now investigate nonlinear feature extractors. For simplicity, we only consider a special case, where the hypothesis space \mathcal{H} is restricted to a reproducing kernel Hilbert space (RKHS) induced by a Mercer kernel $k(\mathbf{x}, \mathbf{x}')$, where $k(\cdot, \cdot)$ is a symmetric positive semi-definite function, $K = (k_{ij}) \in \mathbb{R}^{N \times N}$ is the kernel matrix, $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ denotes its entry, $i, j = 1, 2, \dots, N$. Based on the property of the Mercer kernel, we can assume that the nonlinearly projected features lie in the space spanned by the kernel bases, i.e., $K \times \boldsymbol{\alpha}$, where $\boldsymbol{\alpha}_i = [\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iN}]^\top$ is a projection vector in the kernel space. We have:

$$\begin{aligned} A^* &= \arg \min \text{tr}(A^\top K^\top L K A), \\ \text{s.t.} & : \boldsymbol{\alpha}_m^\top K \boldsymbol{\alpha}_m = 1, \quad m = 1, 2, \dots, M, \end{aligned} \quad (10)$$

where $A = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_M]$ is the kernel space projection matrix. Similarly, the optimal projection vectors are given by the bottom M eigenvectors of $K^\top L K$. Note that the Euclidean distance in \mathcal{H} is given by:

$$d_{\mathcal{H}}(\mathbf{x}, \mathbf{x}') = \sqrt{k(\mathbf{x}, \mathbf{x}) + k(\mathbf{x}', \mathbf{x}') - 2k(\mathbf{x}, \mathbf{x}')}. \quad (11)$$

For a new input instance \mathbf{x} , its projection in the optimal reduced kernel space is given as:

$$\begin{aligned} \mathbf{z} &= \tau(\mathbf{x}) = A^\top \boldsymbol{\kappa}, \\ \boldsymbol{\kappa} &= [k(\mathbf{x}_1, \mathbf{x}), k(\mathbf{x}_2, \mathbf{x}), \dots, k(\mathbf{x}_N, \mathbf{x})]^\top. \end{aligned} \quad (12)$$

This algorithm is referred to as kernel MIE (kMIE).

3.3 Initial Graph and Efficient Approximate Solution

The variational graph embedding algorithm requires initialization of the variational parameters λ_{ij} or equivalently the graph w_{ij} . In this section, we construct

¹ For fast implementation, please refer to [3,24,19] and the references therein.

an initial graph, which is in the near neighborhood of the optimal one and hence makes the convergence of the variational algorithm very fast.

Again, we approximate each kernel term by its first-order Taylor expansion. Since expanding $\exp(-v)$ at v_0 leads to $\exp(-v) \approx \exp(-v_0) - \exp(-v_0)(v - v_0)$, let $v = \frac{\|\mathbf{z}_i - \mathbf{z}_j\|^2}{\delta^2}$, and $v_0 = \frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\delta^2}$, we have:

$$\exp\left(\frac{\|\mathbf{z}_i - \mathbf{z}_j\|^2}{-\delta^2}\right) \approx -\exp\left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{-\delta^2}\right) \|\mathbf{z}_i - \mathbf{z}_j\|^2 / \delta^2 + \text{const.}$$

Integrating this into the objective, we get an initial data graph with edge weights $w_{ij} = \gamma_{ij} e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / \delta^2}$, where $\delta^2 = 2\sigma^2$. That is, the feature space distance $\|\mathbf{z}_i - \mathbf{z}_j\|$ is replaced with the original space distance $\|\mathbf{x}_i - \mathbf{x}_j\|$.

Besides being used to initialize the graph of the variational embedding algorithm, this graph could also be used as an off-the-shelf tool to build other graph-based learners. For instance, solely embedding this graph leads to an efficient non-iterative alternative to MIE (refer to as initial MIE or MIE₀).

3.4 Justifications

3.4.1 Max-Relevance-Min-Redundancy

In feature extraction, our goal is to learn a set of features that are *useful* for building the predictor. Hence, merely extracting features that are most *relevant* to the target concept is suboptimal since the learned features might be *redundant* such that adding them will gain no additional information [8]. Therefore, a good feature extractor should achieve a reasonable balance between maximizing relevance and minimizing redundancy.

It turns out that our proposed algorithm simultaneously (1) maximizes the relevance of the learned features \mathbf{z} to the class label y ; (2) minimizes the redundancy within \mathbf{z} ; and (3) employs a natural tradeoff parameter for perfect balance. This can be seen clearly from the objective function Eq.(4), which consists of two terms: $\mathfrak{J}(\mathbf{z}) = \mathfrak{J}_{\mathbf{z}y} + \eta \mathfrak{J}_{\mathbf{z}}$, where $\mathfrak{J}_{\mathbf{z}y}$, depending on both the features \mathbf{z} and the class label y , is a measure of relevance of \mathbf{z} to y ; $\mathfrak{J}_{\mathbf{z}}$, depending solely on features \mathbf{z} , measures the degree of redundancy within \mathbf{z} ; and $\eta = \mathfrak{J}_y = \sum_c N_c^2 / N^2$, which is invariant to \mathbf{z} , provides a natural compromise between relevance and redundancy.

3.4.2 Max-Discrimination

Another observation is that the proposed algorithms are supervised methods, i.e., they take advantage of the label information to maximize the discriminative ability of the learned features.

Theorem 2. *The weights $\gamma_{i,j}$ have the following properties:*

1. for $\forall i, j : y_i = y_j, \gamma_{i,j} > 0$;
2. for $\forall i, j : y_i \neq y_j, E\{\gamma_{i,j}\} < 0$.

Proof. Without loss of generality, consider a nontrivial case, i.e., $\forall c = 1, \dots, C$, $C \geq 2$, $P_c > 0$ and $N_c \geq 1$.

1. if $y_i = y_j$:

$$\gamma_{ij} = 1 + \frac{1}{N^2} \sum_{c=1}^C N_c^2 - 2 \frac{N_i}{N} = \frac{1}{N^2} \left(\sum_{c \neq i} N_c^2 + \left(\sum_{k \neq i} N_k \right)^2 \right) > 0,$$

2. if $y_i \neq y_j$:

$$E\{\gamma_{ij}\} = E\left\{ \sum_{c=1}^C \frac{N_c^2}{N^2} - \frac{N_i}{N} - \frac{N_j}{N} \right\} = - \sum_{i,j=1}^C (P_i - P_j)^2 - (C-1) \sum_{i=1}^C P_i^2 < 0,$$

completing the proof. \square

From Theorem 2, we can see that instance pairs that are from the same class ($y_i = y_j$) are always assigned positive weights, while pairs from different classes ($y_i \neq y_j$) are expected to get negative weights. As a consequence, the optimization Eq.(7) actually minimizes the distances between patterns with the same label while keeping patterns with different labels as far apart as possible, i.e., maximizing the margin between within-class and between-class patterns.

3.4.3 Locality Preserving

Another appealing properties of the proposed algorithm is that, besides the statistical information used to model the global property underlying the data (e.g., relevance, redundancy and discrimination), it also makes use of local geometric information to restore the submanifold structure from which the data is sampled, or in other words, it is locality preserving. Particularly, the learned graph \mathcal{G} naturally includes a weight term $e^{-\|\mathbf{z}_i - \mathbf{z}_j\|^2 / \delta^2}$ (in the initial graph: $e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / \delta^2}$), which is actually the Gaussian heat weight [19] usually employed to retain the local consistency of the data. This term turns to assign small weights to the instance pairs that are far apart from each other, making sure that the results are mainly affected by neighboring instance pairs and hence sufficiently smooth with respect to the intrinsic structure revealed by the training data.

3.4.4 Connection with LPP and FDA

We provide some theoretical analysis of the connection between our proposed algorithm and two popular dimensionality reduction methods, i.e., the global statistical method FDA and the local geometric method LPP [9]. We show that even the initial solution of MIE (i.e., embedding the initial graph) shares the advantages of both FDA and LPP and mitigates their drawbacks at the same time.

Theorem 3. Maximizing $\hat{\mathfrak{J}}_{\mathbf{z}}$ is equivalent to LPP.

Proof. We have

$$\begin{aligned}\tau &= \arg \max(\hat{\mathfrak{J}}_{\mathbf{z}} = - \sum_{i=1}^N \sum_{j=1}^N w_{ij}^{(0)} \|\mathbf{z}_i - \mathbf{z}_j\|^2) \\ &= \arg \min \sum_{m=1}^M \mathbf{t}_m^\top X^\top L^{(0)} X \mathbf{t}_m,\end{aligned}$$

where $w_{ij}^{(0)} = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\delta^2}$ is the adjacency matrix of a heat graph, $L^{(0)}$ is its corresponding Laplacian matrix. Recall that LPP minimizes exactly the same objective function except that it uses a different normalization constraint (equivalent to using normalized Laplacian). \square

Theorem 4. *If the classes are balanced, i.e., $\frac{N_c}{N} = \frac{1}{C}$, maximizing $\hat{\mathfrak{J}}_{\mathbf{z}y}$ is reduced to a localized version of FDA (LFDA, [20]).*

Proof. We have

$$\begin{aligned}\tau &= \arg \max(\hat{\mathfrak{J}}_{\mathbf{z}y} \propto - \sum_{i=1}^N \sum_{j=1}^N (w_{ij}^{(+)} - w_{ij}^{(-)}) \|\mathbf{z}_i - \mathbf{z}_j\|^2) \\ &= \arg \min \sum_{m=1}^M \mathbf{t}_m^\top X^\top (L^{(+)} - L^{(-)}) X \mathbf{t}_m,\end{aligned}\quad (13)$$

where $w_{ij}^{(+)} = I(y_i = y_j) e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\delta^2}$ and $w_{ij}^{(-)} = (\hat{p}_{y_i} + \hat{p}_{y_j}) e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\delta^2}$ defines two graphs, $L^{(+)}$ and $L^{(-)}$ represents their Laplacian matrices respectively. Assume the classes are balanced, $\forall i, j, \hat{p}_{y_i} = \hat{p}_{y_j} = 1/C$, the Fisher criterion can be rewritten as:

$$\max \sum_{m=1}^M \frac{\mathbf{t}_m^\top S_B \mathbf{t}_m}{\mathbf{t}_m^\top S_W \mathbf{t}_m} \propto \sum_{m=1}^M \frac{\mathbf{t}_m^\top X^\top L^{(-)} X \mathbf{t}_m}{\mathbf{t}_m^\top X^\top L^{(+)} X \mathbf{t}_m} + \text{const.}\quad (14)$$

The equivalence between Eq.(13) and Eq.(14) follows directly from the equivalence of trace ratio and trace difference [7]. \square

To summarize, both LPP and FDA can be viewed as degraded forms of the initial MIE, They maximize solely either $\hat{\mathfrak{J}}_{\mathbf{z}}$ or $\hat{\mathfrak{J}}_{\mathbf{z}y}$. In contrast, MIE simultaneously optimizes both $\hat{\mathfrak{J}}_{\mathbf{z}}$ and $\hat{\mathfrak{J}}_{\mathbf{z}y}$. In addition, instead of combining these two goals in an ad-hoc way, MIE uses a natural parameter $\eta = \hat{\mathfrak{J}}_y = \sum_c \frac{N_c^2}{N^2}$ that is derived from a theoretically optimal criterion to strive for a reasonable balance between these two goals.

As an empirical validation, we test FDA, LPP and the initial MIE on a synthetic 2-D data set [20]. The results are shown in Fig.1. We can see that in an

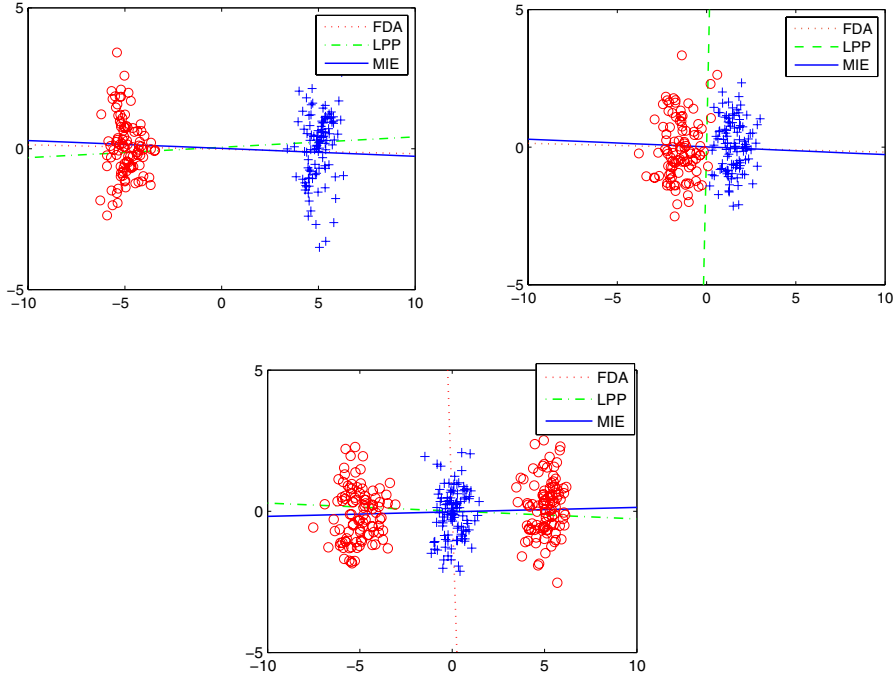


Fig. 1. Typical behaviors of global method (FDA), local method (LLP) and globally-locally consistent method (initial solution of MIE)

ideal case as depicted in the leftmost subfigure, LPP, FDA and MIE are all able to find an excellent projection for the data. In the middle subfigure, while the MIE and FDA can still find a good projection, the geometric and unsupervised method LPP gets a very poor projection such that the data from both classes are totally mixed up. In the rightmost subfigure where the data has multiple modes, the methods that account for locality geometric information (i.e., MIE and LPP) successfully obtain a good feature. However, FDA totally fails, which is not surprising since it is based on Gaussian distribution assumption.

4 Using Bayes Error Rate

We now apply the procedures to the Bayes Error Rate (BER) criterion. From Eq.(2), we have:

$$\mathcal{R} \propto \int_{\mathbf{z}} p(\mathbf{z})(p(y)p(\mathbf{z}|y) - p(c \neq y)p(\mathbf{z}|c \neq y))d\mathbf{z} + \text{const.}$$

Using nonparametric estimator, we have:

$$\begin{aligned}
\min_{\tau \in \mathcal{H}} \mathfrak{J}(\mathbf{z}) &= \int_{\mathbf{z}} p(\mathbf{z})(p(y)p(\mathbf{z}|y) - p(c \neq y)p(\mathbf{z}|c \neq y))d\mathbf{z}. \\
&\approx \sum_{n=1}^N [\hat{P}_{y_n} \sum_{i \in \Omega_n^{(o)}} k(\mathbf{z}_i, \mathbf{z}_n) - (1 - \hat{P}_{y_n}) \sum_{j \in \Omega_n^{(e)}} k(\mathbf{z}_j, \mathbf{z}_n)] \\
&= \sum_{i=1}^N \sum_{j=1}^N r_{ij} k(\mathbf{z}_i, \mathbf{z}_j),
\end{aligned}$$

where $\Omega_n^{(o)} = \{i : y_i = y_n\}$ and $\Omega_n^{(e)} = \{j : y_j \neq y_n\}$ denote the homogenous and heterogeneous index set of \mathbf{z}_n , $\hat{P}_c = N_c/N$, and

$$r_{ij} = \begin{cases} 2\hat{P}_{y_i} & \text{if } y_i = y_j \\ \hat{P}_{y_i} + \hat{P}_{y_j} - 2 & \text{otherwise} \end{cases}$$

The remaining procedures (i.e., kernel approximation and feature construction) are quite straightforward; we will omit the detailed derivation. Eventually, we will derive a variational graph with adjacency $w_{ij} = -r_{ij}\lambda_{ij}$ (similarly, the initial graph $w_{ij} = r_{ij}e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\delta^2}$), and get feature extractors by spectral analysis based on the Laplacian of this graph (the resulting algorithms are referred to as BER Embedding, or BERE/kBERE). It can be easily proved that BERE also has the Max-Discrimination and Locality-Preserving properties.

5 Experiment

We test our proposed algorithms on real-world face recognition tasks. For comparison, PCA, FDA, LPP, MFA (Margin Fisher Analysis,[24]) and their kernel counterparts are selected as baselines, among which PCA and LPP are unsupervised, FDA and MFA are supervised, and LPP and MFA account for local geometric structures.

Three benchmark facial image sets are selected: (1) the **Yale**² data set, which contains 165 facial images of 15 persons, 11 images for each individual; (2) the **ORL**³ data set, which consists of 400 facial images of 40 persons; and (3) the **CMU Pie**⁴ data set, which contains 41368 facial images of 68 individuals. All the three sets of images were taken in different environments, at different times, with different poses, facial expressions and details. In our experiment, all the raw images are normalized to 32×32 . For each data set, we randomly select ν images of each person as training data (referred to as ν train), and leave others for testing. Only the training data are used to learn features. To evaluate the effectiveness of different methods, the classification accuracy of a k -NN classifier on testing data is used as the evaluation metric.

² <http://cvc.yale.edu/projects/yalefaces/yalefaces.html>

³ <http://www.uk.research.att.com/facedatabase.html>

⁴ http://www.ri.cmu.edu/projects/project_418.html

Table 1. Comparison of feature extraction algorithms on face recognition tasks. The results of our proposed algorithms that are better than the performances of baseline methods are highlighted in bold.

Method	Yale			ORL			CMU Pie		
	2train	3train	4train	2train	3train	4train	5train	10train	20train
PCA	.44±.022	.52±.014	.55±.017	.54±.015	.63±.012	.69±.014	.47±.017	.55±.016	.65±.015
FDA	.46±.016	.57±.013	.68±.012	.76±.021	.82±.017	.90±.016	.61±.022	.78±.015	.85±.008
LPP	.50±.019	.61±.018	.67±.015	.70±.018	.78±.017	.86±.014	.62±.022	.75±.016	.84±.013
MFA	.49±.023	.64±.019	.77±.014	.72±.021	.84±.017	.89±.017	.64±.016	.78±.011	.91±.013
MIE ₀	.51±.019	.67±.019	.83±.016	.77±.021	.85±.018	.94±.013	.64±.017	.81±.016	.89±.014
MIE	.50±.016	.67±.018	.85±.018	.75±.019	.88±.018	.92±.017	.65±.019	.83±.018	.93±.015
BERE ₀	.53±.020	.66±.017	.81±.016	.80±.018	.88±.017	.91±.013	.70±.019	.81±.016	.94±.015
BERE	.55±.022	.69±.019	.84±.015	.84±.018	.92±.018	.94±.016	.69±.019	.87±.017	.94±.018
KPCA	.48±.025	.55±.019	.60±.016	.63±.022	.74±.017	.78±.013	.50±.018	.57±.017	.69±.016
KDA	.49±.023	.63±.021	.69±.018	.79±.021	.89±.019	.92±.016	.60±.022	.79±.013	.91±.010
KMFA	.52±.024	.65±.024	.78±.020	.75±.024	.84±.017	.92±.015	.62±.021	.83±.017	.91±.016
kMIE ₀	.55±.026	.70±.021	.84±.018	.86±.022	.91±.017	.94±.015	.69±.018	.83±.016	.93±.017
kMIE	.52±.021	.73±.022	.87±.017	.88±.020	.91±.017	.92±.019	.67±.019	.86±.018	.95±.015
kBERE ₀	.54±.018	.73±.019	.86±.016	.81±.021	.93±.015	.92±.017	.71±.015	.86±.016	.95±.016
kBERE	.57±.021	.72±.019	.89±.018	.83±.022	.95±.019	.95±.016	.74±.019	.89±.016	.94±.018

The only parameter of our algorithms is the bandwidth parameter δ in the initial graph. In our experiments, we adopt the local scaling scheme used in [27]. All the other hyper-parameters (e.g., the number of neighbors k used in k NN and local scaling) are tuned by 5-fold cross validation. The experiments are averaged over 10 random runs. The results are given in Table 1, where each entry represents the mean testing accuracy \pm standard deviation.

From Table 1, we can see that for almost all the entries, our proposed algorithms significantly outperform other baseline methods. Even the algorithms based on the initial graphs (i.e., MIE₀ and BERE₀) perform significantly better than the baselines. Note that the computation complexity of the initial algorithms are of the same order as the baseline methods. The improvements are quite evident. On average, MIE is 36% over PCA, 8% over FDA, and 4% over MFA; BERE is 39% over PCA, 11% over FDA and 6% over MFA. The improvements are even more significant in the kernel case: kMIE (31%,9%,7%) and kBERE(33%,10%,8%) over (KPCA, KDA, KMFA). For most entries in the table, we got p -values less than 0.005. We also observe in our experiment that, when using the initial graph for initialization, the variational graph embedding algorithms (i.e., MIE, BERE) usually converge within 5 iteration steps.

6 Conclusion

In this paper, we have established graph-based feature extraction algorithms based on variational optimization of nonparametric learning criteria. As case studies, we employed two theoretically optimal but computationally intractable feature learning criteria, i.e., Mutual Information and Bayes Error Rate. By nonparametric criteria estimation and kernel term approximation, we reduced the optimization of these criteria to variational graph-embedding problems, which can be solved by an iterative EM-style procedure where the E-Step learns a

variational affinity graph and the M-Step in turn embeds this graph by spectral analysis. The resulting feature learner has several appealing properties such as *maximum discrimination*, *maximum-relevance-minimum-redundancy* and *locality-preserving*. Experiments on benchmark face recognition data sets confirm the effectiveness of our proposed algorithms.

We finally note that our derived graphs (e.g., the initial graphs derived in Section 3 and 4) as well as the approach we used to derive graphs (i.e., *non-parametric learning measure estimation* and *variational approximation of kernel terms*) are not confined to feature extraction scenarios. They might also be useful in a variety of graph-based learning tasks, e.g., semi-supervised learning, relational learning, metric learning. We shall leave such investigations for future research. Sparseness is a desirable property for feature learning, especially for kernel based methods since both the memory for storing the kernel matrix and the training and testing time are typically proportional to the degree of sparsity of the feature extractor. In the future, we would also like to investigate sparse feature learning in the proposed framework.

Acknowledgement. The authors would like to thank the anonymous reviewers for their helpful comments. This work was supported in part by the NSF “Computational methods for nonlinear dimensionality reduction” project (under grant #DMS-0736328), the NSF China grant #60275025 and the MOST of China grant #2007DFC10740.

References

1. Belkin, M., Niyogi, P.: Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In: Proceeding of Advances in Neural Information Processing Systems 14 (NIPS 2001), pp. 585–591 (2001)
2. Bollacker, K.D., Ghosh, J.: Linear Feature Extractors Based on Mutual Information. In: Proceeding of the 13th International Conference on Pattern Recognition (ICPR 1996) (1996)
3. Cai, D., He, X., Han, J.: SRDA: An Efficient Algorithm for Large-Scale Discriminant Analysis. *IEEE Transactions on Knowledge and Data Engineering* 20(1), 1–12 (2008)
4. Choi, E.: Feature Extraction Based on the Bhattacharyya Distance. *Pattern Recognition* 36, 1703–1709 (2003)
5. Fan, J., Li, R.: Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery. In: Proceeding of Regional Conference in Mathematics (AMS 1996), vol. 3, pp. 595–622 (1996)
6. Chung, F.R.K.: Spectral Graph Theory. In: Proceeding of Regional Conference in Mathematics (AMS 1992), vol. 92 (1997)
7. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press, London (1991)
8. Guyon, I., Elissee, A.: An introduction to variable and feature selection. *Journal of Machine Learning Research* 3, 1157–1182 (2003)
9. He, X., Niyogi, P.: Locality Preserving Projections. In: Proceeding of Advances in Neural Information Processing Systems 16 (NIPS 2003) (2003)

10. Hild II, K.E., Erdogmus, D., Torkkola, K., Principe, C.: Feature Extraction Using Information-Theoretic Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(9), 1385–1392 (2006)
11. Jaakkola, T.: Tutorial on Variational Approximation Methods. In: Opper, M., Saad, D. (eds.) *Advanced Mean Field Methods: Theory and Practice*, pp. 129–159. MIT Press, Cambridge (2000)
12. Kaski, S., Peltonen, J.: Informative Discriminant Analysis. In: *Proceeding of the 20th Annual International Conference on Machine Learning (ICML 2003)*, pp. 329–336 (2003)
13. Koller, D., Sahami, M.: Toward Optimal Feature Selection. In: *Proceeding of the 13th International Conference on Machine Learning (ICML 1996)*, pp. 284–292 (1996)
14. Paninski, L.: Estimation of Entropy and Mutual Information. *Neural Computation* 15, 1191–1253 (2003)
15. Peng, H., Long, F., Ding, C.: Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(8), 1226–1238 (2005)
16. Principe, J.C., Fisher III, J.W., Xu, D.: *Information Theoretic Learning*. In: Haykin, S. (ed.) *Unsupervised Adaptive Filtering*. Wiley, Chichester (2000)
17. Roweis, S., Saul, L.K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* 290(22), 2323–2326 (2000)
18. Saon, G., Padmanabhan, M.: Minimum Bayes Error Feature Selection for Continuous Speech Recognition. In: *Proceeding of the 16th Annual Conference on Neural Information Processing Systems (NIPS 2002)*, pp. 800–806 (2002)
19. Saul, L.K., Weinberger, K.Q., Ham, J.H., Sha, F., Lee, D.D.: Spectral Methods for Dimensionality Reduction. In: Chapelle, O., et al. (eds.) *Semisupervised Learning*, MIT Press, Cambridge (2006)
20. Sugiyama, M.: Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis. *Journal of Machine Learning Research* 8, 1027–1061 (2007)
21. Tenenbaum, J.B., Silva, V., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* 290(22), 2319–2323 (2000)
22. Torkkola, K.: Feature Extraction by Nonparametric Mutual Information Maximization. *Journal of Machine Learning Research* 3, 1415–1438 (2003)
23. Weinberger, K.Q., Sha, F., Saul, L.K.: Learning a kernel matrix for nonlinear dimensionality reduction. In: *Proceedings of the 21st Annual International Conference on Machine Learning (ICML 2004)*, pp. 839–846 (2004)
24. Yan, S.C., Xu, D., Zhang, B.Y., Zhang, H.J., Yang, Q., Lin, S.: Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(1), 40–51 (2007)
25. Yang, S.H., Hu, B.G.: Discriminative Feature Selection by Nonparametric Bayes Error Minimization. In: *Knowledge and Information Systems (KAIS) (to appear)*
26. Yang, H., Moody, J.: Data Visualization and Feature Selection: New Algorithms for Nongaussian Data. In: *Proceeding of the 14th Annual Conference on Neural Information Processing Systems (NIPS 2000)*, pp. 687–693 (2000)
27. Zelnik-Manor, L., Perona, P.: Self-tuning Spectral Clustering. In: *Proceeding of the 18th Neural Information Processing Systems (NIPS 2004)*, pp. 1601–1608 (2004)
28. Zhu, X.: *Semi-Supervised Learning Literature Survey*. Technical Report 1530, Department of Computer Sciences, University of Wisconsin-Madison (2005)
29. Zhao, Z., Liu, H.: Spectral feature selection for supervised and unsupervised learning. In: *Proceeding of the 24th Annual International Conference on Machine Learning (ICML 2007)*, pp. 1151–1157 (2007)