Manifold Learning and its Applications:
Papers from the AAAI Fall Symposium (FS-09-04)

Semi-Supervised Learning Using Sparse Eigenfunction Bases

Kaushik Sinha and **Mikhail Belkin**
Dept. of Computer Science and Engineering
Ohio State University, Columbus, OH 43210
{sinhak,mbelkin}@cse.ohio-state.edu

Abstract

We present a new framework for semi-supervised learning with sparse eigenfunction bases of kernel matrices. It turns out that when the *cluster assumption* holds, that is, when the high density regions are sufficiently separated by low density valleys, each high density area corresponds to a unique representative eigenvector. Linear combination of such eigenvectors (or, more precisely, of their Nystrom extensions) provide good candidates for good classification functions. By first choosing an appropriate basis of these eigenvectors from unlabeled data and then using labeled data with Lasso to select a classifier in the span of these eigenvectors, we obtain a classifier, which has a very sparse representation in this basis. Importantly, the sparsity appears naturally from the cluster assumption.

Experimental results on a number of real-world datasets show that our method is competitive with the state of the art semi-supervised learning algorithms and outperforms the natural base-line algorithm (Lasso in the Kernel PCA basis).

1. Introduction

Semi-supervised learning, i.e., learning from both labeled and unlabeled data has received considerable attention in recent years due to its potential in reducing the need for expensive labeled data. However, to make effective use of unlabeled examples one needs to make some assumptions about the connection between the process generating the data and the process of assigning labels. There are two important assumptions popular in semi-supervised learning community the "cluster assumption" (Chapelle, Weston, and Scholkopf 2002) and the "manifold assumption" (Belkin, Niyogi, and Sindhwani 2006). The cluster assumption can be interpreted as saying that two points are likely to have the same class labels if they can be connected by a path passing through a high density area. In other words two high density areas with different class labels must be separated by a low density valley.

In this paper, we develop a framework for semi-supervised learning when the cluster assumption holds. Specifically, we show that when the high density areas are sufficiently separated, a few appropriately chosen eigenfunctions of a convolution operator (which is the continuous counterpart of the kernel matrix) represents the high density areas reasonably well. Under the ideal conditions each high density area can be represented by a single unique eigenfunction called the "representative" eigenfunction. If the cluster assumption holds, each high density area will correspond to just one class label and thus a sparse linear combination of these representative eigenfunctions would be a good classifier. Moreover, the basis of such eigenfunctions can be learned using only the unlabeled data by constructing the Nystrom extension of the eigenvectors of an appropriate kernel matrix.

Thus, given unlabeled data we construct the basis of eigenfunctions and then apply L^1 penalized optimization procedure Lasso (Tibshirani 1996) to fit a sparse linear combination of the basis elements to the labeled data. We provide a detailed theoretical analysis of the algorithm and show that it is comparable to the state-of-the-art on several common UCI datasets.

Rest of the paper is organized as follows. In section 2.we provide the proposed framework for semi-supervised learning and describe the algorithm. In section 3. we provide an analysis of this algorithm to show that it can consistently identify the correct model. In section 4we provide experimental results on synthetic and real datasets and finally we conclude with a discussion in section 5.

2. Semi-supervised Learning Framework

2.1 Outline of the idea

In this section we present a framework for semi-supervised learning under the cluster assumption. Specifically we will assume that (i) data distribution has natural clusters separated by regions of low density and (ii) the label assignment conforms to these clusters.

The recent work of (Shi, Belkin, and Yu 2008a; 2008b) shows that if the (unlabeled) data is clustered, then for each high density region there is a unique eigenfunction of a convolution operator, which takes positive values for points in the chosen cluster and whose values are close to zero everywhere else (no sign change). Moreover, it can be shown (e.g., (Rosasco, Belkin, and Vito 2008)) that these eigenfunctions can be approximated from the eigenvectors of a

kernel matrix obtained from the unlabeled data.

Thus if the cluster assumption holds we expect each cluster to have exactly one label assignment. Therefore eigenfunctions corresponding to these clusters should produce a natural sparse basis for constructing a classification function.

This suggests the following learning strategy:

1. From unlabeled and labeled data obtain the eigenvectors of the Gaussian kernel matrix.
2. From these eigenvectors select a subset of candidate eigenvectors without sign change.
3. Using the Nystrom extension (see (Bengio, Paiement, and Vincent 2003)), extend these eigenvectors to functions defined everywhere.
4. Using the labeled data, apply Lasso (sparse linear regression) in the constructed basis to obtain a classifier.

We now proceed with the detailed discussion of our algorithm and its analysis.

2.2 Algorithm

The focus of our discussion will be binary classification in the semi-supervised setting. Given l labeled examples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l}$ sampled from an underlying joint probability distribution $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$, $\mathcal{X} \subset \mathbb{R}^d, \mathcal{Y} = \{-1, 1\}$, where $\boldsymbol{x}_i$'s are the data points (feature vectors or predictors), y_i's are their corresponding labels (responses) and u unlabeled examples[1] $\{\boldsymbol{z}_i\}_{i=1}^{u}$ drawn iid from the marginal distribution $\mathcal{P}_{\mathcal{X}}$, we choose a Gaussian kernel $k(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\frac{||\boldsymbol{x}-\boldsymbol{z}||^2}{2\omega^2}\right)$ with kernel bandwidth ω to form the Gram matrix K_u where $(K_u)_{ij} = \frac{1}{u}k(\boldsymbol{z}_i, \boldsymbol{z}_j)$. Let $(\lambda_i, \boldsymbol{v}_i)_{i=1}^{u}$ be the eigenvalue-eigenvector pair of the Gram matrix K_u sorted by the non-increasing eigenvalues. The Nystrom extension of the i^{th} eigenvector is given by,

$$\psi_i^u(\boldsymbol{x}) = \frac{1}{\lambda_i\sqrt{u}} \sum_{j=1}^{u} \boldsymbol{v}_i(\boldsymbol{z}_j)k(\boldsymbol{x}, \boldsymbol{z}_j) \tag{1}$$

Even though ideally the unique eigenvectors will have no sign change, in real life, depending on the separation among the high density clusters we allow the unique eigenvectors to have no sign change up to some small precision $\epsilon > 0$, where we say that a vector $\boldsymbol{e} = (e_1, e_2, ..., e_n) \in \mathbb{R}^n$ has no sign change up to precision ϵ if either $\forall i\ e_i > -\epsilon$ or $\forall i\ e_i < \epsilon$. Let N_ϵ be the set of indices all eigenvectors that have no sign change up to precision ϵ. Note that the set N_ϵ and the set $\{1, 2, ..., |N_\epsilon|\}$ are not necessarily same. Our goal is to obtain a function $f(\boldsymbol{x}) = \sum_{i\in N_\epsilon} \beta_i\psi_i^u(\boldsymbol{x})$ that minimizes classification error for the data-label pairs generated from a linear regression model, $y_i = \sum_{j\in N_\epsilon} \beta_j^*\psi_j^u(\boldsymbol{x}_i) + \varepsilon_i$ where ε_is represent stochastic noise (zero mean bounded variance iid random variables) that are independent of y_is and $\boldsymbol{x}_i$s and most of the β_i^*s are zeros. Ideally, to get a sparse solution, we would like to minimize L_0 penalized (on β_is) convex loss function $V\left(\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l}, \boldsymbol{\beta}\right)$. Since such an optimization problem is NP hard, standard approach is to apply a L_1 penalty on β_is. If we select V to be square loss function, we end up solving the L_1 penalized least square or so called Lasso (Tibshirani 1996), whose consistency property were studied in (Zhao and Yu 2006). Thus we would seek a solution of the form

$$\arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \Psi\boldsymbol{\beta})^T(\boldsymbol{y} - \Psi\boldsymbol{\beta}) + \lambda||\beta||_{L^1} \tag{2}$$

which is a convex optimization problem, where Ψ is the $l \times |N_\epsilon|$ design matrix whose i^{th} column is the restriction of $\psi_{a_i}^u, a_i \in N_\epsilon$ to l labeled examples, $\boldsymbol{y} \in \mathbb{R}^l$ is the label vector, $\boldsymbol{\beta}$ is the vector of coefficients and λ is a regularization parameter. Note that solving the above problem is equivalent to solving

$$\arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \Psi\boldsymbol{\beta})^T(\boldsymbol{y} - \Psi\boldsymbol{\beta})\ \ s.t.\ \sum_{i\in N_\epsilon} |\beta_i| \le t \tag{3}$$

because for any given $\lambda \in [0, \infty)$, there exists a $t \ge 0$ such that the two problems have the same solution, and vice versa (Tibshirani 1996). We will denote the solution of Equation 3, by $\hat{\boldsymbol{\beta}}$. Let the set $\mathcal{T}$ contains indices of all nonzero $\hat{\beta}_i$s. The classification function is given by, $f(\boldsymbol{x}) = \sum_{i\in\mathcal{T}} \hat{\beta}_i\psi_i^u(\boldsymbol{x}) = \sum_{i=1}^{u} \mathrm{w}_i k(\boldsymbol{z}_i, \boldsymbol{x})$, where, $\mathrm{w} \in \mathbb{R}^u$ is a weight vector whose i^{th} element is given by

$$\mathrm{w}_i = \sum_{j\in\mathcal{T}} \frac{\hat{\beta}_j\boldsymbol{v}_j(\boldsymbol{z}_i)}{\lambda_j\sqrt{u}} \tag{4}$$

and can be computed while training.

Algorithm for Semi-supervised Learning

Input: $\{\boldsymbol{z}_i\}_{i=1}^{u}$, $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l}$
Parameters: ω, t, ϵ

1. From u unlabeled examples $\{\boldsymbol{z}_i\}_{i=1}^{u}$, generate the kernel Matrix K_u.
2. Select the set of indices N_ϵ of eigenvectors having no sign change up to precision ϵ.
3. Extend each $\boldsymbol{v}_i, i \in N_\epsilon$ to form ψ_i^us using Equation 1.
4. Using $\{\psi_i^u\}_{i\in N_\epsilon}$ and l labeled examples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l}$, form the design matrix Ψ.
5. Solve Equation 3 to get $\hat{\boldsymbol{\beta}}$ and calculate weight vector w using Equation 4.
6. Given a test point $\boldsymbol{x}$, predict its label as $y = sign\left(\sum_{i=1}^{u} k(\boldsymbol{z}_i, \boldsymbol{x})\mathrm{w}_i\right)$

---

3. Analysis of the Algorithm

Before starting the actual analysis, we first note that the continuous counterpart of the Gram matrix is a convolution operator $L_K : L^2(\mathcal{X}, \mathcal{P}_\mathcal{X}) \to L^2(\mathcal{X}, \mathcal{P}_\mathcal{X})$ defined by,

$$(L_K f)(\boldsymbol{x}) = \int_{\mathcal{X}} k(\boldsymbol{x}, \boldsymbol{z})f(\boldsymbol{z})d\mathcal{P}_\mathcal{X}(\boldsymbol{z}) \tag{5}$$

[1] We use the notation $\boldsymbol{z}$ instead of $\boldsymbol{x}$ purely for notational convenience. Now the index goes $\{\boldsymbol{z}_i\}_{i=1}^{u}$ instead of $\{\boldsymbol{x}_i\}_{i=l+1}^{l+u}$. Also, note that, we will use the terms data and examples interchangeably.

The eigenfunctions of the symmetric positive definite operator L_K are denoted by ϕ_i^L.

Next, we briefly discuss the effectiveness of model selection using Lasso (established by (Zhao and Yu 2006)) which will be required for our analysis. Let $\hat{\boldsymbol{\beta}}_l(\lambda)$ be the solution of Equation 2 for a chosen regularization parameter λ. In (Zhao and Yu 2006) a concept of *sign consistency* was introduced which states that Lasso is sign consistent if, as l tends to infinity, signs of $\hat{\boldsymbol{\beta}}_l(\lambda)$ matches with that of $\boldsymbol{\beta}^*$ with probability 1, where $\boldsymbol{\beta}^*$ is the coefficients of the correct model. Note that since we are expecting a sparse model matching zeros of $\hat{\boldsymbol{\beta}}_l(\lambda)$ to the zeros of $\boldsymbol{\beta}^*$ is not enough, but in addition, matching the sign of the non zero coefficients ensures that the true model will be selected. Next, without loss of generality assume $\beta^* = (\beta_1^*, \cdots, \beta_q^*, \beta_{q+1}^*, \cdots, \beta_{|N_\epsilon|}^*)$ has only first q terms non-zero, i.e., only q predictors describe the model and rest of the predictors are irrelevant in describing the model. Now let us write the first q and $|N_\epsilon| - q$ columns of Ψ as $\Psi_{(1)}$ and $\Psi_{(2)}$ respectively. Let $C = \frac{1}{l}\Psi^T\Psi$.

For a random design matrix, sign consistency is equivalent to *irrepresentable* condition (see (Zhao and Yu 2006)). When $\boldsymbol{\beta}^*$ is unknown, in order to ensure that irrepresentable condition holds for all possible signs, it requires that L_1 norm of the regression coefficients corresponding to the irrelevant predictors to be less than 1, which can be written as $\mu_\Psi = \max_{\psi_j^u \in \Psi_{(2)}} \left\| \left(\Psi_{(1)}^T \Psi_{(1)} \right)^T \Psi_{(1)}^T \psi_j^u \right\|_1 < 1$. The requirement $\mu_\Psi < 1$ is not new and have also appeared in the context of noisy or noiseless sparse recovery of signal (Tropp 2004; Wainwright 2006; Zhang 2008). Note that Lasso is sign consistent if irrepresentable condition holds and the sufficient condition needed for irrepresentable condition to hold is given by the following result,

Theorem 1. ***(Zhao and Yu 2006)*** *Suppose $\boldsymbol{\beta}^*$ has q nonzero entries. Let the matrix C' be normalized version of C such that $C'_{ij} = \frac{C_{ij}}{C_{ii}}$ and $\max_{i,j,i\neq j} |C'_{ij}| \leq \frac{c}{2q-1}$ for a constant $0 \leq c < 1$, then strong irrepresentable condition holds.*

In section 3.4we will show that this sufficient condition is satisfied with high probability requiring relatively few labeled examples, as a result the correct model is identified consistently which in turn describes a good classification function.

3.1 Brief Overview of the Analysis

Our analysis lies on three ideas,- (1) sufficient *separation* among high density regions ensuring a sparse model that describes data-label pairs sampled from $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$, (2) *finite sample* result describing how well Nystrom extensions approximate eigenfunctions of L_K using unlabeled examples alone and (3) *concentration* result ensuring model selection consistency of Lasso with high probability using relatively few labeled examples. We first provide a brief overview of these ideas and then present the technical details in the following subsections. All the proofs have been deferred to the Appendix for space limitation.

Note that cluster assumption can also be interpreted as follows,- density function of an individual class has fast tail decay and there is little overlap among the density functions representing different classes. It is shown in (Shi, Belkin, and Yu 2008b) that eigenfunctions of L_K, when L_K is applied to individual density functions separately, are almost preserved when L_K is applied to the combined mixture as well, in case of a mixture of Gaussians, and also in general (Shi, Belkin, and Yu 2008a), provided tails of individual density functions corresponding to each class decay reasonably fast. As a first step of our analysis, in section 3.2 , for a class of probability distributions characterized by fast tail decay, we estimate the separation requirement among the high density regions ensuring that each high density region (class) can be well represented by a unique eigenfunction, which we call "representative" eigenfunction having the property that it has no sign change and it has higher values within a high density region and decays exponentially fast outside the high density region. This allows us to express the classification task in this eigenfunction basis where we look for a classification function consisting of linear combination of representative eigenfunctions only and thus relate the problem to sparse approximation from the model selection point of view, which is a well studied field (Wainwright 2006; Zhang and Huang 2006; Candes and Plan 2007).

Let $N_{\max} = \max_i\{i : i \in N_\epsilon\}$. Assuming that the first $N_{\max}$ eigenvalues of L_K and K_u, sorted in non-increasing order, are simple and bounded away from zero, in the second step, using perturbation results from (Rosasco, Belkin, and Vito 2008) we show that, with high probability, for $i, j \in N_\epsilon$, $\|\psi_i^u - \phi_i^L\|_{L^2(\mathcal{X},\mathcal{P}_\mathcal{X})} = O\left(\frac{1}{\sqrt{u}}\right)$ and accordingly $\langle \psi_i^u, \psi_j^u \rangle_{L^2(\mathcal{X},\mathcal{P}_\mathcal{X})} = O(\frac{1}{\sqrt{u}})$ for all $i, j, i \neq j$ and $\|\psi_i^u\|_{L^2(\mathcal{X},\mathcal{P}_\mathcal{X})} = 1 + O(\frac{1}{\sqrt{u}})$ for all i. This implies that if the number of unlabeled examples u is large enough, then $\langle \psi_i^u, \psi_j^u \rangle_{L^2(\mathcal{X},\mathcal{P}_\mathcal{X})} = \int_\mathcal{X} \psi_i^u(X)\psi_j^u(X) d\mathcal{P}_\mathcal{X}(X) = \mathbb{E}\left(\psi_i^u(X)\psi_j^u(X)\right) \approx 0$ and $\langle \psi_i^u, \psi_i^u \rangle_{L^2(\mathcal{X},\mathcal{P}_\mathcal{X})} = \int_\mathcal{X} [\psi_i^u(X)]^2 d\mathcal{P}_\mathcal{X}(X) = \mathbb{E}\left([\psi_i^u(X)]^2\right) \approx 1$. i.e., these eigenfunctions $\{\psi_i^u\}_{i\in N_\epsilon}$ form an orthonormal basis of $L^2_{N_\epsilon}(\mathcal{X}, \mathcal{P}_\mathcal{X}) \subset L^2(\mathcal{X}, \mathcal{P}_\mathcal{X})$ in the limit as $u \to \infty$, where $L^2_{N_\epsilon}(\mathcal{X}, \mathcal{P}_\mathcal{X})$ is the space spanned by the eigenfunctions of L_K having indices in the set N_ϵ.

So far we have used only unlabeled examples to learn the eigenfunctions ψ_i^us. Next, we use the labeled examples to perform L^1 regularized regression to identify the representative eigenfunctions and their nonzero coefficients. Due to Theorem 1, it is enough to show that the off-diagonal terms of C' are strictly less than $\frac{1}{2q-1}$, where q is the number of columns that describes the sparse model (in case of binary classification $q = 2$, which are the representative eigenfunctions corresponding to each class). Now, any off-diagonal term of the matrix C is given by $C_{ij} = \frac{1}{l}\sum_{k=1}^{l} \psi_{a_i}^u(\boldsymbol{x}_k)\psi_{a_j}^u(\boldsymbol{x}_k), a_i, a_j \in N_\epsilon$ which is empirical version of the expectation $\mathbb{E}\left(\psi_{a_i}^u(X)\psi_{a_j}^u(X)\right)$. In the third step, using standard concentration inequality result we show that C_{ij} is tightly concentrated around its expected value and converges in probability to its expectation exponentially fast

in number of labeled examples. Since for a large enough $u, \mathbb{E}\left(\psi^u_{a_i}(X)\psi^u_{a_j}(X)\right) \approx 0$, it is not hard to show that sufficient condition of (Zhao and Yu 2006) is easily satisfied. In other words, with high probability, only a few labeled examples will be good enough to ensure model consistency of Lasso.

3.2 Separation Requirement

To motivate our discussion we consider binary classification problem where the underlying assumption is that each class has its own probability density function, denoted by $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ respectively and corresponding mixing weights π_1, π_2. Thus, the density of the mixture is $p(\boldsymbol{x}) = \pi_1 p_1(\boldsymbol{x}) + \pi_2 p_2(\boldsymbol{x})$. We will use the following results from (Shi, Belkin, and Yu 2008a) specifying the behavior of the eigenfunction corresponding to the largest eigenvalue.

Theorem 2. *(Shi, Belkin, and Yu 2008a) The top eigenfunction $\phi^L_0(\boldsymbol{x})$ of L_K corresponding to the largest eigenvalue λ_0, (1) is the only eigenfunction with no sign change, (2) has multiplicity one, (3) is non zero on the support of the underlying density, (4) satisfies $|\phi^L_0(\boldsymbol{x})| \leq \frac{1}{\lambda_0}\sqrt{\int k^2(\boldsymbol{x},\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}}$ (Tail decay property), where p is the underlying probability density function.*

Note that the last (tail decay) property above is not restricted to the top eigenfunction alone but is satisfied by all eigenfunctions of L_K. Now, consider applying L_K to the three cases when the underlying probability distributions are p_1, p_2 and p which corresponds to individual classes respectively and the mixture and where the largest eigenvalues and corresponding eigenfunctions are $\lambda^1_0, \lambda^2_0, \lambda_0$ and $\phi^{L,1}_0, \phi^{L,2}_0, \phi^L_0$ respectively. To show explicit dependency on the underlying probability distribution, we will denote the corresponding operators as $L^{p_1}_K, L^{p_2}_K$ and L^p_K respectively. Clearly, $L^p_K = \pi_1 L^{p_1}_K + \pi_2 L^{p_2}_K$. Then we can write, $L^p_K \phi^{L,1}_0(\boldsymbol{x}) = \int k(\boldsymbol{x},\boldsymbol{z})\phi^{L,1}_0(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \pi_1 \int k(\boldsymbol{x},\boldsymbol{z})\phi^{L,1}_0(\boldsymbol{z})p_1(\boldsymbol{z})d\boldsymbol{z} + \pi_2 \int k(\boldsymbol{x},\boldsymbol{z})\phi^{L,1}_0(\boldsymbol{z})p_2(\boldsymbol{z})d\boldsymbol{z} = \pi_1\lambda^1_0\phi^{L,1}_0 + \pi_2 \int k(\boldsymbol{x},\boldsymbol{z})\phi^{L,1}_0(\boldsymbol{z})p_2(\boldsymbol{z})d\boldsymbol{z} = \pi_1\lambda^1_0\left(\phi^{L,1}_0 + \frac{\pi_2}{\pi_1\lambda^1_0}\int k(\boldsymbol{x},\boldsymbol{z})\phi^{L,1}_0(\boldsymbol{z})p_2(\boldsymbol{z})d\boldsymbol{z}\right) = \pi_1\lambda^1_0\left(\phi^{L,1}_0 + T_1(\boldsymbol{x})\right)$ where, $T_1(\boldsymbol{x}) = \frac{\pi_2}{\pi_1\lambda^1_0}\int k(\boldsymbol{x},\boldsymbol{z})\phi^{L,1}_0(\boldsymbol{z})p_2(\boldsymbol{z})d\boldsymbol{z}$. In a similar way we can write, $L^p_K\phi^{L,2}_0(\boldsymbol{x}) = \pi_2\lambda^2_0\left(\phi^{L,2}_0 + T_2(\boldsymbol{x})\right)$ where, $T_2(\boldsymbol{x}) = \frac{\pi_1}{\pi_2\lambda^2_0}\int k(\boldsymbol{x},\boldsymbol{z})\phi^{L,2}_0(\boldsymbol{z})p_1(\boldsymbol{z})d\boldsymbol{z}$. Thus, when $T_1(\boldsymbol{x})$ and $T_2(\boldsymbol{x})$ are small enough then $\phi^{L,1}_0$ and $\phi^{L,2}_0$ are eigenfunctions of L^p_K with corresponding eigenvalues $\pi_1\lambda^1_0$ and $\pi_2\lambda^2_0$ respectively. Note that "separation condition" requirement refers to $T_1(\boldsymbol{x})$, $T_2(\boldsymbol{x})$ being small, so that eigenfunctions corresponding to the largest eigenvalues of convolution operator when applied to individual high density bumps can be preserved when convolution operator applied to the mixture. Clearly, we can not expect $T_1(\boldsymbol{x})$, $T_2(\boldsymbol{x})$ to arbitrarily small if there is sufficient overlap between p_1 and p_2. Thus, we will restrict ourselves to the following class of probability distributions for each individual class which has reasonably fast tail decay.

Assumption 1. *For any $1/2 < \eta < 1$, let $\mathbb{M}(\eta, \mathcal{R})$ be the class of probability distributions such that its density function p satisfies*
1) $\int_{\mathcal{R}} p(\boldsymbol{x})d(\boldsymbol{x}) = \eta$ where $\mathcal{R}$ is the minimum volume ball around the mean of the distribution.
2) For any positive $t > 0$, smaller than the radius of $\mathcal{R}$, and for any point $\boldsymbol{z} \in \mathcal{X} \setminus \mathcal{R}$ with $dist(\boldsymbol{z}, \mathcal{R}) \geq t$, the volume $S = \{\boldsymbol{x} \in (\mathcal{X} \setminus \mathcal{R}) \cap B(\boldsymbol{z}, 3t/\sqrt{2})\}$ has total probability mass $\int_S p(\boldsymbol{x})d\boldsymbol{x} \leq C_1\eta\exp\left(-\frac{dist^2(\boldsymbol{z},\mathcal{R})}{t^2}\right)$ for some $C_1 > 0$.

where the distance between a point $\boldsymbol{x}$ and set $\mathcal{D}$ is defined as $dist(\boldsymbol{x}, \mathcal{D}) = \inf_{\boldsymbol{y}\in\mathcal{D}} ||\boldsymbol{x} - \boldsymbol{y}||$. With a little abuse of notation we will use $p \in \mathbb{M}(\eta, \mathcal{R})$ to mean that p is the probability density function of a member of $\mathbb{M}(\eta, \mathcal{R})$. Now a rough estimate of separation requirement can be given by the following lemma.

Lemma 1. *Let $p_1 \in \mathbb{M}(\eta, \mathcal{R}_1)$ and $p_2 \in \mathbb{M}(\eta, \mathcal{R}_2)$ and let the minimum distance between $\mathcal{R}_1, \mathcal{R}_2$ be Δ. If $\Delta = \Omega^*\left(\omega\sqrt{d}\right)$ then $T_1(\boldsymbol{x})$ and $T_2(\boldsymbol{x})$ can be made arbitrarily small for all $\boldsymbol{x} \in \mathcal{X}$.*

The estimate of Δ in the above lemma, where we hide the log factor by Ω^*, is by no means tight, nevertheless, it shows that separation requirement refers to existence of a low density valley between two high density regions each corresponding to one of the classes. This separation requirement is roughly of the same order required to learn mixture of Gaussians (Dasgupta 1999). Note that, provided separation requirement is satisfied, $\phi^{L,1}_0$ and $\phi^{L,2}_0$ are not necessarily the top two eigenfunctions of L_K corresponding to the two largest eigenvalues but one of them can be quite far down the spectrum of L^p_K depending on the mixing weights π_1, π_2. Next, the following lemma suggests that we can say more about the eigenfunction corresponding to the largest eigenvalue.

Lemma 2. *For any $\frac{e}{1+e} < \eta < 1$, let $q \in \mathbb{M}(\eta, \mathcal{R})$. If ϕ^L_0 is the eigenfunction of L^q_K corresponding to the largest eigenvalue λ_0 then there exists a $C_1 > 0$ such that*
1) For all $\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{R}, |\phi^L_0(\boldsymbol{x})| \leq \frac{\sqrt{(C_1+\eta)}}{\lambda_0}\exp\left(-\frac{dist^2(\boldsymbol{x},\mathcal{R})}{2\omega^2}\right)$
2) For all $\boldsymbol{z} \in \mathcal{R}$ and $\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{R}, |\phi^L_0(\boldsymbol{z})| \geq |\phi^L_0(\boldsymbol{x})|$

Thus, top eigenfunctions corresponding to the largest eigenvalues of each class represent high density region reasonably well, outside high density region they have a lower absolute value decays exponentially fast and provided separation requirement is satisfied they are pretty much preserved in the mixture. These are the eigenfunctions that we will refer to as representative eigenfunction.

3.3 Finite Sample Results

We start with the following assumption.

Assumption 2. *The $N_{\max}$ largest eigenvalues of L_K and K_u, where $N_{\max} = \max_i\{i : i \in N_\epsilon\}$, are simple and bounded away from zero.*

Note that Nystrom extension ψ_i^us are eigenfunctions of an operator $L_{K,u} : \mathcal{H} \to \mathcal{H}$ (see Appendix), where $\mathcal{H}$ is the unique RKHS defined by the chosen Gaussian kernel and all the eigenvalues of K_u are also eigenvalues of $L_{K,u}$. There are two implications of Assumption 2. The first one is due to the *bounded away from zero* part, which ensures that if we restrict to $\psi_i^u \in \mathcal{H}$ corresponding to the largest N eigenvalues, then each of them is square integrable and thus is an element of $L^2(\mathcal{X}, \mathcal{P}_\mathcal{X})$. The second implication due to the *simple* part, ensures that the $N_{\max}$ eigenfunctions corresponding to the $N_{\max}$ largest eigenvalues are uniquely defined and so is the orthogonal projection on them. Note that if any eigenvalue has multiplicity greater than one then the corresponding eigenspace is well defined but not the individual eigenfunctions. Thus, Assumption 2 enables us to compare how close each ψ_i^u is to some other function in $L^2(\mathcal{X}, \mathcal{P}_\mathcal{X})$ in $L^2(\mathcal{X}, \mathcal{P}_\mathcal{X})$ norm sense. Let $g_{N_{\max}}$ be the $N_{\max}^{th}$ eigengap when eigenvalues of L_K are sorted in non increasing order. Then we have the following results.

Lemma 3. *Suppose Assumption 2 holds and the top $N_{\max}$ eigenvalues of L_K and K_u are sorted in the decreasing order. Then for any $0 < \delta < 1$ and for any $i \in N_\epsilon$, with probability at least $(1-\delta)$, $\|\psi_i^u - \phi_i^L\|_{L^2(\mathcal{X},\mathcal{P}_\mathcal{X})} = \frac{2}{g_{N_{\max}}}\sqrt{\frac{2\log(2/\delta)}{u\lambda_i}}$*

Corollary 1. *Under the above conditions, for any $0 < \delta < 1$ and for any $i, j \in N_\epsilon$, with probability at least $(1-\delta)$ the following holds,*

1) $\left|\langle\psi_i^u, \psi_j^u\rangle_{L^2(\mathcal{X},\mathcal{P}_\mathcal{X})}\right| \leq \left(\frac{8\log(2/\delta)}{g^2_{N_{\max}}\sqrt{\lambda_i\lambda_j}}\right)\frac{1}{u} + \left(\frac{\sqrt{8\log(2/\delta)}}{g_{N_{\max}}}\left(\frac{1}{\sqrt{\lambda_i}} + \frac{1}{\sqrt{\lambda_j}}\right)\right)\frac{1}{\sqrt{u}}$

2) $1 - \left(\sqrt{\frac{8\log(2/\delta)}{g^2_{N_{\max}}\lambda_i}}\right)\frac{1}{\sqrt{u}} \leq \|\psi_i^u\|_{L^2(\mathcal{X},\mathcal{P}_\mathcal{X})} \leq 1 + \left(\sqrt{\frac{8\log(2/\delta)}{g^2_{N_{\max}}\lambda_i}}\right)\frac{1}{\sqrt{u}}$

3.4 Concentration Results

Having established that $\{\psi_i^u\}_{i\in N_\epsilon}$ form an orthonormal basis in $L^2_{N_\epsilon}(\mathcal{X}, \mathcal{P}_\mathcal{X})$ in the limit, next, we need to consider what happens when we restrict each of the ψ_i^us to finite labeled examples. Note that the design matrix $\Psi \in \mathbb{R}^{l\times|N_\epsilon|}$ is constructed by restricting the $\{\psi_j^u\}_{j\in N_\epsilon}$ to l labeled data points $\{\boldsymbol{x}_i\}_{i=1}^l$ such that the i^{th} column of Ψ is $\left(\psi_{a_i}^u(\boldsymbol{x}_1), \psi_{a_i}^u(\boldsymbol{x}_2), \cdots, \psi_{a_i}^u(\boldsymbol{x}_l)\right)^T \in \mathbb{R}^l, a_i \in N_\epsilon$. Now consider the $|N_\epsilon| \times |N_\epsilon|$ matrix $C = \frac{1}{l}\Psi^T\Psi$ where, $C_{ij} = \frac{1}{l}\sum_{k=1}^l \psi_{a_i}^u(\boldsymbol{x}_k)\psi_{a_j}^u(\boldsymbol{x}_k)$. First, applying Hoeffding's inequality we establish,

Lemma 4. *For all $i, j \in N_\epsilon$ and $\epsilon_1 > 0$ the following two facts hold.*

$\mathbb{P}\left(\left|\frac{1}{l}\sum_{k=1}^l[\psi_i^u(\boldsymbol{x}_k)]^2 - \mathbb{E}\left([\psi_i^u(X)]^2\right)\right| \geq \epsilon_1\right) \leq 2\exp\left(-\frac{l\epsilon_1^2\lambda_i^2}{2}\right)$

$\mathbb{P}\left(\left|\frac{1}{l}\sum_{k=1}^l \psi_i^u(\boldsymbol{x}_k)\psi_j^u(\boldsymbol{x}_k) - \mathbb{E}\left(\psi_i^u(X)\psi_j^u(X)\right)\right| \geq \epsilon_1\right) \leq 2\exp\left(-\frac{l\epsilon_1^2\lambda_i\lambda_j}{2}\right)$

Next, consider the $|N_\epsilon| \times |N_\epsilon|$ normalized matrix C' where $C'_{ij} = \frac{C_{ij}}{C_{ii}}$, ensuring that the diagonal elements $C'_{ii} = 1$. To ensure that Lasso will consistently choose the correct model we need to show (see Theorem 1) that $\max_{i\neq j}|C'_{ij}| < \frac{1}{2q-1}$ with high probability. Applying the above concentration result and finite sample results we have,

Theorem 3. *Let q be the minimum number of columns of the design matrix $\Psi \in \mathbb{R}^{l\times|N_\epsilon|}$, constructed from l labeled examples, that describes the sparse model. Then for any $0 < \delta < 1$, if the number of unlabeled examples u satisfies $u > \frac{2048q^2\log\left(\frac{2}{\delta}\right)}{g^2_{N_{\max}}\lambda^2_{N_{\max}}}$, then with probability greater than $1-\delta-4\exp\left(-\frac{l\lambda^2_{N_{\max}}}{50q^2}\right)$, $\max_{i\neq j}|C'_{ij}| < \frac{1}{2q-1}$.*

where $\lambda_{N_{\max}}$ is the $N_{\max}^{th}$ largest eigenvalue and $g_{N_{\max}}$ is the $N_{\max}^{th}$ eigengap. Note that role played by the labeled and unlabeled examples in our framework follows similar trend in reducing classification error as reported in the literature (Castelli and Cover 1996; Ratsaby and Venkatesh 1995; Sinha and Belkin 2007; Singh, Nowak, and Zhu 2008), specifically, unlabeled examples help only polynomially fast in estimating the eigenfunctions and labeled examples help exponentially fast in identifying the sparse model consisting of representative eigenfunctions.

4. Experimental Results

4.1 Toy dataset

Here we present a synthetic example in 2-D. Consider a binary classification problem where the positive examples are generated from a Gaussian distribution with mean $(0, 0)$ and covariance matrix $[2\ 0;\ 0\ 2]$ and the negative examples are generated from a mixture of Gaussians having means and covariance matrices $(5, 5)$, $[2\ 1;\ 1\ 2]$ and $(7, 7)$, $[1.5\ 0;\ 0\ 1.5]$ respectively. The corresponding mixing weights are $0.4, 0.3$ and 0.3 respectively. Left panel in Figure 1 shows the probability density of the mixture in blue and representative eigenfunctions of each class in green and magenta respectively using 1000 examples (positive and negative) drawn from this mixture. It is clear that each representative eigenfunction represents high density area of a particular class reasonably well. So intuitively a linear combination of them will represent a good decision function. In fact, the right panel of Fig 1 shows the regularization path for L^1 penalized least square regression with 20 labeled examples. The bold green and magenta lines shows the coefficient values for the representative eigenfunctions for different values of regularization parameter t. As can be seen, regularization parameter t can be so chosen that the decision function will consist of a linear combination of representative eigenfunctions only. Note that these representative eigenfunctions need not be the top two eigenfunctions corresponding to the largest eigenvalues.

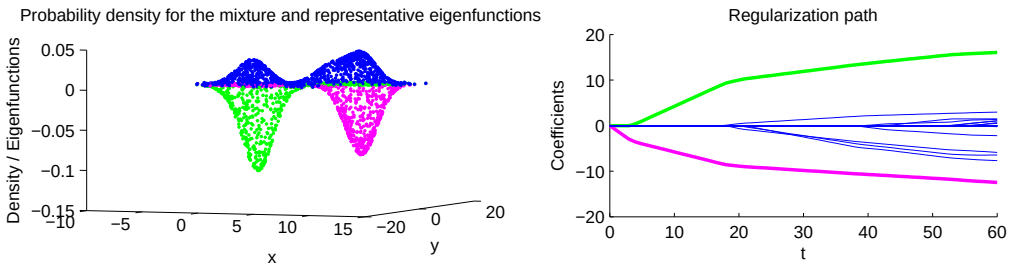

Figure 1: **Left panel:** Probability density of the mixture in blue and representative eigenfunctions in green and magenta. **Right panel:** Regularization path. Bold lines correspond to regularization path associated with representative eigenfunctions.

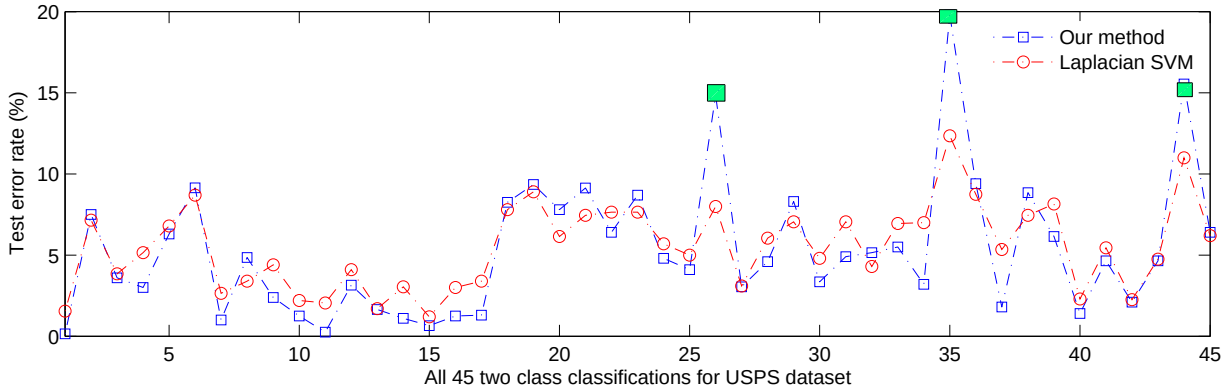

Figure 2: Classification results for USPS dataset

4.2 UCI datasets

In this set of experiment we tested the effectiveness of our algorithm on some common UCI datasets. We compared our algorithm with state of the art semi-supervised learning (manifold regularization) method Laplacian SVM (LapSVM) (Belkin, Niyogi, and Sindhwani 2006), fully supervised SVM and also two other kernel sparse regression methods. In KPCA+L^1 we selected top $|N_\epsilon|$ eigenvectors, and applied L^1 regularization, where as in KPCA_F+L^1 we selected the top 20 (fixed) eigenvectors of K_u and applied L^1 regularization[2]. For both SVM and LapSVM we used RBF kernel. In each experiment a specified number of examples (l) were randomly chosen and labeled and the rest (u) were treated as unlabeled test set. Such random splitting was performed 20 times and the average is reported.

The results are reported in Table 1. As can be seen, for small number of labeled examples our method convincingly outperform SVM and is comparable to LapSVM. The result also suggests that instead of selecting top few eigenvectors, as is normally done in KPCA, selecting them by our method and then applying L^1 regularization yields better result. Table 2 shows that the solution obtained by our method is very sparse, where average sparsity is the average number of non-zero coefficients. The notation A/B represents average sparsity A and number of eigenvectors ($|N_\epsilon|$ or 20).

4.3 Handwritten Digit Recognition

In this set of experiments we applied our method to the 45 binary classification problems that arise in pairwise classification of handwritten digits and compare its performance with LapSVM. For each pairwise classification problem, in each trial, 500 images of each digit in the USPS training set were chosen uniformly at random out of which 20 images were labeled and the rest were set aside for testing. This trial was repeated 10 times. For the LapSVM we set the regularization terms and the kernel as reported by (Belkin, Niyogi, and Sindhwani 2006) for a similar set of experiments, namely we set $\gamma_A l = 0.005$, $\frac{\gamma_I l}{(u+l)^2} = 0.045$ and chose a polynomial kernel of degree 3. The results are shown[3] in Figure2. As can be seen our method is comparable to LapSVM.

We also performed multi-class classification on USPS dataset. In particular, we chose all the images of digits 3, 4 and 5 from USPS training data set (there were 1866 in total) and randomly labeled 10 images from each class. Rest of the 1836 images were set aside for testing. Average prediction accuracy of LapSVM, after repeating this procedure 20

[2]We also selected 100 top eigenvectors and applied L^1 penalty but it gave worse result.

[3]It turned out that the cases where our method performed very poorly, the respective distances between the means of corresponding two classes were very small.

DATA SET	IONOSPHERE d=32, l+u=351			HEART d=75, l+u=303			WINE d=13, l+u=178			BREAST-CANCER d=32, l+u=569		VOTING d=16, l+u=435	
# Labeled Data	l=10	l=20	l=30	l=10	l=20	l=30	l=5	l=10	l=15	l=5	l=10	l=5	l=10
Our Method	**76.55** ±7.42	**78.98** ±7.2	79.14 ±7.05	75.40 ±6.18	**77.38** ±6.04	**79.05** ±1.11	82.29 ±14.07	**96.57** ±6.73	**98.65** ±4.36	**96.88** ±6.92	**99.66** ±1.56	83.62 ±9.33	88.49 ±1.66
LapSVM	73.25 ±6.42	77.64 ±7.40	**79.94** ±6.35	**76.14** ±5.20	77.21 ±2.63	77.29 ±2.85	**82.40** ±14.22	95.33 ±7.66	98.28 ±1.33	95.68 ±8.82	99.05 ±3.52	**88.05** ±5.79	**88.62** ±2.66
SVM	65.16 ±10.87	72.09 ±10.04	79.8 ±9.94	64.61 ±11.63	73.16 ±5.95	76.55 ±4.29	63.47 ±11.57	83.98 ±10.25	88.12 ±11.68	72.83 ±17.56	97.32 ±8.65	81.53 ±16.05	88.51 ±5.88
KPCA+L^1	67.24 ±7.72	74.38 ±8.53	75.23 ±7.55	61.07 ±8.82	67.43 ±6.60	73.11 ±5.59	81.41 ±14.01	90.59 ±9.52	95.18 ±6.74	59.77 ±17.37	73.59 ±12.66	82.61 ±10.32	88.23 ±2.13
KPCA_F+L^1	63.98 ±8.10	67.42 ±8.28	72.29 ±7.56	59.38 ±8.07	66.66 ±7.01	72.33 ±4.91	63.64 ±15.44	76.41 ±12.37	79.07 ±13.08	61.83 ±17.96	73.83 ±14.28	65.12 ±12.08	73.59 ±8.09

Table 1: Classification Accuracies for different UCI datasets

DATA SET	IONOSPHERE	HEART	WINE	BREAST-CANCER	VOTING
Our Method	4.42 / 11	6.33 / 13	3.63 / 7	2.10 / 3	2.00 / 3
KPCA+L^1	6.20 / 11	8.18 / 13	4.27 / 7	2.25 / 3	2.00/ 3
KPCA_F+L^1	8.05 / 20	9.32 / 20	7.25 / 20	4.48 / 20	3.05 / 20

Table 2: Average sparsity of our method for different UCI datasets

times, was 90.14% as compared to 87.53% of our method.

5. Conclusion

In this paper we have presented a framework for spectral semi-supervised learning based on the cluster assumption. We obtain experimental results comparable to the state of the art.

Since our method is closely related to Kernel PCA (KPCA), we would like to point out the important difference. In the case of KPCA, data is projected onto the space spanned by the top eigenvectors corresponding of the kernel matrix and the classification or regression task is performed in that space. For example, this approach was taken in the classification setting in (Blanchard et al. 2004). We note, that since the kernel matrix is computed without any knowledge of the labels, this approach lends itself to a natural semi-supervised algorithm. However, as our experiments confirm, this algorithm does not benefit from the unlabeled data and shows no improvement over a purely supervised algorithm, such as SVM.

Unlike Kernel PCA, our method takes a carefully selected small subset of eigenvectors in an unsupervised manner as a basis, corresponding to the cluster assumption. This selection is crucial to the performance of our method. We then use Lasso with the labeled data to form a classifier that ends up being very sparse in the basis of kernel eigenvectors and benefits significantly from unlabeled data.

References

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *Journal of Machine Learning Research* 7:2399–2434.

Bengio, Y.; Paiement, J.-F.; and Vincent, P. 2003. Out-of-sample Extensions for LLE, Isomap, MDS, Eigenmaps and Spectral Clustering. In *NIPS*.

Blanchard, G.; Massart, P.; Vert, R.; and Zwald, L. 2004. Kernel Projection Machine: A New Tool For Pattern Recognition. In *NIPS*.

Candes, E. J., and Plan, Y. 2007. Near Ideal Model Selection by ℓ_1 Minimization, eprint arxiv:0801.0345.

Castelli, V., and Cover, T. M. 1996. The Relative Value of Labeled and Unlabeled Samples in Pattern Recognition with Unknown Mixing Parameters. *IEEE Transactions on Information Theory* 42(6):2102–2117.

Chapelle, O.; Weston, J.; and Scholkopf, B. 2002. Cluster Kernels for Semi-supervised Learning. In *NIPS*.

Dasgupta, S. 1999. Learning Mixture of Gaussians. In *40th Annual Symposium on Foundations of Computer Science*.

Ratsaby, J., and Venkatesh, S. 1995. Learning From a Mixture of Labeled and Unlabeled Examples with Parametric Side Information. In *COLT*.

Rosasco, L.; Belkin, M.; and Vito, E. D. 2008. Perturbation Results for Learning Empirical Opertors. Technical Report TR-2008-052, Massachusetts Institute of Technology, Cambridge, MA.

Shi, T.; Belkin, M.; and Yu, B. 2008a. Data Spectroscopy: Eigenspace of Convolution Operators and Clustering. Technical report, Dept. of Statistics, Ohio State University.

Shi, T.; Belkin, M.; and Yu, B. 2008b. Data Spectroscopy: Learning Mixture Models using Eigenspaces of Convolution Operators. In *ICML*.

Singh, A.; Nowak, R. D.; and Zhu, X. 2008. Unlabeled Data: Now it Helps Now it Doesn't. In *NIPS*.

Sinha, K., and Belkin, M. 2007. The Value of Labeled

and Unlabeled Examples when the Model is Imperfect. In *NIPS*.

Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B* 58:267–288.

Tropp, J. A. 2004. Greed is Good: Algorithmic Result for Sparse Approximation. *IEEE Trans. Info. Theory* 50(10):2231–2242.

Wainwright, M. 2006. Sharp Thresholds for Noisy and High-dimensional Recovery of Sparsity using ℓ_1-constrained Quadratic Programming. Technical Report TR-709, Dept. of Statistics, U. C. Berkeley.

Zhang, C., and Huang, J. 2006. Model Selection Consistency of Lasso in High Dimensional Linear Regression. Technical report, Dept. of Statistics, Rutgers University.

Zhang, T. 2008. On consistency of feature selection using greedy least square regression. *Journal of Machine Learning Research*.

Zhao, P., and Yu, B. 2006. On Model Selection Consistency of Lasso. *Journal of Machine Learning Research* 7:2541–2563.