Spectral Clustering and Kernel PCA are Learning Eigenfunctions

Yoshua Bengio, Pascal Vincent, Jean-François Paiement
Olivier Delalleau, Marie Ouimet, and Nicolas Le Roux
Département d'Informatique et Recherche Opérationnelle
Centre de Recherches Mathématiques
Université de Montréal
Montréal, Québec, Canada, H3C 3J7
{bengioy,vincentp,paiemeje,delallea,ouimema,lerouxni}@iro.umontreal.ca
http://www.iro.umontreal.ca/∼bengioy
Technical Report 1239,
Département d'Informatique et Recherche Opérationnelle

July 25, 2003

Revised May 6, 2004

Abstract

In this paper, we show a direct equivalence between spectral clustering and kernel PCA, and how both are special cases of a more general learning problem, that of learning the principal eigenfunctions of a kernel, when the functions are from a function space whose scalar product is defined with respect to a density model. This defines a natural mapping for new data points, for methods that only provided an embedding, such as spectral clustering and Laplacian eigenmaps. The analysis hinges on a notion of generalization for embedding algorithms based on the estimation of underlying eigenfunctions, and suggests ways to improve this generalization by smoothing the data empirical distribution.

1 Introduction

Clustering and manifold learning are intimately related: clusters and manifolds both are zones of high density. Up to recently, both tasks have been treated quite separately with different unsupervised learning procedures, but recent work with kernel methods, as well as the results in this paper, are changing that perspective.
Spectral clustering can give very impressive results and has attracted much interest in the last few years (Weiss, 1999; Ng, Jordan and Weiss, 2002). These methods can yield

impressively good results where traditional clustering looking for "round blobs" in the data, such as K-means, would fail miserably. They are based on two main steps: first embedding the data points in a space in which clusters are more "obvious" (using the eigenvectors of a Gram matrix), as seen in Figure 1, and then applying an algorithm to separate the clusters, such as K-means, e.g. as in (Ng, Jordan and Weiss, 2002).

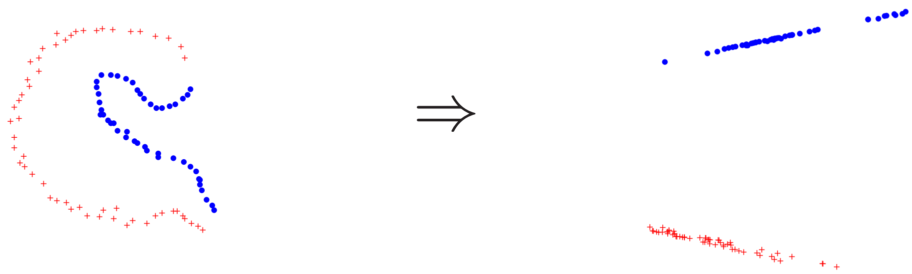

Figure 1: Example of the transformation learned as part of spectral clustering. Input data on the left, transformed data on the right. Colors and cross/circle drawing are only used to show which points get mapped where: the mapping reveals both the clusters and the internal structure of the two manifolds.

One problem with spectral clustering is that the procedure provides a cluster assignment and an embedding for the training points, not for new points. A similar method for dimensionality reduction by spectral embedding has been proposed in (Belkin and Niyogi, 2003), based on so-called Laplacian eigenmaps. Belkin and Niyogi propose to use such transformations in a semi-supervised and transductive setting: the unlabeled test set and the input part of the training set are used to learn a mapping to a more revealing representation, and the transformed training set is used with a supervised learning algorithm.
Kernel PCA is another unsupervised learning method that was proposed earlier and that is based on the simple idea of performing Principal Components Analysis in the feature space of a kernel (Schölkopf, Smola and Müller, 1998). In this paper we study an underlying problem of learning eigenfunctions which allows to draw links between all these methods and propose extensions to them. In particular, we show a direct equivalence between the embedding computed in spectral clustering and the mapping computed with kernel PCA, and how both are special cases of a more general learning problem, that of learning the principal eigenfunctions of a kernel, when the functions are from a function space whose scalar product is defined with respect to a density model. We also show how the formula for extending the embedding to new points can be applied even in the case when the kernel is not positive semi-definite.
A consequence is that a natural mapping is defined, which can be applied to new points, for methods such as spectral clustering and Laplacian eigenmaps for which only an embedding of the training points was available. Another consequence is that the analysis suggests ways to improve the "generalization" obtained on new points, using a smooth of the empirical distribution.

2 Notation and Definitions

Let $D = \{x_1, \dots, x_n\}$ be a data set sampled i.i.d from an unknown distribution with continuous density p and let $\hat{p}$ be the corresponding empirical distribution. Consider the space $\mathcal{H}_q$ of continuous functions f defined everywhere that are square integrable as follows:

$$\int f^2(x)q(x)dx < +\infty$$

where $q(x) \geq 0$ is a weighting function. The scalar product on $\mathcal{H}_q$ is

$$\langle f, g \rangle_q = \int f(x)g(x)q(x)dx.$$

One must note that even though we will use functions for the sake of simplicity, we actually do not work on functions but on equivalence classes: we say two continuous functions f and g belong to the same equivalence class (with respect to q) if and only if $\int (f(x) - g(x))^2 q(x)dx = 0$ (if q is strictly positive, then each equivalence class contains only one function).
We'll consider here two variants of this space. When we choose $q = p$ we obtain the "underlying" Hilbert space $\mathcal{H}_p$. But since p is unknown, we consider an "empirical" Hilbert space $\mathcal{H}_{\hat{p}}$ defined with the weighting function $q = \hat{p}$, for which

$$\langle f, g \rangle_{\hat{p}} = \frac{1}{n}\sum_{i=1}^{n} f(x_i)g(x_i).$$

Let $K(x, y)$ be a symmetric kernel (a 2-argument symmetric function, not necessarily positive semi-definite) with a discrete spectrum, and define an associated linear operator in $\mathcal{H}_q$ as follows:

$$(K_q f)(x) = \int K(x, y)f(y)q(y)dy.$$

If we take $q = \hat{p}$, we obtain the equation

$$(K_{\hat{p}} f)(x) = \frac{1}{n}\sum_{i=1}^{n} K(x, x_i)f(x_i).$$

Note that the above operator application converges (in probability) as $n \to \infty$ to $(K_p f)(x)$ since it is an average and the latter is the corresponding expectation.
In general, the eigensystem for K_q is defined by $K_q f_k = \lambda_k f_k$, which for $K_{\hat{p}}$ rewrites into

$$\frac{1}{n}\sum_{i=1}^{n} K(x_j, x_i)f_k(x_i) = \lambda_k f_k(x_j) \tag{1}$$

for all $x_j \in D$. Note that we adopt the convention that eigenvectors and eigenfunctions have norm 1 in their respective space, and that they are ordered in non-increasing value of the corresponding eigenvalues. Also note that we only care about eigenvectors and eigenfunctions for which the corresponding eigenvalue is not 0.

Let M be the $n \times n$ Gram matrix obtained from D and K by $M_{ij} = K(x_i, x_j)$. Let V be the matrix with the orthonormal eigenvectors v_k of M in its columns, i.e. $V'V = I$ and

$$\sum_{j=1}^{n} K(x_i, x_j) V_{jk} = \hat{\lambda}_k V_{ik}. \tag{2}$$

3 Eigenfunction View of Spectral Embedding Methods

In spectral clustering methods, one often starts from the Gaussian kernel $\tilde{K}$ defined by $\tilde{K}(x, y) = e^{-||x-y||/\sigma^2}$, but a transformation is applied to the corresponding Gram matrix $\tilde{M}$ obtained from $\tilde{K}$ before computing the principal eigenvectors. An example of such transformation (the most successful found in (Weiss, 1999) and the one used in (Ng, Jordan and Weiss, 2002)) is the *divisive normalization*

$$M_{i,j} = \frac{\tilde{M}_{ij}}{\sqrt{S_i S_j}} \tag{3}$$

where $S_i = \sum_j \tilde{M}_{ij}$. Equivalently one can define a normalized kernel K which directly gives rise to M (up to a scaling factor n):

$$K(x, y) = \frac{\tilde{K}(x, y)}{\sqrt{E_x[\tilde{K}(x, y)] E_y[\tilde{K}(x, y)]}}$$

where $E[]$ represents expectation over $\hat{p}$, i.e. an average [1] . In the Laplacian Eigenmaps manifold learning methods (Belkin and Niyogi, 2003), the embedding is obtained from the vectors v that solve a generalized eigenproblem $(S - M)v = \lambda S v$, where S is the diagonal matrix formed by the S_i. This is equivalent to finding the principal eigenvectors of M with the above divisive normalization. This has been shown in (Weiss, 1999) (Normalization Lemma 1) when referring to an earlier spectral clustering method (Shi and Malik, 1997) that also considers the same generalized eigenproblem.

In both spectral clustering and Laplacian eigenmaps, one then obtains an embedding $e_k(x_i)$ for the training points x_i from the principal eigenvectors $v_k = (V_{1k}, V_{2k}, \ldots)'$ of M:

$$e_k(x_i) = V_{ik}.$$

Proposition 1 *Using the above notation, if K is positive semi-definite then the embedding $e_k(x_i)$ obtained with spectral clustering and Laplacian eigenmaps is given by the formula $e_k(x_i) = \frac{1}{\sqrt{n}} f_k(x_i)$, where f_k is the k-th principal eigenfunction of $K_{\hat{p}}$.*

The proof of this proposition can be found in the technical report (Bengio, Vincent and Paiement, 2003), but this proposition is generalized by Theorem 1 proved below.

[1] Better embeddings are usually obtained if we define $S_i = \sum_{j \neq i} \tilde{M}_{ij}$: this alternative normalization can also be obtained with a slightly different kernel.

4 Eigenfunction View of Kernel PCA

Kernel PCA generalizes the Principal Components Analysis approach to non-linear transformations using the kernel trick (Schölkopf, Smola and Müller, 1998), working in the "feature space" $\phi(x)$ of a positive semi-definite kernel written implicitly as a dot product in that space. The principal components are the eigenvectors u_k of the covariance matrix C of the data in feature space: $C = E_x[\phi(x)\phi(x)']$, where $\phi(x)$ must be chosen such that it has zero mean ($E_x[\phi(x)] = 0$, where expectation is over the empirical distribution, in practice). This can be achieved by choosing a normalization of the kernel that centers it in feature space. Starting from an unnormalized kernel $\tilde{K}$ with unnormalized feature space $\tilde{\phi}$, we obtain $\phi(x) = \tilde{\phi}(x) - E_x[\tilde{\phi}(x)]$ with the following *additive normalization*:

$$\begin{aligned} K(x,y) &= (\tilde{\phi}(x) - E_x[\tilde{\phi}(x)]) \cdot (\tilde{\phi}(y) - E_y[\tilde{\phi}(y)]) \\ &= \tilde{K}(x,y) - E_x[\tilde{K}(x,y)] - E_y[\tilde{K}(x,y)] + E_x[E_y[\tilde{K}(x,y)]]. \end{aligned}$$

Once the Gram matrix M is formed from this kernel K and data D, the principal eigenvectors/eigenvalues $(v_k, \hat{\lambda}_k)$ of M are computed. As shown in (Schölkopf, Smola and Müller, 1998), the principal eigenvectors u_k of C ($k \leq n$) are linked to the eigenvectors v_k of M through $u_k = \frac{1}{\sqrt{\hat{\lambda}_k}} \sum_{i=1}^{n} V_{ik}\phi(x_i)$.

The projection $\pi_k(x)$ of a test point x on the k-th principal component can then be obtained as follows:

$$\pi_k(x) = u_k \cdot \phi(x) = \frac{1}{\sqrt{\hat{\lambda}_k}} \sum_{i=1}^{n} V_{ik} K(x_i, x). \tag{4}$$

Proposition 2 *Let $\pi_k(x)$ be the test point projection (eq. 4) on the k-th principal component obtained by kernel PCA with a normalized positive semi-definite discrete spectrum kernel $K(x,y)$. Then*

$$\pi_k(x) = \sqrt{\lambda_k} f_k(x) \tag{5}$$

where λ_k is the k-th eigenvalue of $K_{\hat{p}}$ and f_k is the corresponding eigenfunction.

The proof of this proposition can be found in the technical report (Bengio, Vincent and Paiement, 2003), but this proposition is generalized by Theorem 1 proved below. The similarity between equation 4 (Schölkopf, Smola and Müller, 1998) and the Nyström approximation (Baker, 1977) of the eigenfunctions of K_p (as in eq. 6) has already been pointed out in (Williams and Seeger, 2000).

5 Extension to Kernels with Negative Eigenvalues

Users of spectral clustering and Laplacian eigenmaps may want to use a kernel that is not guaranteed to be positive semi-definite. There are also several other spectral embedding methods which do not guarantee the positive definitness of the Gram matrix M (i.e. of the corresponding kernel K), such as multi-dimensional scaling (MDS) (Cox and Cox, 1994) and ISOMAP (Tenenbaum, de Silva and Langford, 2000). We would like to apply a formula such as eq. 5 to such kernels. The next theorem gives a justification for using such formulae even in the case when the kernel may have negative eigenvalues. If those negative eigenvalues are small (w.r.t. the largest positive eigenvalues), one would just discard them. If they are large, it may actually prove useful to use the corresponding eigenvectors (or eigenfunctions in the following) in order to discover interesting features, as shown in (Laub and Müller, 2003).

Theorem 1 *The eigenfunctions f_k of $K_{\hat{p}}$ (not necessarily positive semi-definite) associated to non-zero eigenvalues are of the form*

$$f_k(x) = \frac{\sqrt{n}}{\hat{\lambda}_k} \sum_{i=1}^{n} V_{ik} K(x, x_i) \tag{6}$$

where the matrix V has the orthonormal eigenvectors v_k of the Gram matrix M in its columns, with eigenvalues $\hat{\lambda}_k$. The eigenvalue λ_k of f_k is $\lambda_k = \hat{\lambda}_k / n$.
For $x_i \in D$ these functions coincide with the corresponding eigenvectors, in the sense that $f_k(x_i) = \sqrt{n} V_{ik}$.

Proof
First, these f_k coincide with the eigenvectors of M at $x_i \in D$. For f_k defined by eq. 6:

$$f_k(x_i) = \frac{\sqrt{n}}{\hat{\lambda}_k} \sum_{j=1}^{n} V_{jk} K(x_i, x_j) = \frac{\sqrt{n}}{\hat{\lambda}_k} \hat{\lambda}_k V_{ik} = \sqrt{n} V_{ik} \tag{7}$$

so that they form an orthonormal family in $\mathcal{H}_{\hat{p}}$:

$$\langle f_j, f_k \rangle_{\hat{p}} = \frac{1}{n} \sum_{i=1}^{n} f_j(x_i) f_k(x_i) = \sum_{i=1}^{n} V_{ij} V_{ik} = \delta_{j,k}. \tag{8}$$

Then for any $x \in D$:

$$(K_{\hat{p}} f_k)(x) = \frac{1}{n} \sum_{i=1}^{n} K(x, x_i) f_k(x_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} K(x, x_i) V_{ik} = \frac{\hat{\lambda}_k}{n} f_k(x) \tag{9}$$

which shows that f_k is an eigenfunction of $K_{\hat{p}}$ with eigenvalue $\lambda_k = \hat{\lambda}_k / n$. $\square$
It is important to note that from eq. 6, eq. 9 holds for any x, even if $x \notin D$. This allows the choice of eq. 6 to be justified by the following proposition:

Proposition 3 *Let K a symmetric kernel with discrete spectrum. Then the operator G_n in $\mathcal{H}_p$ defined by*

$$G_n f = \frac{1}{n} \sum_{i=1}^{n} K(\cdot, x_i) f(x_i)$$

has for eigenfunctions associated to non-zero eigenvalues the functions of the form:

$$f = \frac{f_k}{||f_k||_{\mathcal{H}_p}}$$

with eigenvalues λ_k, where f_k and λ_k are those of Theorem 1, and $|| \cdot ||_{\mathcal{H}_p}$ is the norm in $\mathcal{H}_p$.

Proof
Let f be an eigenfunction of G_n with eigenvalue $\lambda \neq 0$, and

$$f' = \frac{f}{\sqrt{n}||f||_{\mathcal{H}_{\hat{p}}}}. \tag{10}$$

Then for any $x_j \in D$:

$$(G_n f')(x_j) = \frac{1}{n}\sum_{i=1}^{n} K(x_j, x_i) f'(x_i) = \lambda f'(x_j)$$

which means the vector v whose j-th element is $f'(x_j)$ is an eigenvector v_k of the Gram matrix M with eigenvalue $\hat{\lambda}_k = n\lambda$ ($||v|| = 1$ follows from the normalization in eq. 10). We can thus write for any x (not necessarily in D):

$$(G_n f')(x) = \frac{1}{n}\sum_{i=1}^{n} K(x, x_i) V_{ik} = \frac{\hat{\lambda}_k}{n} f'(x)$$

which proves, using eq. 6, that $f'(x) = \frac{f_k(x)}{\sqrt{n}}$. It follows immediately (using $||f||_{\mathcal{H}_p} = 1$) that

$$f(x) = ||f||_{\mathcal{H}_{\hat{p}}} f_k(x) = \frac{f_k(x)}{||f_k||_{\mathcal{H}_p}}. \quad \square$$

An interesting consequence of the above results is that spectral embedding methods which only provided an embedding for the training examples can be extended in order to provide an embedding for out-of-sample examples, using formulae 6 and 7 from the above theorem, at least for the dimensions that correspond to positive eigenvalues. It introduces a notion of *generalization* for embedding procedures: in that context the optimal out-of-sample embedding is the one that would be obtained at a new point x if we had access to an infinite amount of data to estimate the eigenfunctions of K_p in $\mathcal{H}_p$ (the Hilbert space generated with the true generating distribution p of the data as weighting function in the scalar product).

The theoretical justification for this out-of-sample extension stems from several strong results on the convergence and the stability of the eigendecomposition of the Gram matrix M. First of all, (Baker, 1977; Williams and Seeger, 2000) show that the eigenvalues and eigenvectors (when the eigenvalues are distinct) of M converge as $n \to \infty$ and they converge to the eigenvalues and eigenfunctions of the linear operator defined by K_p in $\mathcal{H}_p$. Second, several researchers have studied the stability of this eigendecomposition with respect to the data sample. (Ng, Jordan and Weiss, 2002) have studied the stability of the principal eigenvectors (and thus of spectral clustering) with respect to perturbations of the Gram matrix and the effect of nearby eigenvalues on that stability. (Shawe-Taylor, Cristianini and Kandola, 2002) introduced the use of concentration inequalities to bound the sampling variability of eigenvalue estimation, and (Shawe-Taylor and Williams, 2003) push these results further to give bounds on the kernel PCA reconstruction error, using the linear operator eigensystem used here, i.e. with $q = \hat{p}$. These results also bound the error on the estimation of the subspaces associated with the first k eigenvalues. The formula for extending the embedding to a new data point is the same as the Nyström

formula (Baker, 1977), which has been used successfully for reducing the computational cost of kernel methods by restricting the difficult computations to a subset of the examples (the equivalent of the training set, here).
When we perform the PCA or kernel PCA projection on an out-of-sample point we are taking advantage of the above convergence and stability properties in order to trust that a principal eigenvector of the empirical covariance matrix estimates well a corresponding eigenvector of the true covariance matrix. The same reasoning allows us to apply eq. 6 on an out-of-sample point x. More precisely, we can use Proposition 3 to give an interpretation of the use of eq. 6 to approximate the eigenfunctions of K_p in $\mathcal{H}_p$: first we approximate the operator K_p by G_n (justified by the law of large numbers), then the eigenfunctions of G_n by the f_k, which is again coherent thanks to the law of large numbers, which states $||f_k||_{\mathcal{H}_p}$ is close (with good probability) to $||f_k||_{\mathcal{H}_{\hat{p}}} = 1$ when n is large.

6 Improving Out-of-Sample Generalization of Spectral Embedding

Here, we consider a related question to explore the notion of generalization to new cases with embedding algorithms: *can one get better generalization when using for q a smoother distribution than the empirical distribution?* Indeed, this may yield smoother eigenfunctions, which could be obtained using a rich enough class of functions (such as neural networks) and optimizing them numerically. A training criterion for this purpose is provided by the following proposition, proven in (Bengio, Vincent and Paiement, 2003).

Proposition 4 *If f_i, with i from 1 to $m-1$, are the principal $m-1$ eigenfunctions of the linear operator K_p in $\mathcal{H}_p$, then the function g which minimizes the expected value of $(K(x,y) - g(x)g(y) - \sum_{i=1}^{m-1} \lambda_i f_i(x) f_i(y))^2$ over the joint distribution $p(x)p(y)$ gives the m-th eigenfunction $f_m = g/\sqrt{\lambda_m}$ where $\lambda_m = ||g||^2_{\mathcal{H}_p}$ is its eigenvalue.*

Another simpler solution is to use the smoother density (e.g. obtained from a manifold Parzen windows estimator (Vincent and Bengio, 2003)) to sample a larger data set which will be used to estimate the underlying eigenfunctions.
Application and experimental validation of the previous ideas will be studied elsewhere.

7 Conclusion

In this paper we have established a clear equivalence between the spectral embedding methods used in spectral clustering and Laplacian eigenmaps with the projection computed by the kernel PCA method. Both types of methods are found to estimate eigenfunctions of a linear operator associated with the kernel and with the data. Because previous results have shown how the principal eigenfunctions converge as the amount of data increases, it makes sense to use these eigenfunctions to extend the spectral embedding methods to generalize to new data points. It also makes sense to seek better estimators of these eigenfunctions, in order to obtain a better generalization of the embedding to out-of-sample points.

References

Baker, C. (1977). *The numerical treatment of integral equations*. Clarendon Press, Oxford. 5, 7, 8

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396. 2, 4

Bengio, Y., Vincent, P., and Paiement, J. (2003). Learning eigenfunctions of similarity: Linking spectral clustering and kernel PCA. Technical Report 1232, Département d'informatique et recherche opérationnelle, Université de Montréal. 4, 5, 8

Cox, T. and Cox, M. (1994). *Multidimensional Scaling*. Chapman & Hall, London. 5

Laub, J. and Müller, K.-R. (2003). Feature discovery: unraveling hidden structure in non-metric pairwise data. Technical report, Fraunhofer FIRST.IDA, Germany. 6

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On spectral clustering: analysis and an algorithm. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA. MIT Press. 1, 2, 4, 7

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319. 2, 5

Shawe-Taylor, J., Cristianini, N., and Kandola, J. (2002). On the concentration of spectral properties. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*. MIT Press. 7

Shawe-Taylor, J. and Williams, C. (2003). The stability of kernel principal components analysis and its relation to the process eigenspectrum. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*. MIT Press. 7

Shi, J. and Malik, J. (1997). Normalized cuts and image segmentation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 731–737. 4

Tenenbaum, J., de Silva, V., and Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323. 5

Vincent, P. and Bengio, Y. (2003). Manifold parzen windows. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, Cambridge, MA. MIT Press. 8

Weiss, Y. (1999). Segmentation using eigenvectors: a unifying view. In *Proceedings IEEE International Conference on Computer Vision*, pages 975–982. 1, 4

Williams, C. and Seeger, M. (2000). The effect of the input density distribution on kernel-based classifiers. In *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann. 5, 7