

Provided for non-commercial research and education use.  
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

## Pattern Recognition

journal homepage: [www.elsevier.com/locate/pr](http://www.elsevier.com/locate/pr)

## Approximately harmonic projection: Theoretical analysis and an algorithm

Binbin Lin<sup>a,\*</sup>, Xiaofei He<sup>a</sup>, Yuan Zhou<sup>a</sup>, Ligang Liu<sup>b,a</sup>, Ke Lu<sup>c</sup><sup>a</sup> State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, China<sup>b</sup> Department of Mathematics, Zhejiang University, Hangzhou 310027, China<sup>c</sup> School of Computer Science and Engineering, University of Electronic Science & Technology of China, Chengdu 610054, China

## ARTICLE INFO

## Article history:

Received 5 August 2009

Received in revised form

14 April 2010

Accepted 6 May 2010

## Keywords:

Manifold learning

Dimensionality reduction

Linear projection

Harmonic function

## ABSTRACT

Manifold learning have attracted considerable attention over the last decade. The most frequently used functional is the  $l^2$ -norm of the gradient of the function. In this paper, we consider the linear manifold learning problem by minimizing this functional with appropriate constraint. We provide theoretical analysis on both the functional and the constraint, which shows the affine hulls of the manifold and the connected components are essential to linear manifold learning problem. Based on the theoretical analysis, we introduce a novel linear manifold learning algorithm called approximately harmonic projection (AHP). Unlike canonical linear methods such as principal component analysis, our method is sensitive to the connected components. This makes our method especially applicable to data clustering. We conduct several experimental results on three real data sets to demonstrate the effectiveness of our proposed method.

© 2010 Elsevier Ltd. All rights reserved.

## 1. Introduction

In many cases of interest in machine learning and data analysis, the data is usually represented as points in a very high dimensional space. However, intuitively the data may be generated by structured systems with much fewer degrees of freedom. Various researchers have considered the case when the data is on or near a submanifold of an ambient space. One hopes to estimate intrinsic properties of the manifold from the data points. These problems are typically referred to as *manifold learning*.

The typical work on nonlinear manifold learning includes Locally Linear Embedding [1], ISOMAP [2], Laplacian Eigenmaps [3] and MVU [4]. Some linear manifold learning methods include PCA, LPP [5] and NMF [6]. Belkin and Niyogi propose to use the Dirichlet integral which is the  $l^2$ -norm of the gradient of the function. The Euler–Lagrange equation of the functional is a harmonic equation. Thus the solution of minimizing this functional is a harmonic function. Harmonic function plays an important role in manifold learning. In fact, both isometric mapping on the manifold and linear mapping on ambient Euclidean space are harmonic mappings. Isometry is very difficult (even impossible in many cases) to achieve in real world. An alternative way is to relax this problem to find a harmonic mapping, since isometric is also harmonic. LLE aims to find a locally linear embedding. It actually finds a locally harmonic embedding.

It would be important to distinguish the functions on manifold and functions on ambient space. For a learning task, we usually

hope our function is defined on ambient space. However, all the above mentioned nonlinear manifold learning methods consider the case that the function is defined on the manifold. Compared with nonlinear methods, linear methods have the advantage that the function is defined on ambient space. One criterion for the ideal function is that it should be *good* on both the manifold and ambient space. Based on the functional, since linear functions are always harmonic on ambient space, one hopes then they are as harmonic as possible on the manifold.

For many real world problems, the data manifold is not only one global manifold, precisely one connected component, but numbers of connected components, and the data set are always sparse. Isomap and LLE considered the problem that the manifold has only one connected component. They are trying to “expand” the manifold while keep the geometrical structure. However, there are few discussion on multiple connected components case. Our paper is a trial in this point, and we found one interesting property that we can separate parallel affine hulls. Though the result of nonlinear methods are accurate, the requirement of the graph model is much higher than linear methods. Our approach need approximate the affine hull of the manifold which appears less sensitive to the graph model than Laplacian eigenmaps.

In this paper, we consider the linear methods based on the harmonic framework. We provide theoretical analysis for general linear harmonic methods based on the Dirichlet integral. We show the geometrical meaning of the optimal projections which are closely related to the function defined on ambient space and the affine hull of the manifold. We will show the affine hull of the manifold is essential for linear harmonic methods. PCA happens to be a very good choice for finding the affine hull of the whole manifold. The affine hull of each connected component of the

\* Corresponding author. Tel.: +86 571 88206681; fax: +86 571 88206680.  
E-mail address: [binbinlin@zju.edu.cn](mailto:binbinlin@zju.edu.cn) (B. Lin).

manifold is very important. Global methods like PCA consider only the properties of the whole manifold, whereas linear harmonic methods consider the connected components of the manifold, which is especially important for data clustering. Based on our theoretical analysis, we propose a new linear harmonic method which we call approximately harmonic projections (AHP). The projections are obtained by approximating the Dirichlet integral.

## 2. Theoretical analysis

In this section, we provide theoretical analysis for linear harmonic methods based on the Dirichlet integral.

Let  $\mathcal{M}$  be a smooth, compact,  $m$ -dimensional Riemannian manifold. If the manifold is embedded in  $\mathbb{R}^N$ , the metric is induced by the standard metric on  $\mathbb{R}^N$ . We are trying to find a *good* map from manifold to much lower dimensional Euclidean space. Let  $f$  be a twice differential function,  $f : \mathcal{M} \rightarrow \mathbb{R}$ . One of the most popular criteria is the one preserving locality [3]:

$$\operatorname{argmin}_f \int_{\mathcal{M}} \|\nabla f\|^2 \quad \text{s.t. } \|f\|_{L^2(\mathcal{M})} = 1 \quad (1)$$

### 2.1. Linear projection

Let  $f : \mathbb{R}^N \rightarrow \mathbb{R}$  be continuously differentiable. If  $x^1, \dots, x^N$  are Euclidean coordinates, corresponding to the orthonormal basis  $e_1, \dots, e_N$ , then the gradient [7] is

$$\nabla f = \sum_{i=1}^N \frac{\partial f}{\partial x^i} e_i$$

If we restrict  $f$  on  $\mathcal{M}$ , denote it by  $f_{\mathcal{M}}$ , the gradient of  $f_{\mathcal{M}}$  may be quite different which depends on the metric of the manifold. Let  $u^1, \dots, u^m$  be the local coordinates of the manifold and  $g$  be the Riemannian metric of the manifold, then we have

$$\nabla f_{\mathcal{M}} = \sum_{k,l} g^{kl} \frac{\partial f_{\mathcal{M}}}{\partial u^k} \frac{\partial}{\partial u^l}$$

where  $\{\partial/\partial u^l\}$  is a basis for the tangent space of  $\mathcal{M}$ .

Consider a linear function defined on an ambient space. Let  $f(\mathbf{x}) = \mathbf{a}^T \mathbf{x} = \sum a_i x^i(\mathbf{x})$ , where  $\mathbf{a}$  is an  $N \times 1$  column vector. Linear projection is linear combination of coordinates projection. Thus, it is necessary to measure the goodness for each coordinate projection. In the following, we assume  $f$  is a linear function on  $\mathbb{R}^N$  without specific notation.

As  $f$  is linear on  $\mathbb{R}^N$ ,  $\nabla f$  is a constant vector everywhere:

$$\nabla f = \sum_{i=1}^N a_i e_i = \mathbf{a}$$

As  $f$  is defined on the ambient space, we should represent the gradient by the coordinates of the ambient space. By the linearity of the gradient operator, we have

$$\nabla f_{\mathcal{M}} = \sum_{i=1}^N a_i \nabla x_{\mathcal{M}}^i$$

thus the functional in problem (1) can be rewritten as

$$\int_{\mathcal{M}} \langle \nabla f_{\mathcal{M}}, \nabla f_{\mathcal{M}} \rangle = \sum_{i,j=1}^N a_i a_j \int_{\mathcal{M}} \langle \nabla x_{\mathcal{M}}^i, \nabla x_{\mathcal{M}}^j \rangle = \mathbf{a}^T \mathbf{H} \mathbf{a} \quad (2)$$

where  $H$  is a symmetric matrix whose entry  $H_{ij} = \int_{\mathcal{M}} \langle \nabla x_{\mathcal{M}}^i, \nabla x_{\mathcal{M}}^j \rangle$

The constraint can be computed directly:

$$\int_{\mathcal{M}} f_{\mathcal{M}}^2 = \int_{\mathcal{M}} \sum_{i=1}^N (a_i x_{\mathcal{M}}^i)^2 = \sum_{i,j=1}^N a_i a_j \int_{\mathcal{M}} x_{\mathcal{M}}^i x_{\mathcal{M}}^j = \mathbf{a}^T \mathbf{C} \mathbf{a} \quad (3)$$

where  $C$  is a symmetric matrix whose entry  $C_{ij} = \int_{\mathcal{M}} x_{\mathcal{M}}^i x_{\mathcal{M}}^j$ .

Thus this problem becomes a generalized eigenvector problem:

$$\mathbf{H} \mathbf{a} = \lambda \mathbf{C} \mathbf{a} \quad (4)$$

Here  $H$  and  $C$  play a central role in this problem. Intuitively,  $H$  measures the *smoothness* of the coordinate projection, and  $C$  measures the *representativeness* of the coordinate projection. In the following subsections we give an analysis on both the functional and the constraint.

### 2.2. Affine manifold and affine hull

We will first study the property when the functional achieves its minimum value, i.e.,  $\int_{\mathcal{M}} \|\nabla f_{\mathcal{M}}\|^2 = 0$ . We find it is closely related to the affine hull of the manifold.

Here we give some basic concepts of the *affine manifold* and the *affine hull* [8].

**Definition 2.1.** An affine subspace, or affine manifold, is a set  $V$  such that the (affine) line  $\{\alpha \mathbf{x} + (1-\alpha) \mathbf{x}' : \alpha \in \mathbb{R}\}$  is entirely contained in  $V$  whenever  $\mathbf{x}$  and  $\mathbf{x}'$  are in  $V$  (Note that a single point is an affine manifold).

Take  $\mathbf{v} \in V$ , it is easy to show that  $V - \{\mathbf{v}\}$  is a subspace of  $\mathbb{R}^N$ , which is independent of the particular  $\mathbf{v}$ ; denote it by  $V_0$ . Thus an affine manifold  $V$  is nothing but the translation of some vector space  $V_0$ , sometimes called the direction(-subspace) of  $V$ ; we will also say that  $V_0$  and  $V$  are *parallel*. One can therefore speak of the *dimension* of an affine manifold; it is just the dimension of  $V_0$ .

**Definition 2.2.** To any nonempty set  $S \subset \mathbb{R}^N$ , we can associate the intersection of all affine manifolds containing  $S$ . This gives the affine manifold generated by  $S$ , denoted  $\operatorname{aff}(S)$ : the affine hull of  $S$ .

We should point out an intersection of affine manifolds is still an affine manifold. For the  $\subset$ -relation,  $\operatorname{aff}(S)$  is the smallest affine manifold containing  $S$ . It is not difficult to see the affine hull of certain subset is unique.

Here we give some examples of affine hull. The affine hull of a set of two different points is the line through them. The affine hull of one circle is the plane going through it. For two skew lines (see Fig. 2), the affine hull is a 3-dimensional affine space.

With the definition of the affine hull, we have the following lemma:

**Lemma 2.1.** Let  $S$  be a nonempty subset of  $\mathbb{R}^N$  and denote its affine hull by  $\operatorname{aff}(S)$ . The following statements are equivalent:

1.  $f_S = \text{const.}$
2.  $f_{\operatorname{aff}(S)} = \text{const.}$
3.  $\nabla f \perp \operatorname{aff}(S)$ .

**Proof.** 1  $\Rightarrow$  2. Take  $\mathbf{x}_0 \in S$ , let  $S_0 = S - \mathbf{x}_0$ , and denote the subspace generated by  $S_0$  by  $\operatorname{lin}(S_0)$ . We have  $\operatorname{aff}(S) = \mathbf{x}_0 + \operatorname{lin}(S_0)$ . Assume  $\dim(\operatorname{lin}(S_0)) = l$ , then we can find  $l$  points,  $\mathbf{x}_1, \dots, \mathbf{x}_l \in S$  such that  $\{\mathbf{x}_i - \mathbf{x}_0\}$  is a basis of  $\operatorname{lin}(S_0)$ . This can be obtained by the definition of affine hull. Thus to every point  $\mathbf{y} \in \operatorname{aff}(S)$ , there exist  $\alpha_1, \dots, \alpha_m \in \mathbb{R}$ , such that  $\mathbf{y} - \mathbf{x} = \sum \alpha_i (\mathbf{x}_i - \mathbf{x})$ . Then

$$f(\mathbf{y}) = f(\mathbf{x}) + \sum \alpha_i (f(\mathbf{x}_i) - f(\mathbf{x})) = f(\mathbf{x}).$$

Thus  $f$  is constant on  $\operatorname{aff}(S)$ .

2  $\Rightarrow$  3. For arbitrary  $\mathbf{x}, \mathbf{y} \in \operatorname{aff}(S)$ , as  $\nabla f = \mathbf{a}$

$$\langle \nabla f, \mathbf{y} - \mathbf{x} \rangle = \mathbf{a}^T (\mathbf{y} - \mathbf{x}) = f(\mathbf{y}) - f(\mathbf{x}) = 0.$$

Thus  $\nabla f \perp \operatorname{aff}(S)$ .

3  $\Rightarrow$  1. If  $\nabla f \perp \operatorname{aff}(S)$ , by the previous equation, for arbitrary  $\mathbf{x}, \mathbf{y} \in \operatorname{aff}(S)$ ,  $f(\mathbf{x}) = f(\mathbf{y})$ . Thus  $f$  is constant on  $\operatorname{aff}(S)$ . And consequently it is constant on  $S$ .  $\square$

### 2.3. Trivial solution and affine hull

When the functional vanishes,  $f$  is constant on the manifold. We call it trivial solution or trivial projection. By Lemma 2.1, we have the following proposition:

**Proposition 2.1.** *Let  $\mathcal{M}$  be an  $m$ -dimensional manifold embedded in  $\mathbb{R}^N$ , and denote its affine hull by  $\text{aff}(\mathcal{M})$ . Then the following statements are equivalent:*

1.  $f_{\mathcal{M}} = \text{const}$ .
2.  $f_{\text{aff}(\mathcal{M})} = \text{const}$ .
3.  $\nabla f \perp \text{aff}(\mathcal{M})$ .

When  $f_{\mathcal{M}} = \text{const}$ ,  $\int_{\mathcal{M}} \|\nabla f_{\mathcal{M}}\|^2 = 0$ . So these are trivial optimal solutions. The statement (3) in the above proposition shows the geometric meaning of the trivial optimal solution. Please see Fig. 1. In this case, the whole manifold will be projected to a single point. From another point of view, this is the smoothest projection which is indeed the functional try to find.

On the other hand, the affine hull of the manifold is essential for linear projection. Next we show the functional totally depends on the affine hull of the manifold. Take  $\mathbf{x} \in \text{aff}(\mathcal{M})$ , then  $\text{lin}(\mathcal{M}) = \text{aff}(\mathcal{M}) - \{\mathbf{x}\}$  is a subspace of  $\mathbb{R}^N$ . We expand the basis of  $\text{lin}(\mathcal{M})$  to an orthogonal basis of  $\mathbb{R}^N$ . Let  $\{x^1, \dots, x^N\}$  be the new coordinates of  $\mathbb{R}^N$ . Then  $\nabla f$  can be rewritten as

$$\nabla f_{\mathcal{M}} = \sum_{i=1}^N a_i \nabla x_{\mathcal{M}}^i.$$

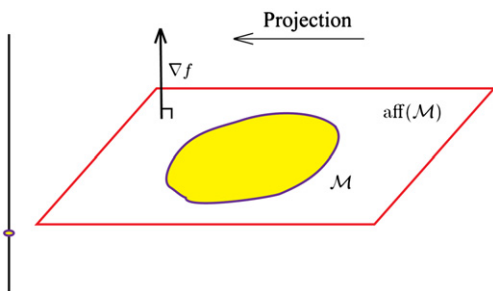
Since  $x_{\mathcal{M}}^i = \text{const}$ , then  $\nabla x_{\mathcal{M}}^i = 0$ ,  $i=1, \dots, N$ . Thus we have

$$\nabla f_{\mathcal{M}} = \sum_{i=1}^N a_i \nabla x^i = \sum_{i=1}^l a_i \nabla x_{\mathcal{M}}^i.$$

Therefore the functional has nothing to do with the orthogonal space to the affine hull. And this conclusion is independent to the topology and number of connected components of the manifold. So it is sufficient to find the projections in the affine hull of the manifold.

Next we prove the problem (4) is well posed in  $\text{aff}(\mathcal{M})$ . We only have to prove the constraint is positive definite. When  $\int_{\mathcal{M}} f_{\mathcal{M}}^2 = 0$ ,  $f_{\mathcal{M}} = 0$ . By Lemma 2.1, we have  $f_{\text{aff}(\mathcal{M})} = 0$ . Thus it is positive definite.

For dimension, we have the following relationship:  $\dim \mathcal{M} \leq \dim \text{aff}(\mathcal{M}) \leq N$ . However, the dimension of the affine hull is independent to the dimension of the manifold. In the extreme case, one dimensional manifold would have very high dimensional affine hull whose dimension even equals to that of the ambient space.



**Fig. 1.** Trivial optimal projection. The optimal function projects the affine hull of the manifold to one point. In this case the gradient of the function on ambient space is orthogonal to the affine hull.

### 2.4. Multiple connected components and clustering

We have shown that the functional totally depends on the affine hull. One natural question is whether there exists nontrivial optimal projection in the affine hull. If  $\mathcal{M}$  has only one connected component, the answer is no. When the manifold has only one connected component, the gradient vanishing implies that the linear function is constant on the whole manifold.

Next we will consider the case when the manifold has multiple connected components. If there exists parallel affine manifolds containing the manifold, we have the following theorem:

**Theorem 2.1.** *Suppose the  $m$ -dimensional manifold  $\mathcal{M}$  contains  $k$  connected components,  $\mathcal{M} = \{\mathcal{M}_1, \dots, \mathcal{M}_k\}$ , and is embedded in  $\mathbb{R}^N$ . Denote their affine hulls by  $\text{aff}(\mathcal{M}_1), \dots, \text{aff}(\mathcal{M}_k)$ , and the corresponding linear spaces by  $\text{lin}(\mathcal{M}_1), \dots, \text{lin}(\mathcal{M}_k)$ . Then the following statements are equivalent:*

1.  $\int_{\mathcal{M}} \|\nabla f_{\mathcal{M}}\|^2 = 0$ .
2.  $f_{\text{aff}(\mathcal{M}_i)} = \text{const}$ ,  $i = 1, 2, \dots, k$ .
3.  $\nabla f \perp \text{lin}(\mathcal{M}_1) \oplus \dots \oplus \text{lin}(\mathcal{M}_k)$ .

**Proof.** We have,

$$\begin{aligned} \int_{\mathcal{M}} \|\nabla f_{\mathcal{M}}\|^2 &= 0 \\ \iff \int_{\mathcal{M}_i} \|\nabla f_{\mathcal{M}_i}\|^2 &= 0 \\ \iff f_{\mathcal{M}_i} = c_i = \text{const}, &\text{ by Lemma 2.1} \\ \iff f_{\text{aff}(\mathcal{M}_i)} = c_i = \text{const}, &\text{ by Lemma 2.1} \\ \iff \nabla f \perp \text{aff}(\mathcal{M}_i), &\text{ since } \text{aff}(\mathcal{M}_i) // \text{lin}(\mathcal{M}_i) \\ \iff \nabla f \perp \text{lin}(\mathcal{M}_i), &\text{ by the definition of affine hull} \\ \iff \nabla f \perp \text{lin}(\mathcal{M}_1) \oplus \dots \oplus &\text{lin}(\mathcal{M}_k). \quad \square \end{aligned}$$

Fig. 2 is an illustration. We need to point out these solutions satisfy the constraint, except for one trivial case  $f=0$ . And  $\int_{\mathcal{M}} \|\nabla f_{\mathcal{M}}\|^2 = 0$  is equivalent to  $\lambda = 0$ , where  $\lambda$  is the eigenvalue of Eq. (4). This can be proved by the fact that  $C$  is positive definite in  $\text{aff}(\mathcal{M})$ .

Denote the multiplicity of the zero eigenvalue by  $s$ , then by the third statement, we have the following equation about the dimension:

$$\dim \text{aff}(\mathcal{M}) = s + \dim \text{lin}(\mathcal{M}_1) \oplus \dots \oplus \text{lin}(\mathcal{M}_k)$$

Thus,

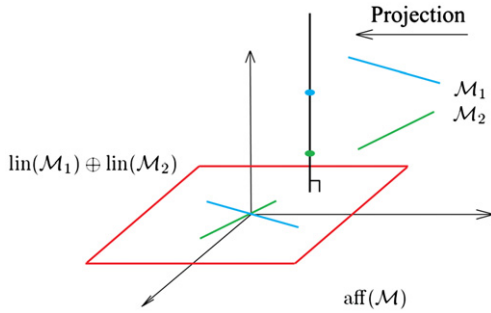
$$\dim \text{lin}(\mathcal{M}_1) \oplus \dots \oplus \text{lin}(\mathcal{M}_k) = \dim \text{aff}(\mathcal{M}) - s \tag{5}$$

The right-hand side is clear and it can be computed from data points. For the left-hand side, since

$$\dim \text{aff}(\mathcal{M}_i) \leq \dim \text{lin}(\mathcal{M}_1) \oplus \dots \oplus \text{lin}(\mathcal{M}_k)$$

it is an upper bound of the dimension of the affine hull of each connected component. Consider Fig. 2,  $\dim \text{aff}(\mathcal{M}) = 3$ ,  $s = 1$ , then we have  $\dim \text{lin}(\mathcal{M}_1) \oplus \dots \oplus \text{lin}(\mathcal{M}_k) = 2$ . It means each connected component can be contained by a two dimensional affine manifold. The affine manifolds may be either parallel or overlapped. This provides an estimation for the dimension of the affine hull of every single connected component. It also reflects the local linear structure of the manifold. Thus linear methods can also discover some local geometry of the manifold. This cannot be done for many other global linear methods such as PCA.

The third statement of this theorem describes the connection between the affine hull of the manifold and the function on the ambient space when it is optimal. In this case, the gradient of the function on ambient space is orthogonal to the affine hull of the connected components of the manifold. In some sense, each connected component with the overlapping affine hull will be



**Fig. 2.** Nontrivial optimal projection. The optimal function projects two skew lines to two different points. In this case the gradient of the projection on ambient space (here is  $\text{aff}(\mathcal{M})$ ) is orthogonal to the plane they span.

collapsed by the optimal projection. Usually data points sampled from different component correspond to different objects, the optimal projection will separate them very well. Thus this functional is very suitable for clustering when there are multiple near parallel connected components. For the affine hull of the connected components overlaps, it cannot guarantee separate different connected components. This is the limitation of linear methods.

2.5. Submanifold of affine manifold

What if  $\dim \text{lin}(\mathcal{M}_1) \oplus \dots \oplus \text{lin}(\mathcal{M}_k) = \dim \text{aff}(\mathcal{M})$ ? By Eq. (5), there is no optimal projection in the affine hull. If the manifold is a compact domain of the affine hull (see Fig. 1), which means  $\dim \mathcal{M} = \dim \text{aff}(\mathcal{M})$ . In this case, the manifold is flat, and we have  $\nabla f_{\mathcal{M}} = \nabla f$

Thus the functional becomes

$$\int_{\mathcal{M}} \|\nabla f_{\mathcal{M}}\|^2 = \int_{\mathcal{M}} \|\nabla f\|^2 = \int_{\mathcal{M}} \mathbf{a}^T \mathbf{a} = \text{vol}(\mathcal{M}) \mathbf{a}^T \mathbf{a}$$

In this case  $H = \text{vol}(\mathcal{M})\mathbf{I}$ , where  $\mathbf{I}$  is the identity matrix. Then Eq. (4) is equivalent to the following equation:

$$\mathbf{a} = \lambda \mathbf{C} \mathbf{a} \iff \mathbf{C} \mathbf{a} = \frac{1}{\lambda} \mathbf{a} \tag{6}$$

Therefore this problem is totally depend on the constraint.

Since  $C_{ij} = \int_{\mathcal{M}} \mathbf{x}_i^i \mathbf{x}_i^j$ , if we discretize the integral into the sum over data points, we have

$$C_{ij} = \sum_k \mathbf{x}_k^i \mathbf{x}_k^j$$

Here  $\{\mathbf{x}_k\}$  are data points. We see that  $C$  is very similar to the covariance matrix. In this case, Eq. (4) maximizes the covariance. However, there are two differences. First, the data points may not be zero centered. We will discuss this point in the next section. Second, when the manifold has multiple connected components, we have

$$C = \sum C_i$$

where  $C_i$  is the constraint matrix of each connected component. While for covariance matrix, it will sum over all data points. Thus Eq. (4) considers more cluster information.

3. Our approach

In this section, we consider how to give a good approximation of the functional and the constraint. We find the affine hull of each connected component is crucial for Eq. (4). We will define

our objective function on graph, as graph is a good linear approximation to the manifold. It also considers local information of the manifold and more importantly the affine hull of the graph is an approximation to the affine hull of the manifold.

Given  $\mathbf{x}_i \in \mathbb{R}^N, i=1,2,\dots,n$ , we aim to find a good representation in a lower dimensional Euclidean space  $\mathbb{R}^d$ . We assume the data points  $\mathbf{x}_i$  reside on a manifold embedded in  $\mathbb{R}^N$ . Then we construct a neighborhood graph, either by  $k$ -nearest neighbors or  $\epsilon$ -neighbors.

In this work we consider more geometrical structure of this graph, including edge length, edge orientation, etc. For each edge, we need to define an orientation, arbitrary but fixed, so that gradient can be computed. For convenience, let  $e_{ij}$  denote a vector starting from  $\mathbf{x}_i$  to  $\mathbf{x}_j$ , i.e.,  $e_{ij} = \mathbf{x}_j - \mathbf{x}_i$ . Denote edge length by  $d_{ij}$ ,  $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$ . Let  $f$  be a linear function,  $f: \mathbb{R}^N \rightarrow \mathbb{R}$  and denote  $y_i = f(\mathbf{x}_i)$ . Then the gradient on each edge is

$$\nabla f_{e_{ij}} = \frac{y_j - y_i}{d_{ij}} \frac{\mathbf{x}_j - \mathbf{x}_i}{d_{ij}}$$

It is important to note that the computation of gradient is independent to the orientation. Different from standard spectral graph methods which mainly consider the connectivity of graph, our approach explicitly make use of the edge length which reflects the geometrical structure of the manifold.

3.1. Approximately harmonic projection

As the structure of the manifold is unknown, it is difficult to give a good partition of the manifold. Thus we use the ‘‘bottom-up’’ strategy. We define the integral on each edge. As it is an approximation to Eq. (1), we call it Approximately Harmonic Projection (AHP). It solves the following problem

$$\min \sum_{i \sim j} \int_{e_{ij}} \|\nabla f_{e_{ij}}\|^2 dt \quad \text{s.t.} \sum_{i \sim j} \int_{e_{ij}} f(\mathbf{x}(t))^2 dt = 1.$$

Here  $t$  is the arc length of  $e_{ij}$ . As  $\|\nabla f_{e_{ij}}\|^2 = ((y_j - y_i)/d_{ij})^2$ , the objective function becomes

$$\begin{aligned} \sum_{i \sim j} \int_{e_{ij}} \|\nabla f_{e_{ij}}\|^2 dt &= \sum_{i \sim j} \int_0^{d_{ij}} \left( \frac{y_j - y_i}{d_{ij}} \right)^2 dt = \sum_{i \sim j} \frac{1}{d_{ij}} (y_j - y_i)^2 \\ &= 2\mathbf{y}^T (D' - W') \mathbf{y} \end{aligned} \tag{7}$$

where  $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ , and  $W'$  is a weight matrix,  $W'_{ij} = 1/d_{ij}$  if  $\mathbf{x}_i$  and  $\mathbf{x}_j$  are connected, otherwise it is set to zero.  $D'$  is a diagonal matrix whose entries are column (or row) sum of  $W'$ .

For each edge  $e_{ij}$ , we have  $f(\mathbf{x}(0)) = y_i, f(\mathbf{x}(d_{ij})) = y_j$ . As  $f$  is linear on each edge, then  $f(\mathbf{x}(t)) = y_i + (t/d_{ij})(y_j - y_i)$ . Hence the constraint will be

$$\begin{aligned} \sum_{i \sim j} \int_{e_{ij}} f(\mathbf{x}(t))^2 dt &= \sum_{i \sim j} \int_0^{d_{ij}} \left( y_i + \frac{t}{d_{ij}} (y_j - y_i) \right)^2 dt \\ &= \frac{1}{3} \sum_{i \sim j} d_{ij} (y_i^2 + y_i y_j + y_j^2) \\ &= \frac{1}{3} \mathbf{y}^T \left( D'' + \frac{1}{2} W'' \right) \mathbf{y} \end{aligned} \tag{8}$$

where  $W''$  is a weight matrix, and  $W''_{ij} = d_{ij}$  if  $\mathbf{x}_i$  and  $\mathbf{x}_j$  are connected, otherwise it is set to zero.  $D''$  is a diagonal matrix whose entries are column (or row) sum of  $W''$ .

Finally, we get our approximately harmonic projection by solving the following minimization problem:

$$\begin{aligned} \min \quad & \mathbf{y}^T (D' - W') \mathbf{y} \\ \text{s.t.} \quad & \mathbf{y}^T (D'' + \frac{1}{2} W'') \mathbf{y} = 1 \end{aligned}$$

By noticing  $y_i = \mathbf{a}^T \mathbf{x}_i$ , it becomes

$$\begin{aligned} \min \quad & \mathbf{a}^T X(D' - W')X^T \mathbf{a} \\ \text{s.t.} \quad & \mathbf{a}^T X(D'' + \frac{1}{2}W'')X^T \mathbf{a} = 1 \end{aligned}$$

Therefore, the minimization problem also turns out to be a generalized eigenvector problem:

$$X(D' - W')X^T \mathbf{a} = \lambda X(D'' + \frac{1}{2}W'')X^T \mathbf{a} \quad (9)$$

### 3.2. PCA and trivial solution

In continuous cases, we have proved the matrix  $C$  is positive definite when it is in the affine hull of the manifold. In discrete cases, we should find the affine hull, and prove this is a well posed problem. It is interesting to note that PCA is a good choice for finding the affine hull.

**Proposition 3.1.** Given  $\mathbf{x}_i \in \mathbb{R}^N$ ,  $i = 1, \dots, n$ . Let  $\mathbf{x}'_i = \mathbf{x}_i - \bar{\mathbf{x}}$ , where  $\bar{\mathbf{x}} = 1/n \sum \mathbf{x}_i$ . Let  $X' = (\mathbf{x}'_1, \dots, \mathbf{x}'_n)$ , and  $f$  be a linear function on  $\mathbb{R}^N$ ,  $f(\mathbf{x}) = \mathbf{a}^T \mathbf{x}$ . Then the following statements are equivalent:

1.  $f(\mathbf{x}_i) = \text{const}$ ,
2.  $f(\mathbf{x}'_i) = 0$  and
3.  $X'X'^T \mathbf{a} = 0$ .

**Proof.** It is not difficult to show statement 1 is equivalent to 2.  $2 \Rightarrow 3$  is straightforward. Then we only need to show  $3 \Rightarrow 2$ .

$$X'X'^T \mathbf{a} = 0 \Rightarrow \mathbf{a}^T X'X'^T \mathbf{a} = 0 \Rightarrow \sum (\mathbf{a}^T \mathbf{x}'_i)^2 = 0$$

Then  $f(\mathbf{x}'_i) = \mathbf{a}^T \mathbf{x}'_i = 0$ .  $\square$

The statement 1 means this is a trivial solution, and 3 shows the trivial solution happens to be the zero eigenvector of PCA. Therefore, we should remove all these trivial solutions. Thus PCA is a very good choice for pre-processing. After PCA, as  $D'' + \frac{1}{2}W''$  is a positive definite matrix, one can prove the constraint is a positive definite. Thus this is a well posed problem.

### 3.3. Translation invariance

For Eq. (9), the left-hand side is translation invariant, as the gradient is translation invariant. The right-hand side is not translation invariant. We can solve this problem by adding one translation term into the projection. Let  $f(\mathbf{x}) = \mathbf{a}^T \mathbf{x} + b$ , where  $b$  is a constant. Then we should add one constraint for computing  $b$ . One natural constraint is  $\int_{\mathcal{M}} f = 0$ . Because constant function is an optimal solution for minimizing the functional, the rest solutions should be orthogonal to it. In discrete form, we have

$$\sum_{i=1}^n y_i = \sum_{i=1}^n (\mathbf{a}^T \mathbf{x}_i + b) = \mathbf{a}^T \left( \sum_{i=1}^n \mathbf{x}_i \right) + nb = 0.$$

If we move the center of data points to the origin,  $f$  only differs by a constant. In this case, we have  $b' = -\mathbf{a}^T (\sum \mathbf{x}_i) / n = 0$ . Hence we only need to move the center of the data points to the origin, and the corresponding translation term vanishes. Thus  $f$  is a linear function in new coordinates. Also it is not difficult to show AHP is rotation invariant.

We summarize our algorithm as follows: First we apply PCA to the data points as pre-processing. Then we construct an adjacency graph, either by  $\epsilon$ -neighborhoods or  $k$ -nearest neighbors. We compute the eigenvalues and eigenvector of Eq. (9). If we want to embed the manifold to  $\mathbb{R}^d$ , let the column vectors  $\mathbf{a}_0, \dots, \mathbf{a}_{d-1}$  be solutions of Eq. (9) ordered by their eigenvalues,  $\lambda_0 \leq \dots \leq \lambda_{d-1}$ .

Then the embedding is as follows:

$$\mathbf{x}_i \rightarrow \mathbf{y}_i = A^T \mathbf{x}_i, \quad A = (\mathbf{a}_0, \mathbf{a}_1, \dots, \mathbf{a}_{d-1})$$

where  $\mathbf{y}_i$  is a  $d$ -dimensional vector, and  $A$  is a  $N \times d$  matrix.

### 3.4. Computational complexity analysis

The complexity of AHP is dominated by three parts:  $k$  nearest neighbor search, matrix multiplication, and solving a generalized eigenvector problem. Consider  $n$  data points in  $N$ -dimensional space. For the  $k$  nearest neighbor search, the complexity is  $\mathcal{O}((N+k)n^2)$ .  $Nn^2$  stands for the complexity of computing the distances between any two data points.  $kn^2$  stands for the complexity of finding the  $k$  nearest neighbors for all data points. The complexity for calculating the matrices  $X(D' - W')X^T$  and  $X(D'' + \frac{1}{2}W'')X^T$  are  $\mathcal{O}(n^2N + nN^2)$ . The third part is solving a generalized eigenvector problem  $Aa = \lambda Ba$ , where  $A$  and  $B$  are  $N \times N$  matrices. To solve this generalized eigenvector problem, we need first to compute the singular value decomposition (SVD) of the matrix  $B$ . The complexity of SVD is  $\mathcal{O}(N^3)$ . Then, to project the data points into  $d$ -dimensional subspace, we need to compute the first  $d$  smallest eigenvectors of an  $N \times N$  matrix, whose complexity is  $\mathcal{O}(dN^2)$ . Thus, the total complexity of the generalized eigenvector problem is  $\mathcal{O}((N+d)N^2)$ . Therefore, the time complexity of the AHP algorithm is  $\mathcal{O}((N+k)n^2 + (n+N+d)N^2)$ . Since  $k \ll n$  and  $d \ll N$ , the complexity of AHP is determined by the number of data points and the number of features.

## 4. Experimental results

In this section, we give several experimental results on real data. We show our method is very suitable for dimensionality reduction and clustering.

### 4.1. Synthetic example

Two synthetic examples are given in Fig. 3. Both of the two data sets correspond essentially to a one-dimensional manifold. The first row shows the case the projection is not orthogonal. In this case, the affine hull of two connected components overlaps which is exactly the ambient space. Our method also find a projection which seems *orthogonal* to the connected component. It is not difficult to see from the mathematical result, as our method always try to collapse the connected component. The first basis of the second example is the optimal case, as it is orthogonal to the affine hull of each circle. Thus it separates two circles very well.

### 4.2. Handwritten digits

To demonstrate the effectiveness of our method, we performed experiments on the USPS handwritten digits database.<sup>1</sup> For each digit from zero to nine, there are 1100 greyscale images. The images are downsampled to  $16 \times 16$  resolution, so the dimension of the ambient space is 256.

Fig. 4 shows the result of using AHP to embed the data set of digit "1" onto a two-dimensional plane. We traverse along the principal direction within the projected manifold. As we can see, the horizontal direction appears to describe the slant of the digits, while the vertical direction describes the change of digit width. The slant and width changes along the paths are quite smooth,

<sup>1</sup> <http://www.cs.toronto.edu/~roweis/data.html>

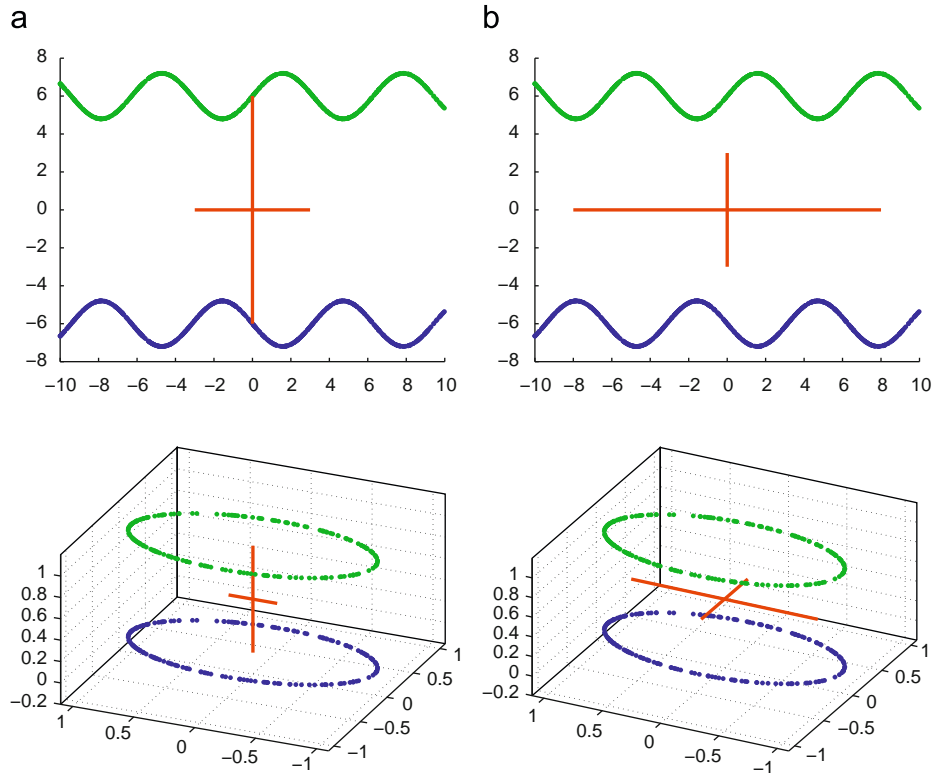


Fig. 3. The first column show the result of AHP, the second column show the result of PCA. The line segments describe the two bases. The first basis shown as a longer line segment, and the second basis is shown as a shorter line segment. (a) AHP and (b) PCA.

thus AHP indeed captures some meaningful structure and this projection is very smooth.

4.3. Manifold of face images

Our second example is the embedding of the manifold of face images. The face images data set used here is the same as that used in [1]. This data set contains 1965 face images taken from sequential frames of a small video. The size of each images is  $20 \times 28$ , with 256 gray levels per pixel. Thus each face image is represented by a point in the 560-dimensional ambient space.

Fig. 5 shows the projection results. The images of faces are projected onto a two-dimensional plane described by the first two coordinates of AHP. The data is clearly divided into two parts (components). The left part are the faces with open mouth, and the right part are the faces with closed mouth. The bottom images correspond to points along the right path (linked by solid line), illustrating one particular mode of variability in pose.

4.4. Face clustering

We applied the AHP to ORL face image data set. The data set contains 400 faces images taken from 40 people, each people having 10 face images. The label information is corresponding to people identification. The size of each image is  $32 \times 32$ , with 256 gray levels per pixel. Thus, each face image is represented by a point in the 1024-dimensional ambient space.

To demonstrate how AHP improves the performance of clustering, we compared seven methods listed below:

- K-means on the original face image matrix (K-means), which is treated as our baseline.

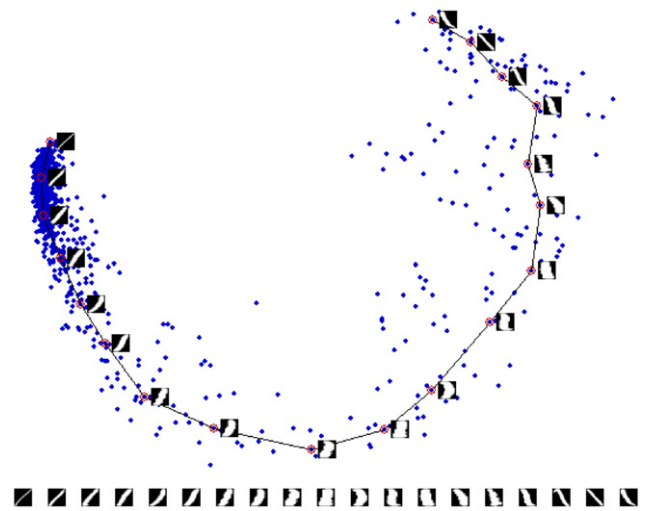
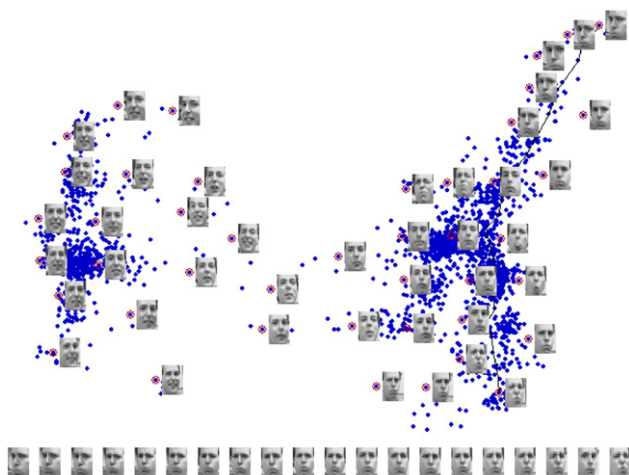


Fig. 4. AHP on handwritten digit "1". Top: embedding of digit images onto  $\mathbb{R}^2$ . Bottom: images corresponding to points along the path linked by solid line.

- K-means after AHP (AHP).
- K-means after PCA (PCA).
- Nonnegative Matrix Factorization based clustering [6].
- K-means after LLE (LLE).
- K-means after ISOMAP (ISOMAP).
- K-means after LE (LE).

Note that, AHP needs to construct a graph on face images. The number of nearest neighbors was set to 5. We use accuracy and mutual information to measure the quality of the clustering



**Fig. 5.** A two dimensional representation of the set of all images of faces using AHP. The bottom images are corresponding to points along the path linked by solid line.

**Table 1**  
Accuracy.

K	K-means	AHP	PCA	NMF	LLE	ISOMAP	LE
2	0.9400	0.9390	0.9220	0.9070	0.9480	0.9410	0.8840
3	0.9060	0.9207	0.8893	0.7693	0.9260	0.9193	0.6520
4	0.8765	0.8955	0.8545	0.7390	0.8650	0.8935	0.6990
5	0.7896	0.8304	0.7816	0.6585	0.8072	0.8352	0.7152
6	0.7477	0.7997	0.7303	0.6160	0.7683	0.8036	0.6930
7	0.7769	0.8171	0.7851	0.6191	0.7874	0.8128	0.7165
8	0.7585	0.7952	0.7555	0.6002	0.7657	0.7895	0.7080
9	0.7364	0.7940	0.7502	0.5898	0.7395	0.7842	0.7077
10	0.7218	0.7764	0.7130	0.5564	0.7120	0.7494	0.6758

results. Please see [9] for the detailed definition of these two standard measures for clustering. Tables 1 and 2 show the experimental results on the ORL face image data set. The evaluations were conducted with different number of clusters, ranging from 2 to 10. For each given cluster number  $K$ , 50 tests were conducted on different randomly chosen clusters, and the average performance was computed over these 50 tests. For each test,  $K$ -means algorithm was applied 20 times with different start points, and the best result in terms of the objective function of  $K$ -means was recorded.

For the dimension of the subspace, we choose the first  $K-1$  dimensions in AHP. As can be seen, AHP performs better than LE in this problem. The main reason is the data points are sparse and there are multiple connected components. In this example, LE is sensitive to the graph structure. For obtaining better result, we have also adjusted the parameter of weight in LE. Moreover, our method is as good as ISOMAP. However the computational cost of ISOMAP is more expensive than our method.

## 5. Conclusion

In this paper, we provide a theoretical analysis on linear harmonic manifold learning methods based on the Dirichlet

**Table 2**  
Mutual Information.

K	K-means	AHP	PCA	NMF	LLE	ISOMAP	LE
2	0.8191	0.8143	0.7589	0.7185	0.8420	0.8169	0.6879
3	0.8400	0.8522	0.8018	0.6251	0.8629	0.8423	0.4763
4	0.8341	0.8377	0.8022	0.6404	0.8360	0.8377	0.6412
5	0.7647	0.7931	0.7518	0.5841	0.8058	0.8023	0.6981
6	0.7474	0.7726	0.7217	0.5718	0.7749	0.7762	0.7088
7	0.7906	0.8140	0.7866	0.6100	0.8148	0.8098	0.7703
8	0.7880	0.8054	0.7795	0.6266	0.8074	0.8021	0.7682
9	0.7826	0.8189	0.7817	0.6294	0.7988	0.8069	0.7751
10	0.7683	0.8062	0.7619	0.6085	0.7797	0.7861	0.7578

integral, which shows that the affine hull of the manifold is essential. And there is a close connection between the affine hull of the manifold and PCA. We show the geometric meaning of optimal projection that the gradient of the function on ambient space is orthogonal to the span of the affine hulls of the connected components. More importantly, it shows the relationship of function on manifold and function on ambient space when it is optimal.

For multiple connected components, the optimal projection will collapse the connected components. Thus it is especially suitable for clustering. If the manifold is a compact domain of its affine hull, then the constraint plays a central role which is very similar to covariance matrix. Based on our theoretical analysis, we propose a new linear dimensionality reduction algorithm called Approximately Harmonic Projection. The experimental results on real data sets are impressive.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 60875044 and 60702072, in part by the National Key Basic Research Foundation of China under Grant 2009CB320801, and in part by the Program for Changjiang Scholars and Innovative Research Team in University (IRT0652, PCSIRT).

## References

- [1] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science* 290 (5500) (2000) 2323–2326.
- [2] J. Tenenbaum, V. de Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science* 290 (5500) (2000) 2319–2323.
- [3] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, *Advances in Neural Information Processing Systems*, vol. 14, MIT Press, Cambridge, MA, 2001, pp. 585–591.
- [4] K.Q. Weinberger, F. Sha, L.K. Saul, Learning a kernel matrix for nonlinear dimensionality reduction, in: *Proceedings of the 21st International Conference on Machine Learning*, Banff, Alberta, Canada, 2004.
- [5] X. He, P. Niyogi, Locality preserving projections, *Advances in Neural Information Processing Systems*, vol. 16, MIT Press, Cambridge, MA, 2003.
- [6] C.-J. Lin, Projected gradient methods for nonnegative matrix factorization, *Neural Computation* 19 (10) (2007) 2756–2779.
- [7] M. Spivak, second ed., *A Comprehensive Introduction to Differential Geometry*, vol. 1, Publish or Perish Inc., Wilmington, DE, 1979.
- [8] J.-B. Hiriart-Urruty, C. Lemarechal, *Fundamental of Convex Analysis*, Springer, Berlin, New York, 2004.
- [9] D. Cai, X. He, J. Han, Document clustering using locality preserving indexing, *IEEE Transactions on Knowledge and Data Engineering* 17 (12) (2005) 1624–1637.