Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Appl. Comput. Harmon. Anal. 21 (2006) 135–144

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

ELSEVIER

Letter to the Editor

Spectral independent component analysis

A. Singer

Department of Mathematics, Program in Applied Mathematics, Yale University, 10 Hillhouse Ave., P.O. Box 208283, New Haven, CT 06520-8283, USA

Available online 26 May 2006

Communicated by Ronald R. Coifman on 12 March 2006

Abstract

Independent component analysis (ICA) of a mixed signal into a linear combination of its independent components, is one of the main problems in statistics, with wide range of applications. The un-mixing is usually performed by finding a rotation that optimizes a functional closely related to the differential entropy. In this paper we solve the linear ICA problem by analyzing the spectrum and eigenspaces of the graph Laplacian of the data. The spectral ICA algorithm is based on two observations. First, independence of random variables is equivalent to having the eigenfunctions of the limiting continuous operator of the graph Laplacian in a separation of variables form. Second, the first non-trivial Neumann function of any Sturm–Liouville operator is monotonic. Both the degenerate and non-degenerate spectrums corresponding to identical and non-identical sources are studied. We provide successful numerical experiments of the algorithm.

1. Introduction

Independent component analysis (ICA) is an important problem in statistics [1–3]. The need for ICA is encountered often in signal processing [4], where the signal must be decomposed into its independent components; this process is also known as blind source separation, or the cocktail party problem. Medical and biological applications include separating the EEG and MEG into different source signals [5], and analyzing DNA microarray data [6], to name just a few.

The linear ICA problem is formulated mathematically as follows. Let the sources $S_1, S_2, \dots, S_n$ be n unknown independent random variables, and A be an $m \times n$ unknown constant mixing matrix. The task is to find the mixing matrix, and thus also the sources, from N different observations of the m-vector X

$$X^{(i)} = AS^{(i)}, \quad i = 1, 2, \dots, N. \tag{1.1}$$

It is well known that A may be assumed to be an orthogonal square matrix ($m = n$), and all sources to have zero mean and unit variance ($\mathbb{E}S_j = 0, \mathbb{E}S_j^2 = 1$). This is achieved by first subtracting the mean value of X and then applying the principal component analysis (PCA) algorithm ([3,7] among others) to the original data, which basically diagonalizes the covariance matrix of X.

E-mail address: amit.singer@yale.edu.

1063-5203/$ – see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.acha.2006.03.003

The starting point of most linear ICA algorithms is casting it as an optimization problem ([1,2] and references therein). For example, the joint differential entropy $H(S)$ of n random variables $S = (S_1, \dots, S_n)$ equals the sum of their marginal entropies $\sum_{j=1}^{n} H(S_j)$ iff the random variables are independent. Furthermore, the joint entropy is invariant to orthogonal transformations, i.e., $H(AX) = H(X)$, for A orthogonal. Therefore, one tries to find an orthogonal transformation A that maximizes the sum of the projected marginal entropies, i.e., $\max_A \sum_{j=1}^{n} H[(A^{-1}X)_j]$. The global search is usually done iteratively by a Newton or a gradient method. However, it suffers from two drawbacks. First, the marginal entropy functional is a non-linear function of the density. Therefore, the calculation of the gradient is complicated. Several approximation schemes have been cleverly devised for this task alone. Second, as usually is the case with global gradient searches, the iterative scheme may find a local maximum instead of the global one. The severity of this problem increases with the number of dimensions n.

In this paper we adapt a different approach to the linear ICA problem that is not based on differential entropy. Our method is based on the graph Laplacian of the data. The eigenvectors of the discrete graph Laplacian are approximations of the continuous backward Fokker–Planck operator, which is the Laplacian operator and a potential term that is a function of the density of the data points. For data originating from an orthogonal transformation of independent sources, the eigenfunctions of the backward Fokker–Planck operator have a separation of variables form, because both the Laplacian operator and the potential are separated. The backward Fokker–Planck operator is separated to n one-dimensional Sturm–Liouville operators, with Neumann boundary conditions. The classical Sturm–Liouville theory guarantees that the first non-trivial eigenfunction of these one-dimensional operators is monotonic. Therefore, exploring the spectrum and eigenvectors of the graph Laplacian enables us to find the orthogonal transformation A.

The paper is organized as follows. Section 2 contains a brief review of the graph Laplacian method and the way it gives rise to the limiting continuous backward Fokker–Planck operator. In Section 3 we show that for the linear ICA problem the Fokker–Planck operator separates to many one-dimensional Sturm–Liouville problems. In Section 4 we make use of the monotonic property of the first non-trivial eigenfunction of any Sturm–Liouville problem to derive a statistic that reveals the lowest frequency independent component. The degenerate case arises when two components have the same distribution, so we look deeper into the spectrum to recover the independent components. In Section 5 we describe how to extract all components, explain the capability of the algorithm to cope with noise, and discuss its limitations in high dimensions. Finally, in Section 6 we provide numerical examples of both the non-degenerate and degenerate cases.

2. The graph Laplacian and the backward Fokker–Planck operator

In this section we briefly review the graph Laplacian method and its connection to the backward Fokker–Planck operator. Graph Laplacians are widely used in machine learning for dimensionality reduction, semi-supervised learning and spectral clustering ([8–15] and references therein). In these setups one is usually given a set of N data points $X^{(1)}, X^{(2)}, \dots, X^{(N)} \in \mathcal{M}$, where $\mathcal{M} \subset \mathbb{R}^n$ is a Riemannian manifold with $\dim \mathcal{M} = d \leqslant n$. The points are given as vectors in the ambient space $\mathbb{R}^n$ and the task is to find the unknown underlying manifold $\mathcal{M}$, its geometry and its low-dimensional representation.

The starting point of spectral methods is to extract an $N \times N$ weight matrix W from a suitable semi-positive kernel k as follows

$$W_{ij} = k\left(\left\|X^{(i)} - X^{(j)}\right\|^2/2\varepsilon\right), \tag{2.1}$$

where $\|\cdot\|$ is the Euclidean distance in the ambient space $\mathbb{R}^n$ and $\varepsilon^{1/2} > 0$ is the bandwidth of the kernel. A popular choice of kernel is the exponential kernel $k(x) = e^{-x}$, though other choices are also possible.

The weight matrix W is then normalized to be row stochastic, by dividing it by a diagonal matrix D whose elements are the row sums of W

$$D_{ii} = \sum_{j=1}^{N} W_{ij} \tag{2.2}$$

and the (negative defined) graph Laplacian L is given by

$$L = D^{-1}W - I, \tag{2.3}$$

where I is a $N \times N$ identity matrix. Next, the top few eigenvalues and eigenvectors of $D^{-1}W$ are computed, to be used for data analysis problems such as dimensionality reduction and clustering.

In the case where the data points $\{X^{(i)}\}_{i=1}^{N}$ are independently and uniformly distributed over the manifold $\mathcal{M}$ the graph Laplacian converges to the continuous Laplace–Beltrami operator Δ_M of the manifold [8,11,12,16]. The manifestation of the last statement is that for a smooth function $f : \mathcal{M} \to \mathbb{R}$ we have [17]

$$\frac{1}{\varepsilon}\sum_{j=1}^{N} L_{ij} f\bigl(X^{(j)}\bigr) = \frac{1}{2}\Delta_M f\bigl(X^{(i)}\bigr) + O\left(\frac{1}{N^{1/2}\varepsilon^{1/2+d/4}}, \varepsilon\right). \tag{2.4}$$

Therefore, the eigenvectors of L are approximations of the eigenfunctions of the Laplace–Beltrami operator Δ_M

$$\Delta_M f(X) = -\lambda f(X), \quad X \in \mathcal{M}, \tag{2.5}$$

with the Neumann boundary conditions

$$\nabla_M f(X) \cdot \nu(X) = 0, \quad X \in \partial\mathcal{M}, \tag{2.6}$$

where $\nu(X)$ is an outer normal unit vector to the boundary of the manifold $\partial\mathcal{M}$.

If the data points are independent identical multivariate random variables, but not necessarily uniformly distributed, then the graph Laplacian converges to a backward Fokker–Planck operator [12,13]. Suppose $p(X)$ is the probability density function of the data points over the manifold, and $U(X) = -2\log p(X)$. Then,

$$\frac{1}{\varepsilon}\sum_{j=1}^{N} L_{ij} f\bigl(X^{(j)}\bigr) \approx \frac{1}{2}\bigl[\Delta_M f\bigl(X^{(i)}\bigr) - \nabla_M U\bigl(X^{(i)}\bigr) \cdot \nabla_M f\bigl(X^{(i)}\bigr)\bigr], \tag{2.7}$$

with the same error estimation as in (2.4), and the eigenvectors of L are approximations of the eigenfunctions of the backward Fokker–Planck operator

$$\Delta_M f(X) - \nabla_M U(X) \cdot \nabla_M f(X) = -\lambda f(X), \quad X \in \mathcal{M}, \tag{2.8}$$

satisfying the Neumann boundary conditions (2.6).

3. Separation of variables

In the linear ICA problem (1.1), the manifold is "planar," that is, $\mathcal{M} = \mathbb{R}^n$, and its dimension is $d = n$. More importantly, the density is a product of n one-dimensional densities

$$p(X) = \prod_{j=1}^{n} p_j(S_j), \tag{3.1}$$

where S_j are the components of $S = A^{-1}X$ and p_j is the density of S_j. Therefore, the potential $U(X)$ is a sum of n one-dimensional potentials

$$U(X) = \sum_{j=1}^{n} U_j(S_j), \tag{3.2}$$

where $U_j(S_j) = -2\log p_j(S_j)$. The Laplacian operator Δ is invariant under orthogonal transformations

$$\Delta = \sum_{j=1}^{n} \frac{\partial^2}{\partial X_j^2} = \sum_{j=1}^{n} \frac{\partial^2}{\partial S_j^2}. \tag{3.3}$$

Therefore, the graph Laplacian $\frac{2}{\varepsilon}L$ approximates the continuous backward Fokker–Planck operator $\mathcal{L}$

$$\mathcal{L} = \sum_{j=1}^{n} \frac{\partial^2}{\partial S_j^2} - \frac{\partial U_j(S_j)}{\partial S_j}\frac{\partial}{\partial S_j}. \tag{3.4}$$

Clearly, the operator $\mathcal{L}$ is separable and can be written as

$$\mathcal{L}=\sum_{j=1}^{n}\mathcal{L}_j, \tag{3.5}$$

where each of the $\mathcal{L}_j$'s is a one-dimensional backward Fokker–Planck operator in the interval $a_j < S_j < b_j$ (the endpoints a_j, b_j may be finite or infinite) with Neumann boundary conditions

$$\mathcal{L}_j=\frac{\mathrm{d}^2}{\mathrm{d}S_j^2}-\frac{\mathrm{d}U_j(S_j)}{\mathrm{d}S_j}\frac{\mathrm{d}}{\mathrm{d}S_j}. \tag{3.6}$$

The eigenvalue problem of $\mathcal{L}_j$ is formulated as a Sturm–Liouville problem

$$\frac{\mathrm{d}}{\mathrm{d}S_j}\left(e^{-U_j}\frac{\mathrm{d}\phi_j}{\mathrm{d}S_j}\right)+\lambda_j e^{-U_j}\phi_j=0, \quad S_j\in(a_j,b_j). \tag{3.7}$$

Therefore, the operator $\mathcal{L}_j$ has an infinite set of eigenfunctions and eigenvalues

$$-\mathcal{L}_j\phi_j^{(k)}=\lambda_j^{(k)}\phi_j^{(k)}, \quad k=0,1,2,\ldots, \tag{3.8}$$

with $0=\lambda_j^{(0)}\leqslant\lambda_j^{(1)}\leqslant\cdots$, and $\lambda_j^{(k)}\to\infty$.

The eigenfunctions of $\mathcal{L}$, denoted

$$\phi^{(k_1,k_2,\ldots,k_n)}, \quad k_j=0,1,2,\ldots,\ j=1,2,\ldots,n, \tag{3.9}$$

are the tensor products of the one-dimensional eigenfunctions

$$-\mathcal{L}\phi^{(\mathbf{k})}(S_1,\ldots,S_n)=\lambda^{(\mathbf{k})}\phi^{(\mathbf{k})}(S_1,\ldots,S_n), \quad \mathbf{k}=(k_1,k_2,\ldots,k_n), \tag{3.10}$$

$$\phi^{(\mathbf{k})}(S_1,\ldots,S_n)=\prod_{j=1}^{n}\phi_j^{(k_j)}(S_j), \tag{3.11}$$

$$\lambda^{(\mathbf{k})}=\sum_{j=1}^{n}\lambda_j^{(k_j)}. \tag{3.12}$$

In particular, $\phi_j^{(0)}=1$ and $\lambda_j^{(0)}=0$ for all $j=1,2,\ldots,n$. Therefore, the first eigenfunction of $\mathcal{L}$ is $\phi^{(\mathbf{0})}=1$ with $\lambda^{(\mathbf{0})}=0$ (here $\mathbf{0}=(0,0,\ldots,0)$), which is not of great interest.

4. The second eigenfunction and the first independent component

4.1. The non-degenerate case

More interesting is the second eigenfunction of $\mathcal{L}$. Suppose that the second eigenvalue is non-degenerate. In such a case, there is a unique $j=\arg\min_{j=1,\ldots,n}\lambda_j^{(1)}$, for which the second eigenvalue is minimal. Denote $e_j=(0,\ldots,0,1,0,\ldots,0)$ the standard unit vector with a single 1 at the jth place. Then, the second eigenfunction of $\mathcal{L}$ is

$$\phi^{e_j}(S_1,S_2,\ldots,S_n)=\phi_j^{(1)}(S_j), \quad \lambda^{e_j}=\lambda_j^{(1)}. \tag{4.1}$$

That is, the second eigenfunction is a function of a single coordinate and this coordinate is exactly one of the independent component sources S_j.

Moreover, we are guaranteed by the Sturm–Liouville theory[1] that the second eigenfunction of a Sturm–Liouville problem with Neumann boundary conditions, such as (3.7), is strictly monotonic, and w.l.o.g., monotonic increasing. In other words,

$$\frac{\mathrm{d}\phi_j^{(1)}(S_j)}{\mathrm{d}S_j}>0 \quad \text{for } S_j\in(a_j,b_j). \tag{4.2}$$

[1] The proof is based on the Prüfer transformation.

The monotonic property motivates us to consider the following vector statistics:

$$Z = \frac{1}{N}\sum_{i=1}^{N} \phi^{e_j}\bigl(X^{(i)}\bigr)X^{(i)} = \frac{1}{N}\sum_{l=1}^{n}\sum_{i=1}^{N} \phi_j^{(1)}\bigl(S_j^{(i)}\bigr)S_l^{(i)} A_l, \tag{4.3}$$

where A_l, $l = 1, \dots, n$, are the columns of the mixing matrix

$$A = \begin{pmatrix} | & & | \\ A_1 & \dots & A_n \\ | & & | \end{pmatrix}.$$

It is convenient to represent Z in the basis corresponding the orthogonal matrix A

$$Z = (Z_1, Z_2, \dots, Z_n), \tag{4.4}$$

where

$$Z_l = \frac{1}{N}\sum_{i=1}^{N} \phi_j^{(1)}\bigl(S_j^{(i)}\bigr)S_l^{(i)} \quad \text{for } l = 1, 2, \dots, n. \tag{4.5}$$

The independence of the sources and their zero mean property $\mathbb{E}[S_j] = 0$ (for all $j = 1, \dots, n$) imply

$$\mathbb{E}Z_l = \mathbb{E}\bigl[\phi_j^{(1)}(S_j)S_l\bigr] = \mathbb{E}\bigl[\phi_j^{(1)}(S_j)\bigr]\mathbb{E}[S_l] = 0 \quad \text{for } l \neq j. \tag{4.6}$$

Therefore, the vector $\mathbb{E}Z$ lies in the A_j direction

$$\mathbb{E}Z = \sum_{l=1}^{n} \mathbb{E}Z_l A_l = \mathbb{E}Z_j A_j = \mathbb{E}\bigl[\phi_j^{(1)}(S_j)S_j\bigr]A_j. \tag{4.7}$$

The Z-statistics enable us to recover the jth column A_j of the mixing matrix A, and the corresponding independent component $S_j = A_j^T X$ provided $\mathbb{E}[\phi_j^{(1)}(S_j)S_j]$ does not vanish.

We therefore now prove $\mathbb{E}[\phi_j^{(1)}(S_j)S_j] > 0$. We evaluate

$$\mathbb{E}\bigl[\phi_j^{(1)}(S_j)S_j\bigr] = \int_{a_j}^{b_j} \phi_j^{(1)}(S_j)S_j p_j(S_j)\,\mathrm{d}S_j \tag{4.8}$$

by integration by parts. To this end, consider the indefinite integral

$$I(S_j) = \int_{a_j}^{S_j} t p_j(t)\,\mathrm{d}t, \tag{4.9}$$

which satisfies

$$I(a_j) = 0, \qquad I(b_j) = \mathbb{E}(S_j) = 0,$$
$$\frac{\mathrm{d}}{\mathrm{d}S_j}I(S_j) < 0 \quad \text{for } S_j < 0, \qquad \frac{\mathrm{d}}{\mathrm{d}S_j}I(S_j) > 0 \quad \text{for } S_j > 0.$$

Therefore,

$$I(S_j) < 0 \quad \text{for } S_j \in (a_j, b_j). \tag{4.10}$$

Integration by parts of (4.8) gives

$$\mathbb{E}\bigl[\phi_j^{(1)}(S_j)S_j\bigr] = -\int_{a_j}^{b_j} I(S_j)\frac{\mathrm{d}\phi_j^{(1)}(S_j)}{\mathrm{d}S_j}\,\mathrm{d}S_j, \tag{4.11}$$

because I vanishes at the end points, $I(a_j) = I(b_j) = 0$. The monotonic property of the second eigenfunction (4.2) together with the negativity of I (4.10) yield the required inequality

$$\mathbb{E}\bigl[\phi_j^{(1)}(S_j)S_j\bigr] > 0. \tag{4.12}$$

The convergence rate of $Z \to \mathbb{E}Z$ as the number of data points $N \to \infty$ is $1/\sqrt{N}$. Indeed, consider the covariance matrix of Z. First, for $l_1 \neq l_2$ at least one of the two indices, say l_2, differs from j, and S_{l_2} is independent of both S_{l_1} and S_j. Combining with (4.5) and (4.6) gives

$$\mathbb{E}[Z_{l_1} Z_{l_2}] = 0 \quad \text{for } l_1 \neq l_2. \tag{4.13}$$

As $\mathbb{E}Z_{l_2} = 0$ (4.6) we conclude that the covariance matrix of Z is diagonal

$$\operatorname{Cov}(Z_{l_1}, Z_{l_2}) = 0 \quad \text{for } l_1 \neq l_2, \tag{4.14}$$

that is, the components of the vector Z are uncorrelated.

We proceed to calculate the variance terms. Each component Z_l is a sum of N i.i.d. variables (4.5). Therefore,

$$\operatorname{Cov}(Z_l, Z_l) = \operatorname{Var} Z_l = \frac{1}{N}\operatorname{Var}\bigl[\phi_j^{(1)}(S_j)S_l\bigr] \tag{4.15}$$

and the standard deviation of each component is proportional to $1/\sqrt{N}$ as asserted.

The observable second eigenvector of the graph Laplacian L, denoted φ_2, approximates the unobservable second eigenfunction ϕ^{e_j} of the backward Fokker–Planck operator $\mathcal{L}$. Thus, computing

$$\tilde{Z} = \frac{1}{N}\sum_{i=1}^{N} \varphi_2\bigl(X^{(i)}\bigr)X^{(i)} \tag{4.16}$$

gives an approximation for the jth column of the mixing matrix A.

Note that two approximation errors are involved here. The first is the approximation error of the eigenvectors of L and $\mathcal{L}$, leading to a difference between Z and $\tilde{Z}$. The second approximation error is due to the difference between Z and $\mathbb{E}Z$. Both approximation errors tend to zero as $1/\sqrt{N}$ as the number of data points $N \to \infty$.

4.2. The degenerate case

The second eigenvalue of $\mathcal{L}$ may be degenerate. This happens when $\min \lambda_j^{(1)}$ is attained for two or more j indices, say j_1 and j_2. This will always be the case when there are only two identically independently distributed (i.i.d.) components. Hereafter we assume that the degeneracy is exactly two; the cases of higher degeneracy are treated in the same spirit. In particular, we suspect that the eigenvalues are degenerate when the numeric values of the first two non-trivial eigenvalues of $\frac{2}{\varepsilon}L$ are close.

Ideally, motivated by the previous non-degenerate case, one would expect the corresponding two eigenfunctions $\phi^{e_{j_1}}, \phi^{e_{j_2}}$ to reveal the two independent components S_{j_1}, S_{j_2} and the corresponding columns A_{j_1}, A_{j_2} of the mixing matrix. However, any linear combination of $\phi^{e_{j_1}}$ and $\phi^{e_{j_2}}$ is also an eigenfunction.

Luckily, the spectrum of $\mathcal{L}$ contains more information that can be used. Specifically, the eigenfunction

$$\phi^{e_{j_1}+e_{j_2}}(S) = \phi_{j_1}^{(1)}(S_{j_1})\phi_{j_2}^{(1)}(S_{j_2}), \tag{4.17}$$

becomes very helpful. We assume that the corresponding eigenvalue $\lambda^{e_{j_1}+e_{j_2}} = \lambda_{j_1}^{(1)} + \lambda_{j_2}^{(1)} = 2\lambda_{j_1}^{(1)} = 2\lambda_{j_2}^{(1)}$ is non-degenerate. This assumption usually holds, except in some special cases. For example, it fails in the case of two normally distributed sources, where the eigenfunctions are the Hermite polynomials and the spectrum is $\mathbb{Z}$ (the harmonic quantum oscillator). In such a case the degeneracy is 3 $(0 + 2 = 1 + 1 = 2 + 0)$ instead of 1. We expect two normally distributed variables to be exceptional, as separation is impossible.

Suppose $\phi^{(1)}, \phi^{(2)}$ are the first two degenerate non-trivial eigenfunctions. We may assume they are orthonormal $(\phi^{(i)^T}\phi^{(j)} = \delta_{ij},\ i, j = 1, 2)$ by applying the Gram–Schmidt procedure. We seek a two-dimensional rotation angle θ that gives the separated eigenfunctions

$$\phi^{e_{j_1}} = \cos\theta\phi^{(1)} - \sin\theta\phi^{(2)}, \tag{4.18}$$

$$\phi^{e_{j_2}} = \sin\theta\phi^{(1)} + \cos\theta\phi^{(2)}. \tag{4.19}$$

Therefore, their tensor product is

$$\phi^{e_{j_1}} \otimes \phi^{e_{j_2}} = \frac{1}{2}\sin 2\theta\big(\phi^{(1)} \otimes \phi^{(1)} - \phi^{(2)} \otimes \phi^{(2)}\big) + \cos 2\theta\phi^{(1)} \otimes \phi^{(2)}. \tag{4.20}$$

The dot product of $\phi^{e_{j_1}+e_{j_2}}$ and $\phi^{e_{j_1}} \otimes \phi^{e_{j_2}}$ is

$$\phi^{e_{j_1}+e_{j_2}} \cdot \big(\phi^{e_{j_1}} \otimes \phi^{e_{j_2}}\big) = \frac{1}{2}\sin 2\theta\phi^{e_{j_1}+e_{j_2}} \cdot \big(\phi^{(1)} \otimes \phi^{(1)} - \phi^{(2)} \otimes \phi^{(2)}\big) + \cos 2\theta\phi^{e_{j_1}+e_{j_2}} \cdot \big(\phi^{(1)} \otimes \phi^{(2)}\big). \tag{4.21}$$

This dot product is extremal for the correct value of rotation angle θ, because the two eigenfunctions point along the same direction (or the opposite directions). Therefore, setting the derivative of (4.21) to zero gives θ

$$\tan 2\theta = \frac{1}{2}\frac{\phi^{e_{j_1}+e_{j_2}} \cdot (\phi^{(1)} \otimes \phi^{(1)} - \phi^{(2)} \otimes \phi^{(2)})}{\phi^{e_{j_1}+e_{j_2}} \cdot (\phi^{(1)} \otimes \phi^{(2)})}. \tag{4.22}$$

In practice we observe the N-element eigenvectors of L which are approximations of the unobservable eigenfunctions of $\mathcal{L}$. However, we do not know which eigenvector is actually $\varphi^{e_{j_1}+e_{j_2}}$, the one that approximates $\phi^{e_{j_1}+e_{j_2}}$. It can be decided according to the expected relation between the approximated eigenvalues, but it may not be of great statistical significance.

Two random vectors in a high-dimensional space are approximately perpendicular. The law of large numbers indicates that the angle α between two random vector satisfies $\cos\alpha = O(1/\sqrt{N})$. Therefore, the statistical test based on (4.21) enables us to find both the eigenvector $\varphi^{e_{j_1}+e_{j_2}}$ and the rotation angle θ. The statistical significance of this test improves as $\sqrt{N}$ as the number of data points N increases. Once the angle θ is recovered, Eqs. (4.18) and (4.19) give $\varphi_{j_1}^{(1)}$ and $\varphi_{j_2}^{(1)}$ and we proceed as in the non-degenerate case to find the columns A_{j_1}, A_{j_2} of the mixing matrix, and the corresponding independent components S_{j_1}, S_{j_2}.

5. The next independent components, Noise and the Curse of dimensionality

We proceed to find the next independent components. For the sake of clarity of exposition alone, we assume the non-degenerate case. Obviously, only the first $n-1$ components actually need to be recovered, because the last component is orthogonal to them. Therefore, for $n = 2$ we are done after recovering the first component, so the case $n \geqslant 3$ is considered next.

The component S_{j_2} that we recover next is the one with the second lowest non-trivial eigenvalue $j_2 = \arg\min_{j \neq j_1} \lambda_j^{(1)}$. However, in the spectrum of $\mathcal{L}$ the corresponding eigenvalue $\lambda_{j_2}^{(1)}$ need not be the second non-trivial eigenvalue. It could happen that other eigenvalues of $\mathcal{L}_{j_1}$ are smaller than $\lambda_{j_2}^{(1)}$, that is, $\lambda_{j_1}^{(r)} < \lambda_{j_2}^{(1)}$ with $r > 1$. Still, we are able to find the corresponding eigenvector $\phi^{e_{j_2}}$ by using the fact that $\phi^{e_{j_1}+e_{j_2}} = \phi^{e_{j_1}}\phi^{e_{j_2}}$. This enables us to identify $\phi^{e_{j_2}}$ and $\phi^{e_{j_1}}$ by the same method elaborated for the degenerate case. Moreover, the problem is even simpler than that of the degenerate case, because there is no rotational degree of freedom, so a comparison of only a few eigenvectors is needed.

Once the second component is revealed then our job is done if there are only three independent components ($n = 3$), or we may proceed in the same manner to find the other components ($n > 3$). Throughout this process we identify the origin of eigenvectors and eigenvalues in the spectrum. Note that the components with smaller eigenvalues, in the low frequency part of the spectrum, are recovered before those with high frequencies. In practice, this property enables us to solve the noisy linear ICA problem

$$X = AS + \mu, \tag{5.1}$$

where μ is the noise. The noise can be treated as another independent component (or components). However, the noise is expected to have a much higher frequency compared to the actual signal S. Therefore, we expect the noise to appear further in the spectrum, so we identify the actual sources before we hit its corresponding eigenvectors. We refer to this feature as de-noising. It is known that the graph Laplacian method is quite robust to noise, and in fact is being used for de-noising in various applications.

The spectral ICA method presented in this paper has its limitations. First, for the graph Laplacian to approximate the Fokker–Planck operator the number of data points needs to be large, especially in high dimensions, for the error term in (2.4) to be small. Second, the number of eigenvectors that need to be well approximated for our method to succeed increases with the dimension. Numerically, the lower eigenvectors are better approximated than the higher ones. Therefore, in high dimensions, we might be able to find only the first few independent components. Finally we remark that the output of the spectral ICA method may be used as an initial guess or input for the classical iterative ICA methods, as it is expected to find an orthogonal matrix nearby the global maximum of the specific functional to be optimized.

6. Numerical examples

6.1. The non-degenerate case

Suppose one independent component is uniformly distributed, while the other is a standard normal variable

$$S_1 \sim U[-\sqrt{3}, \sqrt{3}], \tag{6.1}$$

$$S_2 \sim \mathcal{N}(0, 1). \tag{6.2}$$

Clearly, $\mathbb{E}S_j = 0$ and $\mathbb{E}S_j^2 = 1$ for $j = 1, 2$.

The Neumann eigenfunctions of $\mathcal{L}_1 = \frac{\mathrm{d}^2}{\mathrm{d}x^2}$ satisfy

$$\phi'' = -\lambda\phi, \qquad \phi'(-\sqrt{3}) = \phi'(\sqrt{3}) = 0 \tag{6.3}$$

and are given by

$$\phi_1^{(k)} = \cos k\pi\left(\frac{x}{2\sqrt{3}} + \frac{1}{2}\right), \qquad \lambda_1^{(k)} = \frac{\pi^2 k^2}{12}, \quad k = 0, 1, 2, \ldots. \tag{6.4}$$

The standard normal density $p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ gives rise to the potential $U(x) = -2\log p(x) = x^2 + \log 2\pi$. The corresponding backward Fokker–Planck operator is $\mathcal{L}_2 = \frac{\mathrm{d}^2}{\mathrm{d}x^2} - 2x\frac{\mathrm{d}}{\mathrm{d}x}$. Its Neumann eigenfunctions satisfy

$$\phi'' - 2x\phi' + \lambda\phi = 0, \qquad e^{-x^2}\phi'(x) \to 0 \quad \text{as } x \to \pm\infty. \tag{6.5}$$

The eigenfunctions are the Hermite polynomials [18]

$$\phi_2^{(k)} = H_k(x), \qquad \lambda_2^{(k)} = 2k, \quad k = 0, 1, 2, \ldots. \tag{6.6}$$

The first few Hermite polynomials are

$$H_0 = 1, \qquad H_1 = 2x, \qquad H_2 = 4x^2 - 2, \qquad H_3 = 8x^3 - 12x, \qquad \ldots. \tag{6.7}$$

Note that both $\phi_1^{(1)}(x)$ and $\phi_2^{(1)}(x)$ are monotonic functions as predicted by the Sturm–Liouville theory. The first non-trivial eigenvalue of $\mathcal{L}$ is due to $\mathcal{L}_1$, $\lambda^{e_1} = \lambda_1^{(1)} = \frac{\pi^2}{12} = 0.82\ldots$. Therefore, we expect to recover first the uniform distributed component S_1.

The following numerical experiment was performed. We randomly generated $N = 1000$ points $(S_1^{(i)}, S_2^{(i)})$, $i = 1, 2, \ldots, 1000$, according to (6.1) and (6.2). The points were then rotated by $\pi/4 = 45°$

$$X_1^{(i)} = \left(S_1^{(i)} + S_2^{(i)}\right)/\sqrt{2}, \tag{6.8}$$

$$X_2^{(i)} = \left(S_1^{(i)} - S_2^{(i)}\right)/\sqrt{2}. \tag{6.9}$$

Figure 1 is a scatter plot of the observed points $(X_1^{(i)}, X_2^{(i)})$. The graph Laplacian method was applied with $\varepsilon = 0.2$ after the removal of isolated points. The origin of the isolated points is the tail of the normal distribution. It is difficult to obtain a good approximation of the Laplacian at those points, because the density is too small. Therefore, we omit

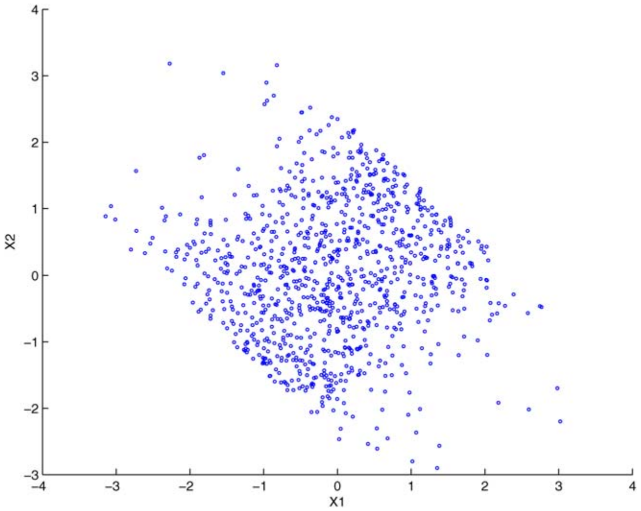

Fig. 1. A scatter plot of $N = 1000$ points $(X_1^{(i)}, X_2^{(i)})$ drawn at random according to (6.1)–(6.2) and (6.8)–(6.9).

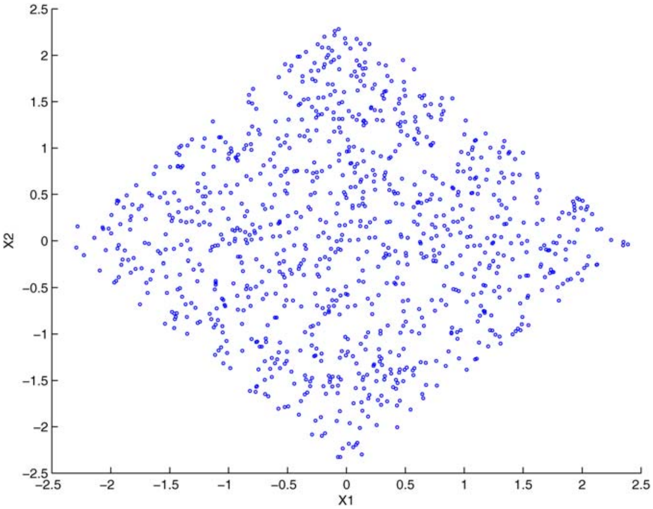

Fig. 2. A scattered plot of $N = 1000$ points $(X_1^{(i)}, X_2^{(i)})$ uniformly distributed in a 45° rotated square.

the isolated points by setting a threshold for the empirical density. The first non-trivial eigenvalue of $\frac{2}{\varepsilon}L$ was found to be $\lambda = 0.90$, which is $O(\varepsilon)$ from the anticipated $\pi^2/12 \approx 0.82$. Computation of $\tilde{Z}_N$ (4.16) resulted in that the rotation angle θ satisfied $\cos\theta = 0.66$ or $\theta \approx 49°$. We repeated this numerical experiment several times just to find that the computed rotation angles approximate the real rotation angles quite well.

6.2. The degenerate case

Suppose both sources are uniformly distributed

$$S_1, S_2 \sim U[-\sqrt{3}, \sqrt{3}] \tag{6.10}$$

and the same rotation of (6.8) and (6.9) is applied. The first two non-trivial eigenfunctions are degenerate with $\lambda^{e_1} = \lambda^{e_2} = \frac{\pi^2}{12}$. The third non-trivial eigenfunction is their tensor product with $\lambda^{e_1+e_2} = \frac{\pi^2}{6}$.

We randomly generated $N = 1000$ points uniformly distributed in the square (6.10) and rotated it by $\alpha = 45°$ (see Fig. 2). We computed the first three non-trivial eigenvectors of the graph Laplacian. The rotation angle θ of the two degenerated eigenvectors with respect to the separated eigenvectors was computed according to the tensor product property (4.22). Then, the separated eigenvectors were calculated according to (4.18) and (4.19) and the Z_N-statistics (4.16) was used to find the rotation angle α. This gave an estimate of $\alpha = 47^\circ$. We repeated the experiment several times to find the predicted rotation angles in good agreement with their actual values.

References

[1] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.

[2] A. Hyvärinen, E. Oja, Independent component analysis: Algorithms and applications, Neural Networks 13 (2000) 411–430.

[3] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer Ser. Statist., Springer, New York, 2001.

[4] P. Common, Independent component analysis, a new concept? Signal Process. 36 (1994) 287–314.

[5] A. Delorme, S. Makeig, EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis, J. Neurosci. Meth. 134 (2004) 9–21.

[6] S. Raychaudhuri, P.D. Sutphin, J.T. Chang, R.B. Altman, Basic microarray analysis: Grouping and feature reduction, Trends Biotechnol. 19 (5) (2001) 189–193.

[7] I.T. Jolliffe, Principal Component Analysis, second ed., Springer Ser. Statist., Springer, New York, 2002.

[8] M. Belkin, Problems of learning on manifolds, Ph.D. dissertation, University of Chicago, 2003.

[9] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, Neural Inform. Process. Syst. 14 (2002) 585–591.

[10] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.

[11] M. Belkin, P. Niyogi, Towards a theoretical foundation for Laplacian-based manifold methods, in: P. Auer, R. Meir (Eds.), Proc. 18th Conf. Learning Theory (COLT), in: Lecture Notes Comput. Sci., vol. 3559, Springer, Berlin, 2005, pp. 486–500.

[12] S. Lafon, Diffusion maps and geometric harmonics, Ph.D. dissertation, Yale University, 2004.

[13] B. Nadler, S. Lafon, R.R. Coifman, I.G. Kevrekidis, Diffusion maps, spectral clustering and eigenfunctions of Fokker–Planck operators, in: Y. Weiss, B. Schölkopf, J. Platt (Eds.), Proc. 2005 Conference, in: Adv. Neural Inform. Process. Syst. (NIPS), vol. 18, MIT Press, Cambridge, MA, 2006.

[14] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, S.W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps, Proc. Natl. Acad. Sci. 102 (21) (2005) 7426–7431.

[15] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, S.W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Multiscale methods, Proc. Natl. Acad. Sci. 102 (21) (2005) 7432–7437.

[16] M. Hein, J. Audibert, U. von Luxburg, From graphs to manifolds—Weak and strong pointwise consistency of graph Laplacians, in: P. Auer, R. Meir (Eds.), Proc. 18th Conf. Learning Theory (COLT), in: Lecture Notes Comput. Sci., vol. 3559, Springer, Berlin, 2005, pp. 470–485.

[17] A. Singer, From graph to manifold Laplacian: The convergence rate, Appl. Comput. Harmon. Anal., in press.

[18] M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions, Dover, New York, 1972.